MCB 137: Physical Biology of the Cell
Homework 2 Solutions

1. The concentration rule of thumb

In the last homework, we worked out the rule of thumb that one molecule per E. coli cell corresponds to a concentration of ≈ 1 nM.

(a) As an application of this idea, how many H^+ ions are there in a bacterial cell if the pH is 7.0?

Solution: We convert pH to concentration of H^+ ions from the relationship

$$[H^+] = 10^{-\mathrm{pH}} M. \tag{1}$$

So for a pH of 7.0, $[H^+] = 10^{-7}$ M, or 10^2 nM. Thus, since we know that 1 molecule per bacterium is equivalent to a concentration of ≈ 1 nM, a pH of 7.0 corresponds with 100 H^+ ions in the cell.

(b) It is very useful to have a sense of how far molecules are apart at a given concentration. Work out a formula that relates the spacing between molecules d to the concentration c. Then, make a plot that shows the distance between molecules as a function of the concentration for concentrations ranging from nM to M.

Solution: Our goal is to determine the average spacing between molecules given a specified concentration. From the molar concentration, we can express the molecular density as

$$\frac{c\,\mathrm{mol}}{L} \cdot \frac{6 \times 10^{23}\,\mathrm{molecules}}{\mathrm{mol}} \cdot \frac{\mathrm{L}}{10^{24}\,\mathrm{nm}^3} = 0.6\,\mathrm{molecules/nm}^3. \tag{2}$$

Inverting this result, we generate the volume of solution occupied by each molecule at molar concentration C M,

$$V = 1.66\,\mathrm{nm}^3/\mathrm{molecule}. \tag{3}$$

The cubed root of this volume thus indicates the average separation between molecules, such that we may conclude

$$d \propto V^{1/3} = \frac{1.18}{c^{1/3}}\,\mathrm{nm}. \tag{4}$$

(c) As an application of your thinking from part (b), explain what the concept of the "critical concentration" is for the polymerization of actin filaments. Then, provide a rough estimate of the mean spacing between actin monomers in a solution at the critical concentration.

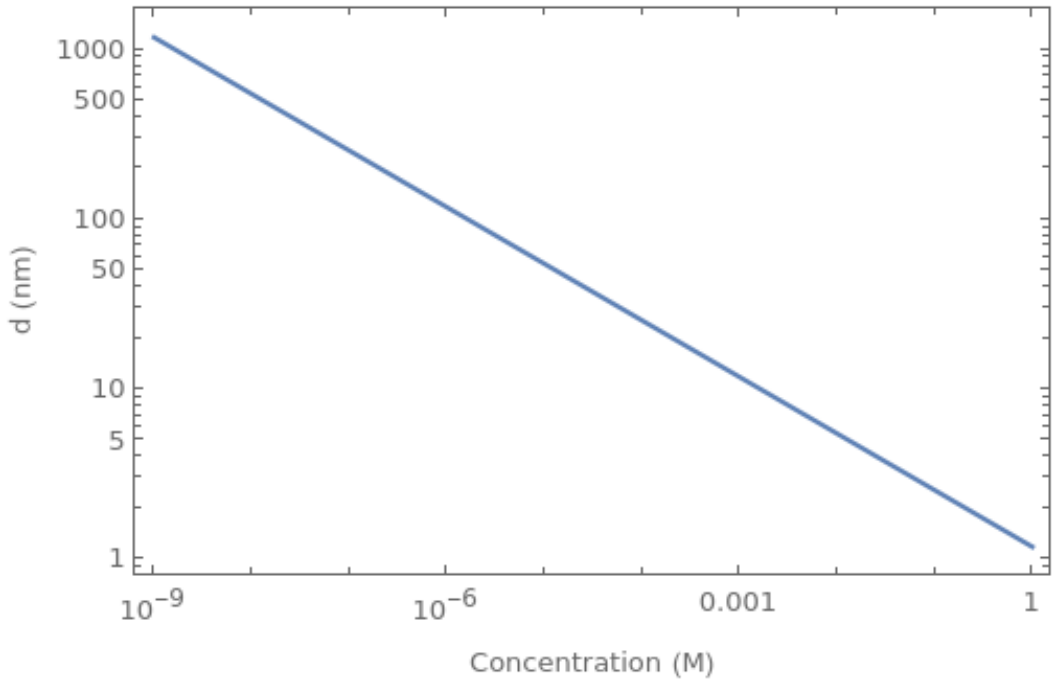

Figure 1: Plot of average separation d between molecules against concentration ranging from nM to M.

Solution: Polymerization in solution can be considered under a simple model: after a nucleation phase where 3-4 monomers randomly interact to form a nucleus which can then be elongated. This elongation is then governed by monomer capture events (monomers polymerize onto one end of the polymer) and monomer escape events (a monomer leaves one end of the polymer). Monomer capture is dependent on the interaction between one end of the polymer and a monomer in solution, and is therefore likelier the smaller the mean separation between monomers in solution - we can capture this by setting the rate of monomer capture to be proportional to the concentration of monomers. Monomer escape, however, does not require interaction with monomers in solution and is therefore independent of the monomer concentration. This can be written as

$$\frac{\mathrm{d}n}{\mathrm{d}t} = k_{on}C - k_{off} \tag{5}$$

where n is the number of monomers that constitute the polymer we are considering, k_{on} (resp. k_{off}) are rate constants for monomer capture (resp. escape), and C is the monomer concentration in solution.

We can then see that the system is in steady state when $C = C_{crit} \equiv \frac{k_{off}}{k_{on}}$. This is referred to as the critical concentration - below this concentration the filaments depolymerize and above this concentration the filaments keep polymerizing, pulling monomers out of solution until the critical concentration is reached.

Actin polymerization, however, is more complicated than this draft model captures. It turns out that the ends of actin filaments behave asymetrically - i.e. actin filaments have a 'plus'

end and a 'minus' end with the 'plus' end having a higher growth (and shrinkage) rate. A better first-order model is given by

$$\frac{\mathrm{d}n}{\mathrm{d}t} = k_{on}^{+}C + k_{on}^{-}C - k_{off}^{+} - k_{off}^{-} \tag{6}$$

where the '+' and '-' superscripts in the rate constants denote the 'plus' and 'minus' end of the filament respectively. (Note: There are more subtleties involved in actin polymerization - e.g. the monomer capture rates also depend on whether the monomer is ADP or ATP bound - but the above is sufficient as a first-order model.)

This system now has three critical concentrations: $C_+ \equiv \frac{k_{off}^{+}}{k_{on}^{+}}$ below which both ends shrink, $C_- \equiv \frac{k_{off}^{-}}{k_{on}^{-}}$ above which both ends grow, and $C_{TM} \equiv \frac{k_{off}^{+}+k_{off}^{-}}{k_{on}^{+}+k_{on}^{-}}$ at which the system reaches steady-state (referred to as 'treadmilling') whereby the 'plus' end grows at the same rate at which the 'minus' end shrinks. For concentrations between C_- and C_+, the 'plus' end grows and the 'minus' end shrinks, with the relative rates of the two processes determining whether or not the filament elongates.

As per BNID 112788, $C_+ \sim 0.06$ μM and $C_- \sim 0.6$ μM, corresponding to a mean separation of ~ 300 nm and ~ 100 nm respectively. C_{TM} lies somewhere between the two, which we estimate to be ~ 0.2 μM and which corresponds to a mean separation of ~ 200 nm.

2. RNA Polymerase and Rate of Transcription

One of the ways in which we are trying to cultivate a "feeling for the organism" is by exploring the processes of the central dogma. Specifically, I want you to have a sense of the number of copies of the key molecular players in the central dogma as well as the rates at which they operate. Further, I argue that it is critical you have a sense of *how* we know these numbers.

(a) If RNA polymerase subunits β and β' together constitute approximately 0.5% of the total mass of protein in an *E. coli* cell, how many RNA polymerase molecules are there per cell, assuming each β and β' subunit within the cell is found in a complete RNA polymerase molecule? The subunits have a mass of 150 kDa each. (Adapted from problem 4.1 of Schleif, 1993.)

Solution: We have discussed measurements showing that the dry weight of an *E. coli.* cell is roughly 30% of its total mass, half of which is protein. The total mass of the cell is estimated to be a picogram from the assumption that its density is nearly that of water and its volume roughly 1 μm^3. It is given in the problem that the β and β' subunits together have a mass of 300 kDa and comprise 0.5% of the protein mass. The number of β, β' subunit pairs is assumed to equal the number of RNA polymerases (RNAP). Putting all this together yields

$$\#\text{ of RNAP} = \frac{\text{total mass of RNAP in cell}}{\text{mass per RNAP}} = \frac{0.15 \times 10^{-12}\text{ g} \times 0.005}{3 \times 10^5\text{ Da} \times 1.6 \times 10^{-24}\text{ g/Da}} = 1.5 \times 10^3. \tag{7}$$

(b) Rifampin is an antibiotic used to treat *Mycobacterium* infections such as tuberculosis. It inhibits the initiation of transcription, but not the elongation of RNA transcripts. The time evolution of an *E. coli* ribosomal RNA (rRNA) operon after addition of rifampin is shown in Figure 3.36(A)–(C). An operon is a collection of genes transcribed as a single unit. Use the figure to estimate the rate of transcript elongation. Use the beginning of the "Christmas-tree" morphology on the left of Figure 3.36(A) as the starting point for transcription.

Solution: Comparing Fig. 3.36(A) and Fig. 3.36(B), one sees that 40 seconds after rifampin addition roughly 1.5 kb of the DNA from the start site has become free of RNAP. The micrographs are aligned well enough that one can assume the left edge in all of them is the start site. Assuming that the last RNAP to initiate transcription did so at nearly the same time as rifampin addition, one can infer that this RNAP transcribed 1.5 kb of DNA in 40 seconds, implying an elongation rate of

$$\text{Elongation Rate} = \frac{1.5\text{ kb}}{40\text{ seconds}} \approx 0.04\text{ kb/sec} \tag{8}$$

Making the same comparison of fig. 3.36(A) and 3.36(C), indicates an elongation rate of 3.5 kb/70 sec = 0.05 kb/sec, or roughly 50 nucleotides/sec.

(c) Using the calculated elongation rate estimate the frequency of initiation off of the rRNA operon. These genes are amongst the most transcribed in *E. coli*.

Solution: The operon is roughly 6 kb long. Given the elongation rate in (b), one RNAP would require $\frac{6\text{ kb}}{0.05\text{ kb/sec}} = 120$ seconds to complete a transcript.
To estimate the rate at which transcripts of the operon are made, one needs the number of RNAP on the operon at any one time. Looking at the micrograph, one can make a rough count that under normal conditions there are 10 – 20 RNAP per kilobase and that the operon is roughly 6 kb long. This implies roughly

$$6\text{ kb} \times 15\text{ RNAP/kb} = 90\text{ RNAP} \tag{9}$$

on the operon, and if each RNAP requires 120 seconds to complete transcription, then the initiation frequency of the operon is 90/120s = $0.75s^{-1}$. This corresponds to a production rate that just over ~ 1 transcript per second.

Alternatively, we can notice that the mean spacing between RNAPs is roughly 1000nt/15 $\approx$ 67nt. The average speed of each such RNAP (from the previous part of the problem) is 50 nt/s. Hence, the initiation frequency is

$$\frac{50 \text{ nt/s}}{67 \text{ nt}} \approx 0.75 \text{ s}^{-1} \tag{10}$$

which is the same as our first result.

3. A feeling for the complete blood count (CBC) test.

Typical results for a complete blood count (CBC) are shown in Table 1. Assume that an adult has roughly 5 L of blood in his or her body. Based on these values estimate:

(a) the number of red blood cells.

Solution: From the table, we see that a typical red blood cell count is around 5 $\times$ 106 cells per L. Scaling this up to the 5 L of blood, we get a total number of red blood cells,

$$5\text{L} \times \frac{10^6 \ \mu\text{L}}{\text{L}} \times 5 \times 10^6 \frac{\text{cells}}{\mu\text{L}} \approx 2 \times 10^{13} \text{ cells}$$

(b) the percentage in volume they represent in blood.

Solution: Red blood cells have a volume of around 100 fL (BNID:110805). This means the total volume of all 20 trillion red blood cells is

$$100 \frac{\text{fL}}{\text{cell}} \times 2 \times 10^{13} \times \frac{\text{L}}{10^{15}\text{fL}} = 2 \text{ L}$$

This means that the red blood cells take up two-fifths, or about 40% of the total blood volume. Coincidentally, the hematocrit value from the table corresponds to this value as empirically determined, and we see that 40% is right on the money for the actual value.

(c) their mean spacing.

Solution: Converting our 5 L of blood into a length scale more meaningful for cells, we get that we have $5 \times 10^{15} \mu m^3$ of blood. Dividing by the number of cells, we get

$$\frac{5 \times 10^{15} \mu m^3 \text{blood volume}}{2 \times 10^{13} \text{cells}} = 250 \mu m^3$$

blood volume that each cell is "allotted". We can alternatively think of this as each cell getting an $\approx 6 \times 6 \times 6 \mu m$ box to call its own, meaning there is $\approx 6 \mu m$ spacing between cells.

(d) the total amount of hemoglobin in the blood.

Solution: Reading off the table, we see that hemoglobin is around $15 g/dL$. Scaling up to 5 L of blood, we get a total hemoglobin mass of

$$15\frac{\text{g}}{\text{dL}} \times 10\frac{\text{dL}}{\text{L}} \times 5L = 750\text{g}$$

(e) the number of hemoglobin molecules per cell.

Solution: To convert the mass of hemoglobin to a number of hemoglobin molecules we simply need to divide by the mass of a hemoglobin molecule. To estimate this, we harken back to Problem 2, where the typical amino acid is 100 Da and the typical protein, meaning the typical protein is around 30 kDa in mass. If we recall that hemoglobin is actually made of 4 subunits, we might adjust our estimate of hemoglobin mass to be four times larger, or around 120 kDa. (It turns out that this is a bit of an over estimate, since each subunit isn't as big as a "typical" protein, but this is sufficient for an order-of-magnitudes estimate). This gives us

$$\frac{750 \text{ g of hemoglobin}}{120 \text{ kDa}} \times \frac{6 \times 10^{20}\text{kDa}}{\text{g}} \approx 4 \times 10^{21}\text{hemoglobin molecules}$$

Finally, the number of hemoglobin per cell is

$$\frac{4 \times 10^{21}\text{hemoglobin}}{2 \times 10^{13}\text{cells}} = 2 \times 10^{8}\text{hemoglobin/cell}$$

(f) the number of white blood cells in the blood.

Solution: Again reading off the table, we see that a typical value for white blood cells is 8×10^3 per μL. Scaling up to 5 L of blood, we get a total number

$$5L \times \frac{10^6 \mu L}{L} \times 8 \times 10^3 \frac{\text{cells}}{\mu L} \approx 4 \times 10^{10}\text{cells}$$

4. Testing the model of nucleolus scaling

Test	**Value**
Red blood cell count (RBC)	Men: $\approx(4.3\text{–}5.7) \times 10^6$ cells/μL
	Women: $\approx(3.8\text{–}5.1) \times 10^6$ cells/μL
Hematocrit (HCT)	Men: $\approx(39\text{–}49)\%$
	Women: $\approx(35\text{–}45)\%$
Hemoglobin (HGB)	Men: $\approx(13.5\text{–}17.5)$ g/dL
	Women: $\approx(12.0\text{–}16.0)$ g/dL
Mean corpuscular hemoglobin (MCH)	$\approx(26\text{–}34)$ pg/cell
MCH concentration (MCHC)	$\approx(31\text{–}37)\%$
Mean corpuscular volume (MCV)	$\approx(80\text{–}100)$ fL
White blood cell count (WBC)	$\approx(4.5\text{–}11) \times 10^3$ cells/μL
Differential (% of WBC):	
Neutrophils	$\approx(57\text{–}67)$
Lymphocytes	$\approx(23\text{–}33)$
Monocytes	$\approx(3\text{–}7)$
Eosinophils	$\approx(1\text{–}3)$
Basophils	$\approx(0\text{–}1)$
Platelets	$\approx(150\text{–}450) \times 10^3$ cell/μL

Table 1: Typical values from a CBC. (Adapted from R. W. Maxwell, Maxwell Quick Medical Reference, Tulsa, Maxwell Publishing Company, 2002.)

In class, we discussed the scale and scaling of various cellular structures and processes. In particular, we talked about the scaling of the size of the nucleolus with the size of the nucleus itself in the C. elegans embryo. Using a simple model, we derived an expression for the number of molecules that make the nucleolus (in this case FIB-1 molecules) given by

$$M = \left(\frac{N}{V} - \frac{k_{off}}{k_{on}}\right)V$$

where N is the total number of FIB-1 molecules inside the nucleus, k_{on} is the rate of FIB-1 incorporation into the nucleolus, k_{off} is the rate with which FIB-1 molecules detach from the nucleolus, and V is the nuclear volume.

We explored two types of experiments. First, we discussed an experiment in which the normal course of development leads to the progressive reduction of cell—and nuclear—size. In this case, the FIB-1 concentration within each nucleus remains constant such that we can rewrite Equation 1 as

$$M = (c_{tot} - c_*)V$$

where $c_{tot} = N/V$ is the FIB-1 concentration and $c_{=k_{off}/k_{on}}$ is the critical concentration at which the nucleolus forms. A second experiment relied on altering the expression of genes that lead to the formation of C. elegans embryos with larger or smaller cells. The assumption is that, in these mutants, the total number of FIB-1 molecules N will not

change, but its nuclear concentration will. As a result, we can write Equation 1 for this case of constant number as

$$M = N - c_* V$$

(a) Read the paper by Weber and Brangwynne (provided on the course website) and, in one short paragraph, explain how they managed to change the size of cells within the embryo and how they ensured that, for all embryo sizes, the total number of FIB-1 molecules remained constant.
The two types of experiments captured by Equations 2 and 3 give us an opportunity to test the predictive power of our model. Specifically, note that Equation 3 predicts that, for the fixed FIB-1 number experiment, the y-intercept of the scaling of the nucleolus with nuclear volume will be given by N while the slope will be c.

Solution: Weber and Brangwynne tested the size of the nucleolus depending on the total number of FIB-1 molcules. The key to their experiment was to keep the number of FIB-1 fixed and change the embryo size. They achieved this via the RNAi process, which allowed them to change the size of the embryo. They ensured the FIB-1 molecules remained constant via 3D timelapse fluorescence imaging of GFP tagged FIB-1 molecules.

(b) Write Python code to plot the data (provided on the course website) and perform a manual linear fit to the data in order to estimate the value of N and c.

Solution:

$$c_* = 0.028$$

$$N = 8$$

(c) Now, use the parameters inferred in (b) to predict the scaling of nucleolar size versus nucleus volume for the fixed FIB-1 concentration experiment. Specifically, draw a plot where you overlay the experimental data with your theoretical prediction.

Solution:

$$M = (c_{tot} - c_*)V$$

$$c_{tot} - c_* = 0.08$$

.

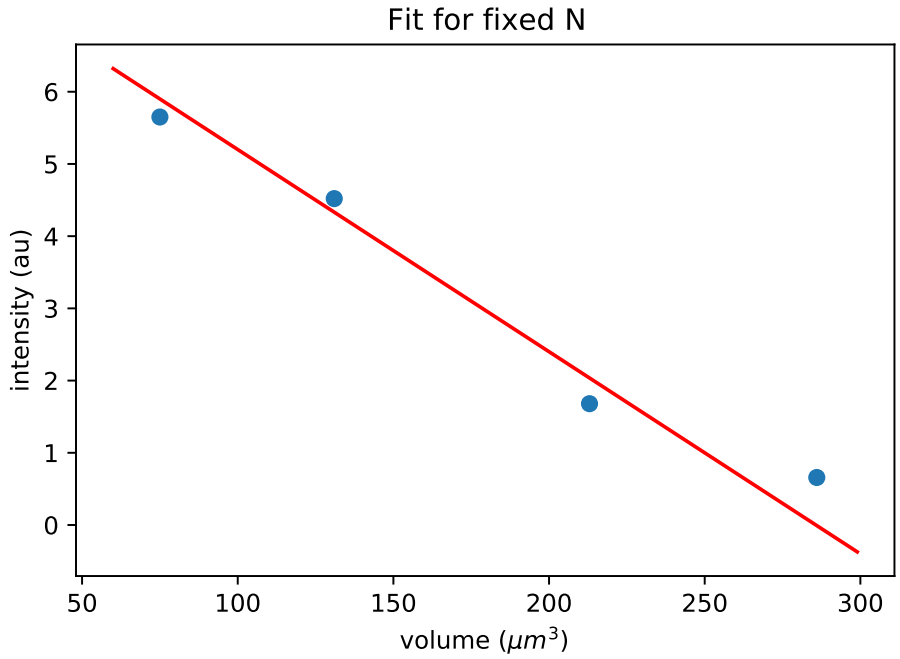

Figure 2: Plot for fixed N

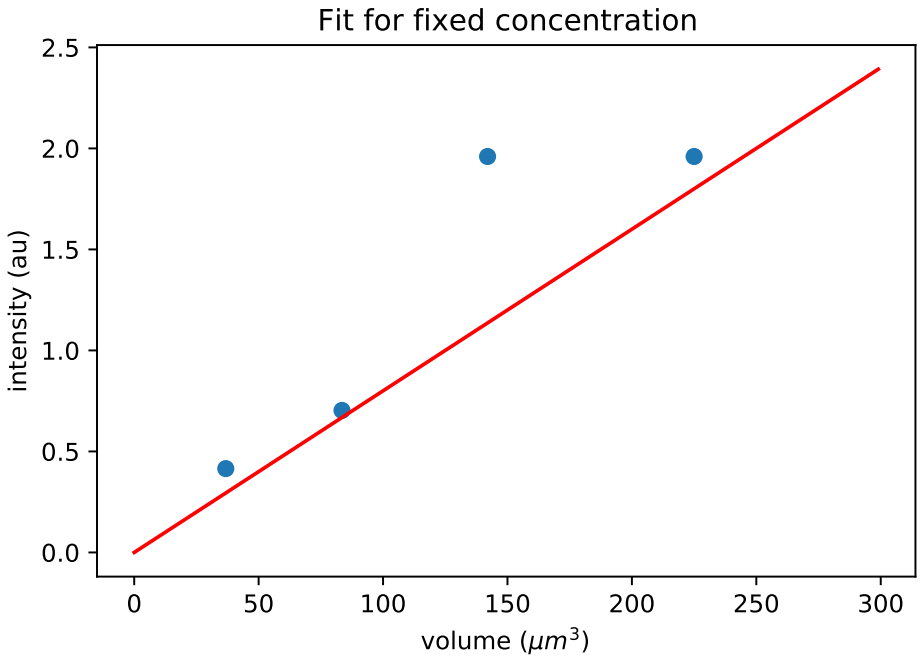

Figure 3: Plot for fixed concentration

5. Post-Translational Modifications and "nature's escape from genetic imprisonment"

In a very interesting article ("Post-translational modification: nature's escape from genetic imprisonment and the basis for dynamic information encoding"), Prof. Jeremy Gunawardena discusses how we should think about post-translational modifications as a way of expanding the natural repertoire of the 20-letter amino acid alphabet. Similarly, Prof. Christopher Walsh (also at Harvard) wrote a whole book entitled "Posttranslational Modifications of Proteins: Expanding Nature's Inventory", again making the point that by adding chemical groups to proteins we can significantly change their properties.

(a) Provide at least one mechanistic idea about how adding a chemical group to a protein can alter its structure or function. Your answer should be offered in less than a paragraph, but should be concrete in its assertions about how these modifications change the protein. Why does Gunawardena refer to this process of post-translational modification as "escape from genetic imprisonment"?

Solution: There are many ways in which adding a chemical group can affect the structure of a protein. For instance, it could promote dimerization by providing an energetically favorable surface for two binding events. Alternatively, adding a charged chemical group could cause increased electrostatic repulsion within the protein. This could cause the protein to "open up," which, among other effects, could alter function by allowing access to previously occluded binding pockets.

(b) As a toy model of the combinatorial complexity offered by post-translational modifications, let's imagine that a protein has N residues that are able to be phosphorylated (NOTE: please comment on which residues these are - the answer is different for bacteria and eukaryotes). How many distinct states of the protein are there as a result of these different phosphorylated states? Make an approximate estimate of the mass associated with a phosphate group and what fraction of the total mass this group represents. Similarly, give some indication of the charge associated with a phosphate group. What ideas do you have about how we can go about measuring these different states of phosphorylation?

Solution: The state of the protein will be determined by which residues are phosphorylated, not merely how many. Therefore, for N residues, there will be 2^N states. For bacteria, the most commonly phosphorylated amino acids are histidine, serine, threonine, and tyrosine , while for eukaryotes they are serine, threonine, and tyrosine. A phosphate group is composed of a phosphorus and four oxygen atoms. Its mass it thus roughly 100 Daltons, which is approximately the same as that of an amino acid.

A typical protein is composed of 300 amino acids, and so, the fraction of the total mass that one represents is $100\,\text{Da}/30,000\,\text{Da} = 5 \times 10^{-3}$. A phosphate group has a charge of $-2e^-$ when bound to a residue, which can be very important for protein's function. The addition of a phosphate group will greatly affect mass to charge ratio, so the use of mass spectrometry would be a very powerful technique for measuring the number of phosphorylated amino acids.

(c) In this part of the problem, we make a very crude estimate of the number of sites on a protein that are subject to phosphorylation. To do so, imagine that the protein is a sphere with N residues. How does the radius of that sphere depend upon the number of residues in the protein? Given that estimate, what is the number of residues that are on the surface? Given that number, what fraction of those are phosphorylatable? Remember, these are crude estimates. Work out these results for a concrete case of a typical protein with roughly 400 amino acids.

Solution: In our toy model we can assume that each residue is itself a sphere. The crude scaling of the radius of our spherical protein will go as $N^{1/3}$. However, we can do a little better. If the amino acid spheres are maximally close packed, then the volume of the protein will be roughly $0.75N \times V_{aa}$. Therefore, the radius of our protein will be given by $R = (0.75N)^{1/3} R_{aa}$. Given that a typical protein has 400 amino acids and a radius of 2.5 nm, the radius of a single amino acid becomes $R_{aa} \approx 0.4$ nm. Therefore, the scaling of the protein radius with the number of amino acids will go approximately as

$$R(N) \approx 0.4 \times (0.75N)^{1/3} \text{ nm}. \tag{11}$$

As a very crude estimate for the total number of residues on the surface of the protein, we can take the total surface area and divide it by the 2D projection area of a single residue, taht is

$$\begin{aligned} S_{\text{array}} &= 4\pi R^2, & (12)\\ S_{aa}^{\text{proj}} &= \pi R_{aa}^2, & (13)\\ N_{\text{surface}} = \frac{S_{\text{protein}}}{S_{aa}^{\text{proj}}} &= 4 \times (0.75N)^{2/3} \approx 3N^{2/3}. & (14) \end{aligned}$$

Taking $N = 400$ for a typical protein, we find that the number of surface residues to be

$$N_{\text{surface}} \approx 150. \tag{15}$$

As discussed in part (b), 4 or 20 amino acids are commonly phosphorylated in bacteria. Making the crude assumption that all amino acids are equally represented on the surface, only 20% of the resiudes, or $150 \times 0.2 \approx 30$, will on average be phosphorylatable. This means that a total of $2^{30} \approx 10^6$ phosphorylation states are available to proteins.

(d) Let's close out these estimates by thinking about a bacterial cell. If all 3×10^6 proteins in such a cell can be phosphorylated with the number of different phosphorylation states that you estimated above, how many distinct cells could we make with all of these different states of phosphorylation.

Solution: If each protein is distinguishable, the total number of cells that could be created will be

$$N_{\text{cells}} = (2^{30})^{3\times 10^6} = 10^{2\times 10^7}, \tag{16}$$

since there are effectively 10^8 sites available for phosphorylation. However, if we assume instead that each protein is indistinguishable from any other, then the situation is identical to choosing 3×10^6 proteins to make a cell from the possible 2^{30} phosphorylation states (with replacement). The number of different cells in this case becomes (see the "Combinations with Repetition" section here: https://www.mathsisfun.com/combinatorics/combinations-permutations.html)

$$\begin{aligned} N_{\text{cells}} = \binom{2^{30}+3\times10^6-1}{3\times10^6} &\approx \frac{(10^6+3\times10^6)!}{(10^6)!(3\times10^6)!} \\ &\approx \frac{(4\times10^6)^{10^6}}{(10^6)!} \\ &\approx 10^{10^6}. \end{aligned} \tag{17}$$

Taking the geometric mean of the upper and lower limits, we obtain

$$(10^{2\times10^7}\times10^{10^6})^{1/2} \approx 10^{10^7}, \tag{18}$$

which is an "astronomically large" number.

6. The pandemic elephant in the room.

(a) What is the information density of the SARS-CoV-2 virus? What I mean is that there are a certain number of bits of information contained in the viral genome, so you can report a density of bits/nm^3.

Solution: A bit is binary (0 or 1), giving two unique options. However, RNA, such as that carried by SARS-CoV-2, contains up to four unique nucleotides (A, C, G, U). In order to ensure uniqueness to represent each nucleotide, two bits are needed (00, 01, 10, 11). For SARS-CoV-2, which has a roughly ≈ 30 kb genome, the total information is approximately 6×10^4 bits. Because the virion has a diameter of ≈ 100 nm, by modeling the virion as a sphere, we come to a volume of

$$V_{virion} = \frac{4\pi}{3}r^3 = \frac{4\pi}{3}(50\ nm^3)^3 \approx 5\times10^5\ nm^3$$

Then the information density of the virus is

$$\rho_{virion} = \frac{I}{V_{virion}} = \frac{6\times10^4\text{ bits}}{5\times10^5\text{ nm}^3} \approx 10^{-1}\ \frac{\text{bits}}{\text{nm}^3}$$

(b) What is the information density of a typical hard drive for backing up our laptops?

Solution: External drives have improved in storing more data in smaller units. Suppose we took a 1 TB external solid state drive (SSD). The Samsung T5 Portable SSD is roughly 75 mm long, 60 mm tall, and 10 mm thick, giving it a volume of roughly $\approx 5 \times 10^4$ mm^3, or 5×10^{22} nm^3 for later comparison to part (a).

We note that 1 terabyte is $\approx 10^{12}$ bytes while 1 byte is 8 bits. Thus, the information density of a SSD is

$$\rho_{SSD} \approx \frac{8 \times 10^{12} \text{ bits}}{5 \times 10^{22} \text{ nm}^3} \approx 10^{-10} \frac{\text{bits}}{\text{nm}^3}$$

Here, we see that the virion is about nine orders of magnitude more information dense than a typical SSD. Non-solid state external hard drives, which store data in a larger container, are even less information dense than an SSD, illustrating just how tightly the viral information is compacted in the virion.

(c) Given your answer to part (a), how many SARS-CoV-2 viruses would it take to capture all the information in the Library of Congress? How much volume would such a "library" take up? Could the whole Library of Congress fit into one 5 mL tube, a one L flask, one shelf of a -80 freezer?

Solution: The collection of print materials at the Library of Congress is said to fill over 800 miles of bookshelves. We will use this number as the basis for our estimation. This length corresponds roughly with 1300 km of bookshelves. Thinking back to the sizes of bookshelves at the local public library, each bookshelf has roughly eight shelves, giving $\approx 10^7$ meters of shelves.

If I think of a package of standard printer paper, I know that there are 500 sheets of paper. Furthermore, the package is roughly 5 cm in thickness. Given that there are two sides of a sheet of paper, this means that written text is ≈ 200 pages/cm. This suggests $2 \times 10^4 \frac{\text{pages}}{\text{m}} \times 10^7 \text{m} \approx 2 \times 10^{11}$ pages of written text. I estimate that words are typically more than three letters but less than eight letters long, giving me a geometric mean of 5 letters in a word. Finally, I expect more than 100 words on a page, but fewer than 1000 words, giving me ≈ 300 words on each page. Putting this all together, this suggests $2 \times 10^{11} \frac{\text{words}}{\text{page}} \times 5 \frac{\text{letters}}{\text{word}} \approx 3 \times 10^{14}$ letters. If the Latin-based alphabet (including accents applied to various non-English languages) is the predominant alphabetical system (note that the Library of Congress includes written texts across over 450 languages), then we can resort to representing each letter as 1 byte (or 8 bits). As a result, we come to $\approx 3 \times 10^{14}$ bytes (300 Tb) of information, or about 10^{15} bits of information. Considering that the Library of Congress also contains various other non-written pieces such as sheet music, maps, and photographs, this may be a lower bound of the amount of information in the Library. Nevertheless, we will press on with this rough estimate.

We know that each virion contains 6×10^4 bits of information, we see that there would have to be $\approx 2 \times 10^{10}$ virions to capture all of the information from the Library of Congress.

With each virion having a volume of $\approx 5 \times 10^5 \text{nm}^3$, this means a total volume of $\approx 10^{16} \text{nm}^3$. $1\text{mL} = 1\text{cm}^3 = 10^{21}\text{nm}^3$,so this equates to a volume of 10^5 mL,or 0.01μL,which is below the volumes of standard P2 pipettors in the lab!