PHYS 176/276 Quantitative Molecular Biology

Problem Set #1

Thursday, Jan 26, 2023

1. **Making of a bacterial cell:** An *E. coli*cell is placed in a test tube containing 5 ml sterile growth medium, consisting of 0.5% w/v glucose ($C_6H_{12}O_6$), 5mM ammonium (NH_4Cl), various inorganic ions, and kept shaken (for good aeration) in a 37°C water bath shaker. After a short transient (to be neglected here), it is observed to grow exponentially with 60 min doubling time until it runs out of either the carbon or nitrogen source. Below you are asked to make a number of *estimates*. You may only use the information provided here except for basic chemical properrties (e.g., molar mass) or otherwise specified. Please report your numerical answers as well as the mathematical expressions.

   (a) The biomass of an *E. coli*cell is approximately its dry weight, about 0.3 pg. Given the chemical composition of biomass is ∼50% carbon and ∼14% nitrogen, estimate how many molecules of glucose and ammonium are required to make a cell if all C and N consumed end up in the cellular biomass. Based on this estimate, what is the maximum cell density this tube of culture can reach? Express your answer in the unit of OD (optical density), with 1 OD∼10^9 cells/ml. From this, work out work out how many mM of glucose and ammonium does it take to synthesize 1 OD worth of biomass, and thus the rate of glucose and ammonium consumption per OD of cells. How long would it take for the tube to reach the maximum cell density starting from a single cell? from 10^6 cells/ml (a typical starting culture density)? When the culture is saturated, how much culture volume does each cell have to itself? Compare to the volume of an *E. coli*cell ($1\mu m^3$), how "crowded" is the culture at saturation?

   (b) Let us next estimate the energy cost of making the above cell. The dominant cost of biosynthesis turns out to be in protein synthesis. It takes 4 ATP molecules to extend a nascent polypeptide chain by one amino acid. Given that about half of the cells biomass is in proteins, how many ATP does it cost to synthesize all the proteins in a cell? (You may assume that half of the protein mass is in Carbon and that one amino acid contains on average 5 carbon atoms.) Given the ATP hydrolysis energy of ∼30 kJ/mol, what is the energy cost to synthesize all the proteins in one cell? of all cells in the tube? If this amount of energy is used to heat up 5 ml of water, how much would the temperature change? (You need to look up the specific heat of water.)

   (c) The ATP concentration is maintained at ∼4mM in the cell. From your answer to part (b), estimate the rate ATP is being drained to perform protein synthesis to maintain exponential growth at 60 min per doubling. How long would it take to deplete the ATP pool if it is not being supplied? Theoretically, one glucose molecule can maximally generate ∼30 ATPs using "respiration" under aerobic conditions. Estimate the minimal rate at which glucose molecules need to be supplied to sustain the energy consumption for protein synthesis for a single cell. Express this demand in unit of mM/h for a growing culture at density of 1 OD. Compare this rate to that needed to supply biomass synthesis you worked out in part (a).

(d) It turns out that *E. coli*(and many other microbes) actually generates energy from glucose inefficiently, using "fermentation" rather than respiration even in aerobic condition. For *E. coli*, the aerobic fermentation of glucose generates 12 ATP accompanied by the excretion of 2 acetic acid molecules for each glucose molecule used for energy generation. In this case, what is the rate of glucose uptake needed to supply the energy requirement for biosynthesis for 1 OD worth of cells growing exponentially at 60 min per doubling? How does this compare to the rate of glucose uptake needed for the synthesis of biomass? What is the rate of acetate accumulation in the medium? Express your answer in mM acetate/OD/h. Find the acetate concentration in the medium when growth stops. Look up the chemical properties of acetate and express the answer in volume fraction. Compare to table vinegar which is $4 \sim 8\%$ of acetate by volume.

2. **Equilibrium binding of protein with DNA:** A protein (P) can bind with a short piece of DNA sequence (S) to form a complex PS in solution. Assume that in a test tube there is a *total* amount of P and of S, with concentrations $[\mathrm{P}]_{\mathrm{tot}}$ and $[\mathrm{S}]_{\mathrm{tot}}$, respectively. In this problem, we will find how the fraction of DNA sequences bound by the protein depends on the total concentrations $[\mathrm{P}]_{\mathrm{tot}}$ and $[\mathrm{S}]_{\mathrm{tot}}$, which are variables set by the experimenter (or by the cell if the test tube is a cell).

   (a) Let the concentration of the complex formed by the protein and DNA sequence (PS) be denoted as [PS]. Justify in words why the fraction of DNA sequence occupied can be written as $f \equiv [\mathrm{PS}]/[\mathrm{S}]_{\mathrm{tot}}$.

   (b) Thermodynamically, the concentration of the complex [PS] is related to the *free* concentrations of P and S, denoted as [P] and [S], respectively, through the dissociation constant K_d, as

$$K_d = \frac{[\mathrm{P}] \cdot [\mathrm{S}]}{[\mathrm{PS}]}.$$

   Find the occupied fraction f in terms of the free concentrations [P], [S] and K_d. Note that the occupied fraction f is independent of the free sequence concentration [S]. Try to rationalize this result.

   (c) Despite the simplicity of the result in (b), it is not directly useful because the experimenter controls the total concentrations, not the free concentrations. Next, we want to find the occupied fraction f in terms of the total concentrations $[\mathrm{P}]_{\mathrm{tot}}$, $[\mathrm{S}]_{\mathrm{tot}}$ and K_d. To do so, first write down $[\mathrm{P}]_{\mathrm{tot}}$ and $[\mathrm{S}]_{\mathrm{tot}}$ in terms of [P], [S], and [PS]. Then, rewrite the definition of K_d in (b) in terms of $[\mathrm{P}]_{\mathrm{tot}}$, $[\mathrm{S}]_{\mathrm{tot}}$, and [PS], and solve for [PS]. Finally find the quotient $[\mathrm{PS}]/[\mathrm{S}]_{\mathrm{tot}}$. Note that the occupied fraction is dependent on the total sequence concentration $[\mathrm{S}]_{\mathrm{tot}}$.

   (d) Derive analytically that for $[\mathrm{P}]_{\mathrm{tot}} \gg [\mathrm{S}]_{\mathrm{tot}}$, the dependence of the occupied fraction f on $[\mathrm{P}]_{\mathrm{tot}}$ obtained in (c) becomes the same as the dependence of f on [P] obtained in (b). Explain intuitively why this result should be expected.
   (Hint: For the derivation, it will be useful to use Taylor's expansion $\sqrt{1-x} \approx 1 - x/2$ for $x \ll 1$.)

   (e) Suppose $K_d = 10$ nM. Plot the occupation fraction f against $[\mathrm{P}]_{\mathrm{tot}}$ for $[\mathrm{P}]_{\mathrm{tot}} = 1 - 100$ nM, for $[\mathrm{S}]_{\mathrm{tot}} = 1, 10, 100$ nM. Plot also the approximate form derived in (d). Comment on your numerical results in light of the analytical results derived above.

3. **Binding energy matrix:** Mnt is a dimeric transcription factor which binds to a 17bp DNA segment. The binding energy matrix $G_i(b)$ was measured by the Stormo lab and is reproduced below for the half site from position 10 to 17. The binding energies for the other half (position 1-8) can obtained as the *reverse complement* of those shown here. Position 9 is a neutral position which does not affect Mnt-DNA binding.

Table 1: Binding energy matrix

position	10	11	12	13	14	15	16	17
A	1.8	2.4	1.6	1.0	0	2.1	0.8	1.1
C	2.4	1.9	4.2	2.1	0.3	0	0	0
G	0	1.6	0	0	1.2	3.2	1.0	1.2
T	3.0	0	2.2	2.2	0.6	2.2	0.7	0.3

(The numbers are expressed in units of $k_BT \approx 0.6$ kcal/mole)

(a) Given that $G^{ns} - G^* \approx 16k_BT$, find the effective dissociation constant $\widetilde{K}^*$ for the strongest binder in the presence of genomic DNA that is $5 \cdot 10^6$ bp in length.
[Hint: approximate the genomic DNA as a random string of nucleotides with equal distribution of $\{A, C, G, T\}$.]

(b) Approximate the non-zero entries of the binding energy matrix by one parameter, ϵ, and find the smallest value of ϵ that would result in the same $\widetilde{K}^*$.

(c) The target sequence is located within the following segment of DNA

```
5' - TCTACGATCCACTGTCGACTCGACTGCCGTAT - 3'
```

Compute and plot the binding energy $G_j = \min(G_j^{sp}, G^{ns})$ as a function of the position j, the position in the sample sequence that the first position of the Mnt motif aligns to. Repeat the plot for $G^{ns} - G^* \approx 30k_BT$. Attach your computer code (or show your method if performed otherwise) and comment on your findings.

4. **Multiple target sites.**

[Those who have not had a basic course in Statistical Mechanics need not attempt this problem.]

Suppose a transcription factor can bind to N_0 distinguishable target sites with the same specific binding free energy G_0, and N additional distinguishable background sites with the non-specific binding free energy G_{ns}. Suppose that there are M TF molecules in the cell, with $N \gg \{M, N_0\}$, and assume all TFs are associated with the DNA. Derive an expression for the probability $P_A(M; N, N_0)$ that a particular target site, site A, is occupied by the following procedure.

(a) Write down the Boltzmann weight $W(M-m,m)$ that $M-m$ proteins are bound to the background sites and m proteins are bound to the target sites.

(b) Given that m proteins are bound to the N_0 target sites, what is the probability, $f(m,N_0)$, that the site A is occupied?

(c) Argue that P_A is given by

$$P_A(M;N,N_0) = \frac{\sum_{m=0}^{N_0} f(m,N_0)W(M-m,m)}{\sum_{m=0}^{N_0} W(M-m,m)}$$

Simplify this expression for $M \gg N_0$ and cast it into the form $P_A = \dfrac{1}{1+\widetilde{K}_A/M}$

[Hint: For $N \gg n$, $N!/(N-n)! \approx N^n$]

Find $\widetilde{K}_A$ in terms of the other variables introduced above. On which variable does it not depend for $M \gg N_0$?

(d) Would the dependence of $\widetilde{K}_A$ on N_0 affect the "programmability" of the binding affinity?

(e) Finally, we'll see just how large M needs to be before the above approximation becomes valid. Write the exact expression for $\widetilde{K}_A$ starting from the full form of P_A for $N_0 = 2$ and show that it is dependent on M.

For $N = 10^7$ bp, and $G^{ns} - G_0 = 16k_BT$, plot $\widetilde{K}_A$ as a function of M for $N_0 = 2$. Find the range of M where $\widetilde{K}_A$ approaches the M-independent limit.