**Quantitative Microbiology**

Problem Set #2: Transcription and Transcriptional Regulation

due: Thursday, February 9, 2023

1. **Transcription by RNA polymerase:** In this problem, we will think through quantitatively the gross numbers involved in bacterial transcription, for the synthesis of typical (mRNA-encoding) genes (parts (a)-(c)) and of genes encoding ribosomal RNA (rRNA) (parts (d), (e)) in exponentially growing cells. For concreteness, let us consider *E. coli* cells growing with a doubling time of 60 min. Please report your numerical answers as well as the mathematical expressions.

Table 1: Pertinent parameter values

| | | |
|---|---|---|
| $\varepsilon_m$ | speed for transcribing mRNA | $\sim$50 nt/sec |
| $\varepsilon_r$ | speed for transcribing rRNA | $\sim$85 nt/sec |
| $\ell_m$ | physical size of RNAp when transcribing mRNA | $\sim$55 nt |
| $\ell_r$ | physical size of RNAp when transcribing rRNA | $\sim$40 nt |
| $\tau_m$ | half-life of mRNA | $\sim$2 min |
| $\delta_m$ | mRNA degradation rate | $\ln 2/\tau$ |
| $T$ | cell doubling time | 60 min |
| $\lambda$ | dilution rate due to growth | $\sim \ln 2/T$ |
| $g$ | average gene copy number | 1.5 |
| $G_m$ | number of mRNA-encoding genes/cell | $\sim$4500 |
| $G_r$ | number of rRNA-encoding genes/cell | 7 |
| $L_m$ | average length of mRNA-encoding genes | $\sim$1000 nt |
| $L_r$ | length of rRNA-encoding gene | $\sim$4500 nt |

**(a)** The maximal speed of transcriptional elongation by RNAp is $\sim$50 nt/sec when they are transcribing mRNA-encoding genes. Given the physical size of RNAp (covering $\sim$55 nt), how many RNAp can be maximally packed on a gene (typically $\sim$1000 nt in length)? what is the maximal rate at which mRNA transcripts can be synthesized for these genes?
[Hint: think about how frequently can a new transcript be initiated given that an RNAp must move out of the promoter region before another RNAp can initiate another transcript.]

**(b)** Given the half-life of 2 min for a typical mRNA transcript, what is the maximal copy number of such a transcript in the steady state (where mRNA synthesis is balanced by mRNA degradation)? If every gene is transcribed at this rate, how many RNAp would be needed in a cell to make these mRNA?
[You may assume that there are 4500 genes in the chromosome of *E. coli*, the vast majority of which are mRNA-encoding, and there are on average 1.5 copies of chromosome per cell (for typical growing cells which are in between cell division).]

**(c)** One actually finds only $\sim$2 transcripts/cell for a typical actively transcribed mRNA-encoding gene. How frequently are these genes being transcribed if their half-lives are

as given above? If you take a snapshot, how many RNAp would you find on each such gene? Suppose one-quarter of the genes in the genome are transcribed at this rate (and the others not actively transcribed), on average how many RNAp would you expect to be engaged in transcribing these genes in a cell?

Now, consider ribosomal RNA (rRNA) synthesis. Each ribosome contains 3 pieces of rRNA (known as 5S, 16S, and 23S rRNA) which are encoded by 3 genes. For the purpose of this exercise, we will refer to them as a single "rRNA-encoding gene" (or rrn gene) with total length being ∼4500 nt. There are 7 identical copies of such rrn genes in the chromosome of *E. coli*. For simplicity, we assume again on average 1.5 copies of chromosomes per growing cell, so that there are ∼10 rrn genes per cell. The transcriptional elongation speed of RNAp on rRNA genes is faster (∼85 nt/sec) and the physical size is smaller (∼40 nt), due to dedicated mechanisms.

**(d)** What is the maximal number of rRNA the cell can synthesize in one cell-doubling? How many RNAp would it take? Given that each rRNA is incorporated into one ribosome, and that the ribosomes are stable (i.e., negligible degradation), how many ribosomes do you expect there to be in a cell if they are synthesized at the maximal rate?
`[Note: even though the ribosomes are not degraded, they are still ‘‘diluted’’ due to cell growth.]`

**(e)** There are actually 15,000 ribosomes per cell. What is the initiation rate of rRNA synthesis at each rRNA gene? what is the average separation between successive RNAp? and how many RNAp are involved in rRNA synthesis?

**(f)** Based on your answers above, how many RNAp are engaged in transcribing mRNA genes? and how many transcribing rRNA genes? What is the total amount of nucleotides in mRNA? in rRNA? what about the total flux of nucleotide towards mRNA synthesis and rRNA synthesis? Comment on your results.

2. **Regulatory function:** The input-output characteristics of transcriptional control is described by a regulatory function $\mathcal{R}$, defined as the ratio of the transcription rate in the presence of the relevant regulatory proteins to that in the absence of the regulatory proteins. Consider the following (Hill) form of the regulatory function for a promoter controlled by a single species of transcription factor A:

$$\mathcal{R} = \frac{1 + \omega q_A^n}{1 + q_A^n}$$

where $q_A = [\mathrm{A}]/K_A$, and $n$ is the Hill coefficient.

**(a)** Plot $\mathcal{R}$ vs.[A] on both linear-linear and log-log plots for $K_A = 100$ nM, $\omega = \{3, 10, 30, 100\}$, and $n = \{1, 2, 4\}$ over the range [A] =1-1000 nM. Indicate on the plots the capacity and sensitivity of the regulatory functions.

**(b)** The "sensitivity" $s$ of a promoter is defined as the maximum log-log derivative of the regulatory function. Find an expression for $s$ in terms of the parameters $\omega$, and $n$. Plot $s(\omega)/n$ vs. $\ln\omega$ for $\omega$ =1-1000. Comment on the results.

3. **Transcriptional AND-gate:** As discussed in class, a gene $g$ is regulated by two TFs, TF1 and TF2, binding to operators O1 and O2 respectively. O2 is located between O1 and the core promoter. The effective dissociation constants of the operators and RNAP are $K_1$, $K_2$ and $K_p$. TF1 and TF2 bound to O1 and O2 interact with each other with a cooperativity factor $\omega_{12}$ and TF2 bound to O2 interacts with promoter-bound RNAP with a cooperativity factor $\omega_{2p}$. There is no interaction between an activator bound to O1 and the RNAP.

   *Below, we study which parameter values would provide an AND-like response.*

   **(a)** Write down the truth table for a Boolean AND gate. Describe how the promoter of $g$ could act as a transcriptional AND gate.

   **(b)** Write down the expression for the probability $\mathcal{P}$ of promoter occupation by RNAP assuming that the thermodynamic approximation holds. Assuming further that $q_p = [\text{RNAP}]/K_p \ll 1$ and $\mathcal{P} \ll 1$, derive the regulatory function $\mathcal{R}$ and write it in terms of $q_1 = [\text{TF1}]/K_1$, $q_2 = [\text{TF2}]/K_2$, $\omega_{12}$ and $\omega_{2p}$.

   **(c)** Along the line [TF1] = [TF2], $\mathcal{R}$ increases from 1 (at [TF1] = [TF2] = 0) to some maximum value $\omega$. To have a good AND gate, this transition must be steep. Show that a Hill coefficient of 2 can be achieved if $\omega_{12} \gg \omega_{2p} \gg 1$. I.e., along the diagonal [TF1] = [TF2], $\mathcal{R}$ can be written in the form

$$\mathcal{R} \approx \frac{1 + \omega\left([\text{TF1}]/K\right)^2}{1 + \left([\text{TF1}]/K\right)^2}$$

   with $K = K_1\sqrt{r/\omega_{12}}$, provided that $r = K_2/K_1$ is in the range $1 \lesssim r \lesssim \omega_{2p}$.

   [Hint: Generally, the numerator and denominator of $\mathcal{R}$ are polynomials of the form $1 + ax + bx^2$. The Hill form $1 + bx^2$ is approximately obtained if the linear term $ax$ is small compared to 1 and to $bx^2$ for the entire range of $x$. This algebraic exercise can be skipped by those from the more biological background.]

   **(d)** What should $K$ be if we want the transition to be at [TF1] = [TF2] $\approx$ 300 nM?

   **(e)** If TF2 is present at high concentrations ($\sim$ 1000 nM) and [TF1] = 0, the expression of the transcriptional AND gate should be low ($\mathcal{R}$ small). What is the best choice for $r$ given the constraints demonstrated in (c)?

   **(f)** Assume $\omega_{12} = 150$ and $\omega_{2p} = 20$; use the values you just selected for $r$ and $K$ to compute the corresponding $K_1$ and $K_2$.

4. **Regulation of the *lac* promoter:** In this problem, you are asked to estimate the parameters describing the control of the *lac* promoters of *E. coli*, by analyzing the data of Oehler et al. (EMBO J 13: 3348-3355, 1994).

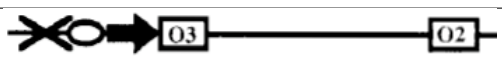
**(a)** The first set of data generated in the experiment of Oehler et al. was for mutants where the operators at positions O2 and O3 are both mutated completely (crosses in first column of Table below).

Fold repression, defined as the ratio of gene expression without and with the Lac repressor, was measured for two different levels of Lac tetramers ([R] = 50 nM and [R] = 900 nM) for 3 strains of cells whose operator site at position O1 bears the sequence of the operators O1, O2, O3, respectively. (The CRP site is indicated by the ellipse. For this problem you may ignore its effect.)

| Promoter | [R] = 50 nM | [R] = 900 nM |
|---|---|---|
|   | 200 | 4700 |
|   | 21 | 320 |
|   | 1.3 | 16 |

Find the effective in vivo dissociation constants $K_1$, $K_2$ and $K_3$ for the 3 operator sites O1, O2 and O3 respectively. Compare to the value of $K_1 = 1$ pM found by von Hippel's *in* vitro experiment.

**(b)** For the second set of data, the operator at position O3 is removed, while the operators at positions O1 and O2 are replaced by various combinations of operators as shown below. For each of these strains, fold-repression was measured for the two concentrations of repressors:

| Promoter | [R] = 50 nM | [R] = 900 nM |
|---|---|---|
|   | 2300 | 6800 |
|   | 360 | 560 |
|   | 6.8 | 15 |

Now, looping can be an issue. Using the above two sets of data to find the best estimate of the loop parameter $C_{L12}$ (defined as in the lecture).

**(c)** Using the values of $K_1$, $K_2$, $K_3$, and $C_{L12}$ estimated above, plot the repression factor as a function of [R] on a log-log plot. Draw the data in the table in part (b) on the same plot, thereby seeing how the prediction of the thermodynamic model with DNA looping compares with the observed data quantitatively. Explain what differ in the 3 regions you see in the log-log plot. Under physiological conditions *in vivo*, the Lac repressor is expressed at $\sim$10 tetramers per cell. Which region is the natural system in and what is the expected fold-repression in conditions with no or saturating amount of IPTG (inducer of Lac repressor)?