Bi/BE/CS 183 2022-2023
Instructor: Lior Pachter
TAs: Tara Chari, Meichen Fang, Zitong (Jerry) Wang

Problem Set 9

Submit your solutions as a single PDF file via Canvas by **8am Friday March 10th**.

- If writing up problems by hand, please use a pen and not a pencil, as it is difficult to read scanned submission of pencil work. Typed solutions are preferred.
- For problems that require coding, Colab notebooks will be provided. Please copy and save the shared notebook and edit your own copy, which you should then submit by including a clickable link in your submitted homework. Prior to submission make sure that you code runs from beginning to end without any error reports.

Problem 1 (50 points)

In this question, you will investigate a set of simple, nontrivial models of gene regulation, encompassing the initiation of transcription and production of mRNA molecules. You will derive the steady state dynamics for a model of bursty transcription, which happens to result in a negative binomial distribution of RNA in living cells. 'Bursty' transcriptional dynamics in particular are common in bacterial and mammalian systems, outlined in [**?**], where RNA molecules are transcribed in distinct bursts, separated by periods of time (see Fig. 1). The distribution of molecules produced in each burst was additionally shown to be geometric (see Fig. 2).

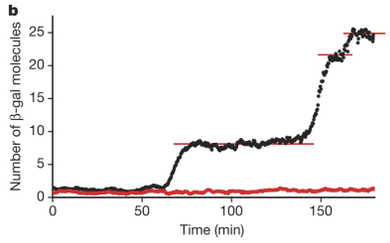

Figure 1: Number of β-galactosidase molecules produced, from the β-gal gene in a single cell. Red line denotes a blank background. See also BE150 Bursty Expression

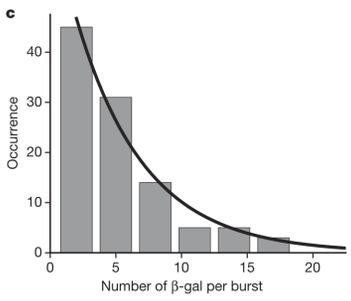

Figure 2: Number of β-galactosidase molecules produced per burst.

To begin the modeling process, we will first define the behavior of a gene that randomly switches between two promoter states: active state A and inactive state I. This is known as a telegraph model.

$$\begin{aligned} I &\xrightarrow{k_{on}} A, \\ A &\xrightarrow{k_{off}} I. \end{aligned} \tag{1}$$

The promoter transitions from state I to state A with rate k_{on}, and from state A back to state I with rate k_{off}. We are interested in the probability of the promoter being state A and I at steady state (i.e. when $dP_I/dt = dP_A/dt = 0$). Recall from Lecture **17** that this process is governed by a matrix $\mathcal{A}$, such that $\partial_t P = \mathcal{A}P$, or:

$$\frac{d}{dt}\begin{bmatrix} P_I(t) \\ P_A(t) \end{bmatrix} = \begin{bmatrix} -k_{on} & k_{off} \\ k_{on} & -k_{off} \end{bmatrix}\begin{bmatrix} P_I(t) \\ P_A(t) \end{bmatrix} \tag{2}$$

(a) (10 points) Set the LHS of Equation 2 to zero and solve for the steady-state probabilities P_I and P_A in terms of k_{on} and k_{off}. Note that $P_I + P_A = 1$.

Now we will first define the more general stochastic transcription model. We append transcription and degradation reactions to the telegraph system in Equation 1:

$$\begin{aligned} I &\xrightarrow{k_{on}} A, \\ A &\xrightarrow{k_{off}} I \\ A &\xrightarrow{k_A} A + \mathcal{T} \\ \mathcal{T} &\xrightarrow{\gamma} \varnothing \end{aligned} \tag{3}$$

where $\mathcal{T}$ denotes a single mRNA molecule. Note we assume transcription only occurs in the active state. **For a particular duration of the active state τ, the number of mRNA molecules transcribed is distributed according to a Poisson distribution with rate parameter τk_A.**

(b) (15 points) Find the distribution for mRNA burst size by integrating over all possible durations of the active state $\tau \in (0, \infty)$. Namely, show that the probability of x transcripts produced while the gene promoter is active is equal to $\alpha_x = (1-p)p^x$ where $p = \frac{k_A}{k_A + k_{off}}$, meaning that the mRNA burst size takes on a geometric distribution. (Hint: the residence time of every state of a Markov chain is exponentially distributed).

If now we assume the transcription rates are high but the active state of the promoter is short-lived such that multiple mRNA molecules can be produced instantaneously (taking $k_{off}, k_A \to \infty$ while keeping the mean of the mRNA burst distribution finite), we approach the 'bursty' limit. We can write down the chemical master equation that governs the number of RNA n as:

$$\partial_t P(n,t) = k_{on}\left[\sum_{x=0}^{n} \alpha_x P(n-x,t) - P(n,t)\right] + \gamma\left[(n+1)P(n+1,t) - nP(n,t)\right], \tag{4}$$

where α_z is the probability of a burst containing z mRNA molecules part. Our goal is to demonstrate that the limiting distribution $P(n)$ is negative binomial. In principle, we can set the LHS of Equation 4 to zero, plug in the negative binomial distribution for $P(n)$, and show that the equality holds for some parameter values. Unfortunately, the NB distribution has unwieldy combinatorial factors, and this approach rapidly becomes intractable. Instead, we can use the *probability generating function* (PGF):

$$G(z,t) := \sum_{n=0}^{\infty} P(n,t) z^n, \tag{5}$$

where $z \in \mathbb{C}$. By performing the summation over all terms of Equation 4, we can convert it to a partial differential equation (PDE):

$$\frac{\partial G}{\partial t} = k_{on} \left[F(z) - 1\right] G + \gamma \left[1 - z\right] \frac{\partial G}{\partial z}, \tag{6}$$

where $F(z)$ is the PGF of the mRNA burst size distribution.

(c) (10 points) Write down the PGF of the burst distribution $F(z)$ by plugging the result from (b) into the definition in Equation 5, and calculating the sum.

Now, we can use the PGF of the NB distribution:

$$G(z) = \left(\frac{1}{1 - \theta(z-1)}\right)^r \tag{7}$$

Now we will find *which* values of θ and r give us the correct solution.

(d) (15 points) Plug in $F(z)$ from (c) and $G(z)$ from Equation 7 into Equation 6, and set its LHS to zero. Solve the resulting equation and calculate θ and r in terms of k_{on}, γ, and b (where b is the mean of the mRNA burst size).

Problem 2 (50 points)

In this question, we will compare our results to stochastic simulations and see how well the various approximations work. The core of the simulation code is already implemented. It remains for you to write the *propensity functions* that determine the instantaneous rates of the reactions. This amounts to writing down the individual contributions to the efflux rates in the matrix $\mathcal{A}$. For example, consider the constitutive production system with transcription rate k_A, degradation rate γ, and n molecules of RNA i.e. there is no inactive state. At each instant, it will assign a propensity of k to the transcription reaction and γn to the degradation reaction.

You will also investigate the inverse problem, of fitting parameters given data from a stochastic system. You will use a Markov Chain Monte Carlo approach to approximate parameters for the steady state distribution of a system, given data points sampled from the system.

Problem 2 notebook

Your edited version of the notebook *must be submitted* for this problem. Reminder to check that your notebook runs all the way through with the the `Runtime` → `Restart` and `Runtime` → `Run All` commands.