Bi/BE/CS 183 2021-2022
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 **10am Friday March 11th**.

- 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.

  See the class Colab Tutorial on how to produce a shareable link for your notebook.

Problem 1 (35 points)

In this question, we will investigate one of the simplest nontrivial models of gene regulation. First, we consider the behavior of a gene that randomly switches between two promoter states: active state A and inactive state I.

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

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) (3 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 $P_I + P_A = 1$.

Most genes in bacteria and mammals demonstrate *bursty* RNA transcription: there are intermittent periods of high transcription, separated by long periods of no production at all. This means state A is infrequent, but has a high transcription rate k_A. To understand the dynamics of this condition, 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. This is also known as constitutive transcription model. 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) (12 points) Find the distribution of 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 $p = \frac{k_A}{k_A + k_{off}}$, and conclude 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 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 can write down the chemical master equation that governs the number of RNA n as:

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

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. These parameter values will give us the necessary scaling. Unfortunately, the NB distribution has unwieldy combinatorial factors, and this approach rapidly becomes inelegant. Instead, we can get rid of the combinatorial factors by treating the *probability generating function* (PGF):

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

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

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

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

(c) (5 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:

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

It remains to figure out *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).

Congratulations! You have now demonstrated that there are at least two mechanisms that can lead to a negative binomial distribution of RNA in living cells: "extrinsic" cellular heterogeneity and "intrinsic" bursting transcriptional dynamics.

Problem 2 (45 points)

In this question, we will apply the CME to solve a simple model of protein translation. To analyze it properly, we should write down a multi-species CME that accounts for (potentially bursty) RNA transcription and degradation, as well as protein translation and degradation. This turns out to be intractable even for the simplest, constitutive model:

$$\begin{aligned} \varnothing &\xrightarrow{k} \mathcal{T} \\ \mathcal{T} &\xrightarrow{\gamma} \varnothing \\ \mathcal{T} &\xrightarrow{k_t} \mathcal{T} + \mathcal{P} \\ \mathcal{P} &\xrightarrow{\gamma_p} \varnothing, \end{aligned} \tag{8}$$

Fortunately, we can make approximations. Proteins have a much longer half-life than RNA, so the variation in RNA levels has a relatively small effect on the relevant timescale. As a zeroth-order approximation, we can write down the following *mean-field* approximation:

$$\begin{aligned} \varnothing &\xrightarrow{k_t\langle T\rangle} \mathcal{P} \\ \mathcal{P} &\xrightarrow{\gamma_p} \varnothing, \end{aligned} \tag{9}$$

where k_t is the translation rate per molecule of RNA, $\langle T\rangle$ is the average number of RNA transcripts, and γ_p is the protein degradation rate. This is mathematically identical to the constitutive transcription model, with birth rate $k_t\langle T\rangle$ and degradation rate γ_p.

The distribution of the protein counts thus follows a Poisson distribution with mean $k_t\langle T\rangle/\gamma_p$, $Poisson(k_t\langle T\rangle/\gamma_p)$. This *discrete* distribution is somewhat unwieldy to evaluate when protein copy numbers are very high, often the case in living cells. To simplify evaluation, we use *continuous* distributions that effectively approximate the protein PMF.

(a) (4 points) As the mean $k_t\langle T\rangle/\gamma_p := \alpha$ increases, what Gaussian distribution does the discrete PMF converge to?

Even if the assumption that RNA is produced and degraded much faster than proteins holds, the approximations are not equally effective. The constitutive Poisson distribution is best for the *microscopic* regime, where protein numbers are low. The distribution in (a) is best for the *macroscopic* regime, where proteins numbers are high. We can fill in the range between these extremes, the *mesoscopic* regime, which works fairly well when protein abundance is moderate.

From the derivation in lecture, we can write down the following stochastic differential equation (SDE) that tracks the number of proteins N_t:

$$dN_t = (k_t\langle T\rangle - \gamma_p N_t)dt + \begin{bmatrix}\sqrt{k_t\langle T\rangle} & -\sqrt{\gamma_p N_t}\end{bmatrix}\begin{bmatrix}dW_t^{(1)} \\ dW_t^{(2)}\end{bmatrix}, \tag{10}$$

where $dW_t^{(1)}$ and $dW_t^{(2)}$ are two independent processes that account for the randomly firing reactions. Note that the SDE has a deterministic term, which tracks the mean of the protein number; the random term essentially quantifies fluctuations about this mean.

The SDE in Equation 10 is in the standard drift-diffusion form:

$$dN_t = \mu(N_t)dt + \boldsymbol{\sigma}(N_t)d\mathbf{W}_t, \tag{11}$$

where $\boldsymbol{\sigma}$ is the *diffusion coefficient*, here the 1×2 matrix in Equation 10. We would like to find the distribution of proteins in the mesoscopic regime, which is given by the law of N_t as $t \to \infty$. The easiest way to do this is by converting the SDE into a Fokker-Planck equation, which tracks the evolution of *probability densities* $p(x)$ rather than values N_t.

The FPE has the following form:

$$\partial_t p(x,t) = -\partial_x[\mu(x)p(x)] + \partial_x^2[D(x)p(x)], \text{ such that} \tag{12}$$

$$D(x) = \frac{1}{2}\boldsymbol{\sigma}\boldsymbol{\sigma}^T. \tag{13}$$

In our case, D(x) is a scalar function.

(b) (9 points) The variable $x \in (0, \infty)$ represents the amount of protein molecules. Write down $\mu(x)$ and $D(x)$ in terms of their functional dependence on current state x, based on Equation 10 and 11.

(c) (14 points) Write $\partial_x \ln p(x)$ as a function of α and x. Plug the solution to (b) into Equation 12. Since we are concerned with the steady state, we can start with $0 = -[\mu p] + \partial_x[Dp]$ instead of the full second-order FPE (assuming $\lim_{x\to+\infty} p(x) = 0$).

(d) (9 points) Integrate $\partial_x \ln p(x)$ to obtain $p(x)$ as $Cf(x)$, where C is a normalization constant you do not need to compute.

(e) (9 points) We have obtained an unnormalized probability distribution. Instead of doing another integral, we can exploit the fact that $p(x)$ is closely related to the gamma distribution:

$$f_\Gamma(x) \propto (x - c)^{a-1}e^{-(x-c)/b}, \tag{14}$$

where a is the *shape* parameter, b is the *scale* parameter, and c is the *location* parameter. Rewrite the $f(x)$ you have obtained in (d) in the same form as Equation 14. Calculate a, b, and c that yield $f(x)$.

Congratulations! You have developed a non-Gaussian mesoscopic approximation to the constitutive transcription process, and written it in a form that can be calculated by any numerical package with gamma distributions.

Problem 3 (20 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. At each instant, it will assign a propensity of k to the transcription reaction and γn to the degradation reaction.

The Problem notebook is here.

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.