Problem set 5: Finite number fluctuations and random walks, models with continuous time

BIO-369

Prof. Anne-Florence Bitbol
EPFL

Thanks to Benjamin Martin for proposing Problems 1-3.

In this problem set, we will first derive the Gillespie algorithm. Then, we will use it to simulate the simple gene expression model studied in the lecture. We will also elaborate on how to analyze and model experimental data regarding gene expression.

1 Gillespie algorithm

The aim of this problem is to introduce the Gillespie algorithm, an exact algorithm allowing to simulate a process where independent reactions occur randomly with given rates [1]. This algorithm is extremely useful to simulate chemical systems where reactions occur, taking into account finite size effects. Its scope is general, and for instance it can also be employed to simulate microbial populations (we could use it to simulate a birth-death process similar to the Moran process but with continuous time instead of discrete time).

a) Consider an event that happens randomly with constant rate k. Consider a small time interval δt such that at most one event happens during it. What is the probability that one event occurs during one such time interval? What is the name of the probability distribution describing the number of events occurring during one such time interval? Assume that one event occurred between 0 and δt. Explain why this does not affect the probability that one event occurs between δt and $2\delta t$.

b) Denoting by $N(t)$ the number of events that happen between time 0 and time t, explain why the probability distribution of $N(t)$ is the Poisson distribution with mean kt. For this, you can cut the time interval $[0, t]$ in small time intervals of duration δt and use the result of the previous question.

c) Starting at time 0, let T denote the first time (after time 0) when the event happens. Note that T is a random variable because the event happens at random times. Consider $t > 0$. Explain why the probability $P(T > t)$ that T is larger than t is equal to the probability $P(N(t) = 0)$. Give the expression of $P(T > t)$.

d) What is the probability density $p(t)$ of the first time when the event happens? Recall that, by definition, $p(t)$ is such that $P(t < T \leq t + dt) = p(t)\, dt$. What is the name of this probability distribution?

   Here, we just demonstrated a fundamental point of the Gillespie algorithm: at each time point, the time at which the next event happens can just be drawn in the distribution found in the last question.

e) Now assume that two different and independent events 1 and 2 can happen, one with rate k_1 and the other with rate k_2. What is the probability that one event (whatever its type) occurs in a small time interval δt? How should we draw the time at which the next event, *whatever its type*, occurs?

f) Now we know how to figure out *when* the next event happens. How should we decide *which* of the two possible events (1 or 2) is the one that happens at this time? How can this be implemented in a simulation, starting from drawing a random number between 0 and 1 in a uniform distribution?

g) How would you simulate the process and the trajectory of the system (showing the state of the system versus time), performing events one after the other?

2 Simple gene expression model: simulation and data analysis

First, assume the following gene expression system:

$$\begin{aligned} \emptyset &\xrightarrow{k_{RNA}} RNA \\ RNA &\xrightarrow{k_{decay}} \emptyset \end{aligned}$$

a) Write the master equation on the probability $P_n(t)$ to be have n RNA molecules (with $n = 0, 1, 2, \dots$) at time t for this model. Give the expression of the distribution P_n satisfying this equation at steady state. What are the mean, the variance and the Fano factor of this distribution?

b) Using Python, write a Gillespie algorithm to simulate the previous system. You can initialize the number of RNAs to 0 and use $k_{RNA} = 17.370\, molecules.h^{-1}$ and $k_{decay} = 0.1368\, h^{-1}$. You can simulate the system for 150 hours. Use the algorithm to generate ten trajectories, and plot them.

c) Modify your previous algorithm so that it only returns the number of RNA molecules at the final timepoint t_f (using the same parameters as previously, i.e $k_{RNA} = 17.370\, molecules.h^{-1}$, $k_{decay} = 0.1368\, h^{-1}$ and $t_f = 150\, h$). Plot the histogram of the number of RNA molecules at the final timepoint, using at least 500 trajectories. Compute the mean, the variance and the Fano factor of these results. Compare with your theoretical predictions from question 1a.

d) In a recent paper Desai et al. [2] inferred the dynamics of gene expression for a given gene (Nanog) using single-molecule fluorescence *in-situ* hybridization (smFISH), a method allowing to quantify the number of a target mRNA in single cells. The data they obtained are given in `smFISH_data.csv`. You can use the following command to import the data as a `numpy array`: `data = np.genfromtxt("smFISH_data.csv", skip_header=1, delimiter="\n")`.

Using the data, plot the histogram of the number of RNA molecules at equilibrium. Compute the average, variance and Fano factor. Are the results consistent with the previous model (questions 1a-c)?

3 Modeling gene expression systems beyond the Poisson model

A widely used model for gene expression in mammalian cells is the random telegraph model [3], where the gene can be in two distinct states, namely ON and OFF. The set of reactions can be written as follows:

$$\begin{aligned} OFF &\xrightarrow{k_{ON}} ON \\ ON &\xrightarrow{k_{OFF}} OFF \\ ON &\xrightarrow{k_{RNA}} ON + RNA \\ RNA &\xrightarrow{k_{decay}} \emptyset \end{aligned}$$

a) Write down the master equations for this system. For this, you will need to take into account the fact that the state of the systems needs to be described by two variables. The first one is the gene state (OFF or ON), which you can denote by a binary variable i, with $i = 0$ meaning OFF and $i = 1$ meaning ON. The second one is n, the number of RNA molecules. Thus, you can enumerate the possible transitions from state $(0, n)$, leading to an equation on $P_{0,n}$, and then do the same for $P_{1,n}$, which will yield two different equations for each n.

Let us introduce

$$\nu_{ON} \doteq \frac{k_{ON}}{k_{ON} + k_{OFF}}\,. \tag{1}$$

Let us admit (see [3] for a proof) that, at equilibrium, the average number of RNA molecules $\langle RNA \rangle$, and the Fano factor FF of the system are given by:

$$\begin{aligned} \langle RNA \rangle &= \nu_{ON} \times \frac{k_{RNA}}{k_{decay}}, \\ FF &= 1 + \frac{(1-\nu_{ON}) \times k_{RNA}}{k_{ON} + k_{OFF} + k_{decay}}. \end{aligned}$$

b) How can the previous equation for the average number of RNA molecules be interpreted?

Hint: first think about the interpretation of ν_{ON}, and then use question 1 a to interpret k_{RNA}/k_{decay}.

c) In Python, write a Gillespie algorithm to simulate the system. To do so, you can modify the algorithm from question 1c to incorporate the new states and transitions.

Desai et al. [2] used maximum likelihood estimation (we will study this method later in this class) to infer the parameters $k_{ON}, k_{OFF}, k_{RNA}, k_{decay}$ from the data, assuming the random telegraph model. The inferred parameter values are given in Table 1.

Parameter	Value
k_{ON}	0.572
k_{OFF}	0.758
k_{RNA}	40.389
k_{decay}	0.1368

Table 1: Kinetic rates values inferred from data using maximum likelihood estimation [2]. All the kinetic rates are given in h^{-1} or in $molecules.h^{-1}$.

d) Simulate the system using these parameter values and, as previously, plot the histogram of the number of RNA molecules at the final timepoint. Here, take $t_f = 75\,h$. Also plot the experimental data on the same graph. Compute the average, variance and Fano factor. How does it compare to the experimental data?

e) Setting $k_{ON} \gg k_{OFF}$, what do you expect? Simulate and conclude.

Hint: you can compare the random telegraph model with $k_{ON} \gg k_{OFF}$ to a Poisson model with the same average, and use 2a.

4 Additional problem: Population dynamics

This problem was previously given as part of the final exam of this class.

In this problem, we are going to consider a simple model that describes the dynamics of a population of bacteria. Let n be the number of bacteria in our population. All bacteria are assumed to be identical. Each bacteria can reproduce with division rate b and die with rate d. Let $P_n(t)$ be the probability that n bacteria exist at time t. Consider a very small time interval δt, such that at most one event occurs during this time.

a) During δt, how can the number n vary? List all possibilities, giving the probability that each of them happens.

b) What happens if n hits 0?

c) What algorithm would you propose to use to simulate this model? You do not need to describe it in detail.

d) What are the similarities and differences between this model and the simplest model of gene expression where the protein synthesis reaction has rate k and each protein has a decay rate d?

e) Assuming $n \geq 1$, write down a differential equation on $P_n(t)$. Justify. How is this type of equation called?

f) Write down a differential equation on $P_0(t)$. Justify.

g) At steady state, what is P_1 equal to? What about P_n for $n > 1$? And P_0? Given this, what do you expect to happen to the population at long times?

h) Show that the average number $\langle n \rangle$ of bacteria satisfies the equation

$$\frac{d\langle n \rangle}{dt} = (b - d)\langle n \rangle \,. \tag{2}$$

i) Solve Eq. 2 to obtain the time evolution of $\langle n \rangle$. From this result, what do you expect to happen to the population at long times? Compare to your result from question g) and discuss.

j) What is the equation satisfied by n if it is very large? Compare it to Eq. 2 and comment.

k) What changes if n is not very large compared to the case where it is very large?

l) Instead of taking a constant division rate b, another possibility is to assume a division rate written as $b(1 - n/K)$, where K is a constant. The death rate d remains constant. What is the equation satisfied by n if it is very large? What would n be equal to at steady state? Comment on the effect of this modification of the division rate.

References

[1] Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, December 1977. Publisher: American Chemical Society.

[2] Ravi V. Desai, Xinyue Chen, Benjamin Martin, Sonali Chaturvedi, Dong Woo Hwang, Weihan Li, Chen Yu, Sheng Ding, Matt Thomson, Robert H. Singer, Robert A. Coleman, Maike M. K. Hansen, and Leor S. Weinberger. A DNA repair pathway can regulate transcriptional noise to promote cell fate transitions. *Science*, 373(6557):eabc6506. Publisher: American Association for the Advancement of Science.

[3] J. Peccoud and B. Ycart. Markovian Modeling of Gene-Product Synthesis. *Theoretical Population Biology*, 48(2):222–234, October 1995.