Traffics jams on RNA
Dynamical modelling of living systems 7.5hp

Lucas Hedström* and Ludvig Lizana†
IceLab, Umeå University

INTRODUCTION

Proteins in the cell are produced by means of translation. Free-floating ribosomes within the cell volume attach to sequences of mRNA, which are then transcribed by matching each codon pair along the mRNA to a corresponding amino acid (AA), forming a protein. Even though translation is complex, we can gain understanding using simple models imitating traffic flows along roads can reproduce non-trivial results.

One important aspect is the study of burstiness. In a standard poisson process every event happens uniformly over time. For a bursty process, events occur intermittently during shorter time periods. This is a behaviour that has been measured in living cells [1]. One way to quantify burstiness is by using the Fano factor, which if larger than one usually indicates a bursty process [2].

In this lab, you will simulate protein production by ribosomes. Initially you will start with a very simple stochastic model where each mRNA is assumed to produce one protein before decaying, then you will simulate protein production from multiple mRNAs using a traffic model, where the traffic speed is determined by the sequence of codons along the mRNA.

Lab report

You will write a lab report for this lab that, without including figure captions, should consist of a maximum of 1500 words (roughly 3–5 pages). You are allowed one figure panel per task, but each figure panel can consist of more than one figure. Make sure that the figures are well formatted in a vector format with clear axis labels, legends and proper font size.

TASKS

Task 1: Simple protein production

To get a baseline model of 'burstiness' we will first simulate a simple stochastic model where we simply assume that each mRNA produces x number of proteins in their lifetime.

An mRNA is produced every 10 minutes which produces, on average, one protein. Proteins decay every 30 minutes. We can ignore the mRNA and only consider that a protein is produced and decayed with certain rates, which are

$$k_p^{\text{mRNA}} = \frac{1}{600}\ [s^{-1}], \quad k_d = \frac{1}{1800} N_p\ [s^{-1}]$$

where N_p is the number of proteins. *Ponder: Why is* $k_d \sim N_p$*?*

Goals

- Express the protein production as an ODE and solve for the expected number of proteins.
- Stochastically simulate the system for a reasonable time frame and plot the number of proteins versus time. *Hint: The probability of an event occuring is the rate times the timestep.* Calculate mean number of proteins and the Fano factor $= \frac{\sigma^2}{\mu}$, where σ^2 and μ is the variance and mean of the number of proteins over time. Discuss the results.
- Repeat the same two tasks, but but this time let each mRNA produce on average 10 proteins. Compare the results and discuss.

Plotting

For this first task, we'd like to take a second to explain how we want the plot to be. The most common mistake that students make is **not combining plots**. So, in this task, we want to see 1(!) plot where you plot the number of proteins versus time with two lines representing the systems where a different number of proteins is created for each mRNA.

For the remainder of this course, **we expect you to always try to combine plots in this manner**.

Task 2: The traffic model

In order to account for varying rates of attachment, movement, detachment and a limited ribosome count we will extend our simulations to a traffic model. In this model, the mRNA acts as our road and the ribosomes as our vehicles moving along this road, combining amino-acids to create new proteins after leaving the road.

At the moment we focus only on one mRNA with a zero decay rate. We let this mRNA be L codons long. We define

(i) α $[s^{-1}]$ and β $[s^{-1}]$ as the rates for a ribosome to attach at the start and detach at the end

(ii) The ribosomes move from codon i to $i+1$ with a rate q_i $[s^{-1}]$.

Note that two ribosomes cannot be on the same codon, so they cannot pass each other. To get good results in the following tasks, you should let the mRNA be at least ~ 30 codons long.

Goals

- Write a stochastic simulation to simulate the traffic model. *Hint: For each time step, randomly go through all ribosomes and the start site and let them act based on the rates. If you go through them sequentially, there might be som preferential behaviour due to the non-symmetric system.* Calculate the Fano factor and plot the number of proteins created versus time using the same protein decay rate as before (one protein is created every time a ribosome detaches from the end). **Important:** Make sure that you do not include the initial transient when you calculated the Fano factor.

- The protein current is defined as $J = \beta\rho_L$, where ρ_L is the occupational probability at the last site. *Ponder: How can you calculate an occupational probability with discrete time steps?* Using $q_i = \alpha = 1$ plot the current versus the mean-field results for $\alpha \in [0, 1]$ with both $\beta = 0.5$ and 0.25 [3]. Discuss the different mean-field regimes and how they relate to your model.

Task 3: Burstiness in the traffic model

The traffic model introduces some interesting variations when changing β and α, but is not much to talk about when looking at the protein production over time. Now we're going to introduce a lot more complexity to our model.

In our cells, we do not have an infinite amount of ribosomes. Rather, the mRNAs have to share a finite number of ribosomes that can actively translate sequences. Consider a number of free ribosomes N_r which are not translating any mRNA. We redefine the system as

(i) ribosomes enter with a rate αN_r

(ii) when the ribosome exit as β, they are returned to the free bulk

Goals

- Extend your model to consider a finite number of ribosomes. Let $\alpha = 0.9$, $\beta = 2$ with an initial pool of free ribosomes $N_r = 4$. Simulate, calculate the Fano factor and plot the number of proteins against time. **Important:** At this point, make sure that you use a low enough timestep as to not have probabilities much higher that 50%.

As we discussed in the first task, an important aspect in the system is the production and decay of mRNA. We will now add it to our system as

(i) the system starts with no mRNA. At every timestep, one mRNA is produced/decayed with the rates k_p^{mRNA} and k_d^{mRNA}.

(ii) when an mRNA decays, all of the ribosomes on it will return to the free bulk

(iii) the initial pool of free ribosomes is 4 times the average number of mRNA. *Ponder: Can you calculate this from the rates?*

Goals cont.

- Extend your model so that you can have multiple mRNA which are all translated in parallel. Let $k_p^{\text{mRNA}} = \frac{1}{600}$ $[s^{-1}]$ and $k_d^{\text{mRNA}} = \frac{1}{300} N_{\text{mRNA}}$ $[s^{-1}]$, where N_{mRNA} is the number of mRNA. *Hint: This goal is very programming language dependent. In an object-oriented language like Python you could define an mRNA class that has its own sequence of codons that ribosomes slide along. If classes are not available, you can use structs to represent mRNA.* Simulate using the same α and β as before. Calculate the Fano factor and plot the number of proteins against time. Compare against the last goal.

Task 4: Optimal codons

Up until now we've considered a constant translation rate $q_i = 1$. However, this is not true in real life. In fact, the ribosomes translate certain codons faster than other [4]. Furthermore, certain codons code for the same amino acids. This means that a sequence can produce the same protein but with different translation rates, dependent on the sequence of codons. We will investigate this for this last lab.

For this task you need a sequence consisting of ~ 30 codons. If you do not want to find your own sequence, use the following 30-codon sequence

AGCGCGCGGUCACAACGUUACUGUUAUCGAUCCGGUCGAAAAACU
GCUGGCAGUGGGGCAUUACCUCGAAUCUACCGUCGAUAUUGCUGA

or pick one from RegulonDB and convert it to RNA[1] [5].

On canvas you will find three JSON-files

(i) `rates_from_codon.json` — A dictionary with the codons as keys and rates as values.

(ii) `aminoacid_from_codon.json` — A dictionary with the codons as keys and amino acids as values.

(iii) `aminoacid_codon_groups.json` — A dictionary with the amino acids as keys and lists of codons as values.

You can use these files in your code to convert the codons to rates, but also to randomise a codon to another codon that produces the same amino acid. `json`-files are relatively easy to parse in most programming languages. *e.g.*, use `jsondecode` in MATLAB or the package `json` in Python. If you want to define the rates yourself, you can get all relevant info from the work by Mitarai et. al. [4].

Goals

- Using the given sequence (or your own), randomize the codons as to produce the same protein (the same amino-acid sequence). For the original sequence, and the randomized sequences, simulate using the same model as before. For each sequence, calculate the protein current J. Plot a histogram of J with the J for the original sequence clearly marked. What do you observe? Discuss the benefit/cost of having a slow/fast sequence of codons.

- Based on the simulations, pick the top 25% of systems that have the highest J, and the bottom 25%. Calculate and plot the mean q_i for the top and bottom performers, along with the original sequence. Plot these q_i versus position (1–L). *Hint: You should get three lines, corresponding to the top and bottom 25% performers, along with the original sequence. The x-axis should be from 1–L with the corresponding q_i value on the y-axis.* Also plot a linear fit of J versus $\mathbb{E}(q)$. What do you observe?

- From the last excersises you should have a clear understanding of what affects J. However, this is not what we see in cells. Could you give an explanation to why? How could you adjust the model to take this into account? Motivate why this would more accurately capture the dynamics in the cell.

* lucas.hedstrom@umu.se
† ludvig.lizana@umu.se

[1] M. Dobrzyński and F. J. Bruggeman, Elongation dynamics shape bursty transcription and translation, Proceedings of the National Academy of Sciences **106**, 2583 (2009).
[2] U. Fano, Ionization yield of radiations. ii. the fluctuations of the number of ions, Phys. Rev. **72**, 26 (1947).
[3] B. Derrida, E. Domany, and D. Mukamel, An exact solution of a one-dimensional asymmetric exclusion model with open boundaries, Journal of statistical physics **69**, 667 (1992).
[4] N. Mitarai and S. Pedersen, Control of ribosome traffic by position-dependent choice of synonymous codons, Physical biology **10**, 056011 (2013).
[5] A. Santos-Zavaleta, H. Salgado, S. Gama-Castro, M. Sánchez-Pérez, L. Gómez-Romero, D. Ledezma-Tejeida, J. S. García-Sotelo, K. Alquicira-Hernández, L. J. Muñiz-Rascado, P. Peña-Loredo, *et al.*, Regulondb v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in e. coli k-12, Nucleic acids research **47**, D212 (2019).

[1] Make sure there's no stop codons except at the end.