Cell type Analysis from scRna-seq Data Achieved from a Mixture MOdel (CARDAMOM): vignette

Elias Ventre[1,2,3,4,†] and Ulysse Herbach[5, ‡]

[1] *Laboratoire de Biologie et Modélisation de la Cellule, École Normale Supérieure de Lyon, CNRS, UMR 5239, Inserm, U1293, Université Claude Bernard Lyon 1, 46 allée d'Italie F-69364 Lyon, France*
[2] *Inria Center Grenoble Rhône-Alpes, Équipe Dracula, Villeurbanne, France*
[3] *Univ Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5208, Institut Camille Jordan, Villeurbanne, France*
[4] *Current address: The University of British Columbia, Vancouver, Canada*
[5] *Université de Lorraine, CNRS, Inria, IECL, F-54000 Nancy, France*
[†] *eventre@math.ubc.ca*
[‡] *ulysse.herbach@inria.fr*

CARDAMOM (Cell type Analysis from scRna-seq Data achieved from a Mixture MOdel), is an algorithm for inferring a GRN from timestamped scRNA-seq data, which crucially exploits the notions of metastability and transcriptional bursting. It has been developed in Ventre. [1], and benchmarked and applied to real data in Ventre et ql.[2]. In the latter, it has been shown that combined with the simulation algorithm HARISSA [3], the method allows to calibrate a mechanistic model in order to reproduce accurately experimental timestamped scRNA-seq data.

The aim of this vignette is to present the main functions of the method CARDAMOM, as well as to detail the small improvements that have been performed in Ventre, Herbach et al. [2] with respect to the first publication [1]. The actual package is available at the following address https://github.com/eliasventre/cardamom, and the previous version at https://gitbio.ens-lyon.fr/eventr01/cardamom.

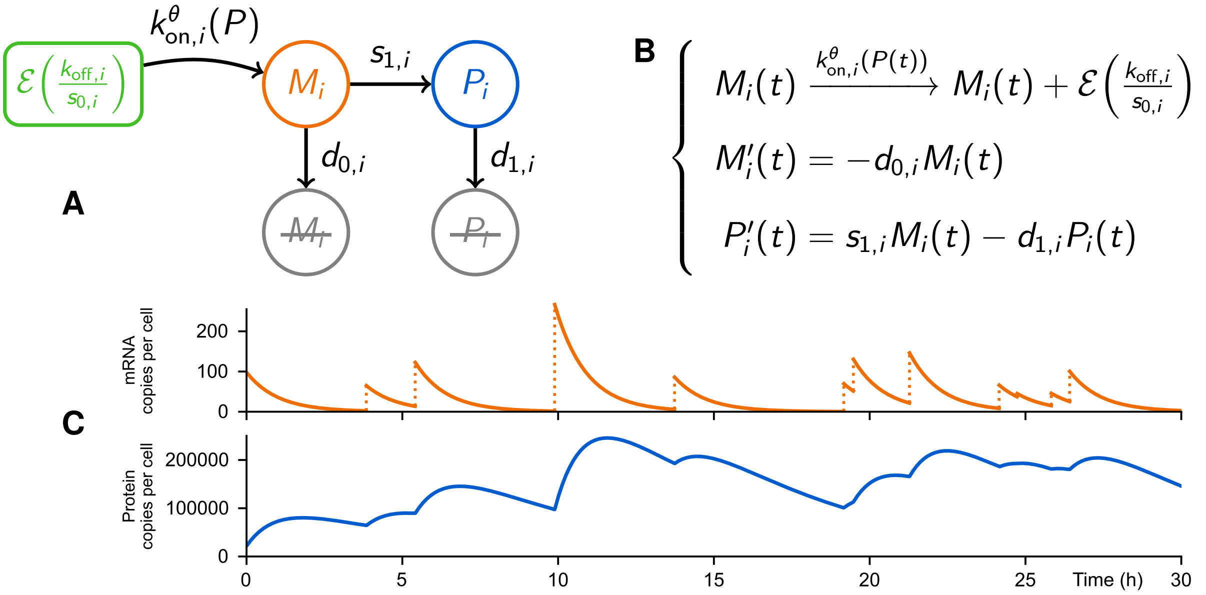
Brief description of the model

This paragraph can be found on the Method section of [2].
The model used throughout this article is based on a hybrid version of the well-established two-state model of gene expression [4], where a gene is described by the state of a promoter, which can be either *on* or *off*. If the promoter is *on*, mRNAs are being transcribed at a rate s_0, which are then translated into proteins at a rate s_1. Degradation of both mRNAs and proteins occurs at a rate d_0 and d_1, respectively. The transitions between the on and off states occur at times of rates k_{on} and k_{off}. We consider the *bursty* regime of this model ($k_{\mathrm{on}} \ll k_{\mathrm{off}}$), corresponding to short active periods with high transcription rates, as experimentally observed [5–8]. In this regime, mRNA is then transcribed by bursts of tens to hundreds of molecules. The random times at which these bursts occur are still described by an exponential distribution of parameter k_{on}, and their random size by an exponential distribution with mean s_0/k_{off}. This model is compatible with experimental single-cell data, as steady-state mRNA levels follow for each gene a Gamma distribution, in line with continuous single-cell data [9].

The key idea is to incorporate this model into a network: the burst rate for each gene i is given by a gene-specific function $k^{\theta}_{\mathrm{on},i}(P)$, where P is the vector of protein quantities (Figure 1). This function depends on proteins through a GRN, represented by an n-by-n matrix $\theta = (\theta_{ij})$ where n is the number of genes in the network. The value of $k^{\theta}_{\mathrm{on},i}(P)$ then corresponds to the transcriptional burst frequency of gene i given protein levels P. Each parameter θ_{ij} encodes the interaction $j \to i$ with its direction, sign, and intensity. Recent work suggests that burst sizes are smaller and more uniform than previously anticipated [7] therefore leaving more room for burst frequency modulation [10] as a mechanism for gene expression regulation. We therefore consider that interactions come mainly from the modulation of burst frequencies $k^{\theta}_{\mathrm{on},i}$ and that for any gene i, the rates $k_{\mathrm{off},i}$ do not depend on P. The burst frequencies can be represented by sigmoid functions [3] as a simplification of the mechanistic form used in [4, 11]:

$$k^{\theta}_{\mathrm{on},i}(P) = k_{0,i} + (k_{1,i} - k_{0,i}) \left(1 + \exp\left(-\beta_i - \sum_{j=1}^{n} \theta_{ij} P_j\right)\right)^{-1} \tag{1}$$

where $k_{0,i}$ (resp. $k_{1,i}$) is the minimal (resp. maximal) burst frequency of gene i and β_i is the basal activity of gene i, which can be also considered as the constant activity of a set of genes that are not measured but act on the network.



$$\begin{cases} M_i(t) \xrightarrow{k^{\theta}_{\mathrm{on},i}(P(t))} M_i(t) + \mathcal{E}\left(\frac{k_{\mathrm{off},i}}{s_{0,i}}\right) \\ M_i'(t) = -d_{0,i} M_i(t) \\ P_i'(t) = s_{1,i} M_i(t) - d_{1,i} P_i(t) \end{cases}$$

Figure 1: Figure S1 from [2]. Graphical (**A**) and mathematical (**B**) descriptions of the mechanistic model for the dynamics of a gene i. (**C**) Bursts of mRNA occur at random times with rate $k_{\mathrm{on},i}$ and their size follows an exponential distribution $\mathcal{E}(k_{\mathrm{off},i}/s_{0,i})$. The variables M_i and P_i describe respectively the mRNA and protein quantities associated to gene i in the cell. The vector of protein levels is denoted by $P = (P_1, \cdots, P_n)$ while θ denotes the GRN which couples the genes together through functions $k_{\mathrm{on},i}$.

Description of the model parameters

We provide here a list of the main variables of the mechanistic model. *It can be found on the Appendix A. of [1].*

- θ a matrix defining the interactions between genes, corresponding to a matrix with diagonal terms defining external stimuli,

- $k_{0,i}$ is the basal rate of expression of gene i,
- $k_{1,i}$ is the maximal rate of expression of gene i,
- β_i is the basal activity of gene i, which can be also considered as the constant activity of set of genes which are not measured and act on the network,
- $d_{0,i}$ is the degradation rates for mRNAs of gene i,
- $d_{1,i}$ is the degradation rate for proteins of gene i,
- $s_{0,i}$ is the creation rate for mRNAs of gene i,
- $s_{1,i}$ is the creation rate for proteins of gene i,
- $k_{off,i}$ is the exponential rate of switching from state **on** to state **off** for the promoter of gene i,
- $k_{on,i}$ is a sigmoidal function depending on the global protein field $P \in \Omega$, defined in (1), which characterizes the exponential rate of switching from state **off** to state **on** for the promoter of gene i,
- $c_i = \frac{k_{off,i}}{s_{0,i}}$ is the exponential rate at which mRNAs of gene i are created at every burst.

The following table precise the way these parameters are encoded in an object of the class NetworkModel in the package CARDAMOM.

Parameters	Variable in the object NetworkModel
k_0	self.a[0]
k_1	self.a[1]
c	self.a[2]
d_0	self.d[0]
d_1	self.d[1]
θ	self.inter
β	self.basal

Note that the creation rates s are not directly implemented because s_0 is contained in c and s_1 can always be fixed to $S1 = s_1 := cd_0d_1/k_1$ which is the value for which the proteins scale to 1 (see Ventre. [1] Section 4.2). As the proteins are not observed in practice, this does not change the mRNAs counts obtained by simulating the model.

Description of CARDAMOM

The principle of CARDAMOM is based on a two-step procedure:

- In a first step, we find the set of parameters defining the mixture of negative binomial distributions that best fits the data. The only parameter that can vary over time is the mixture proportion parameter μ_t (allowing to estimate, for each gene i, the mean burst size $s_{0,i}/k_{\text{off},i}$ and the values $k_{0,i}$ and $k_{1,i}$ of the typical modes associated to the function $k_{\text{on},i}$). Note that all model parameters are estimated except the degradation rates, which are constant for each gene and scale the dynamics of gene expression.

- In a second step, we calibrate the basal activity and interaction parameters β_i and θ_{ij} in order to approximate this mixture distribution. The interaction parameters θ_{ij} are then updated at each timepoint sequentially to match the mixture parameters.

Organisation of CARDAMOM

The functions of CARDAMOM are splitted into two directories:

- The first one is located at **cardamom/model** and contains the file **base.py**, that contains itself all the functions associated to an object of the class NetworkModel for calibrating the mechanistic model. The scRNA-seq data are given as an input to the function **fit**.

- The second one is located at **cardamom/inference** and contains the files **kinetics.py** and **network.py**. **kinetics.py** contains the functions allowing to perform the first step of CARDAMOM (finding the mixture parameters), and **network.py** contains the functions allowing to perform the second step of CARDAMOM (finding the associated network).

The function **fit** then uses the two main functions of the second directory, which are **infer_kinetics** and **inference_optim**.

- **infer_kinetics**: This function infer the parameters of a mixture of Negative Binomial. In the first version of CARDAMOM [1], it was achieved using a MCMC algorithm on all the data observed at each timepoint grouped together. Following a simpler idea developed in [3], this function now simply fits the parameters of a Negative Binomial on the data observed at each timepoint and for each gene separately, using a variational method. For every gene, we consider that the rates $k_{0,i}$ and $k_{1,i}$ correspond respectively to the minimal and maximal values of these inferred parameters, following the assumption that every gene reaches its maximal frequency during one of the observed timepoints.

- **inference_optim**: Thus function returns the inferred network (basal + inter), as well as the time at which each edge has been detected with the strongest intensity, as detailed in [1], Section 4. We have slightly modified the way each we penalized each coefficient, introducing weighting between the diagonal coefficients and the rest. This has been calibrated by cross-validation on a set of reference networks, and should not be changed for new networks.

References

[1] E. Ventre. “Reverse engineering of a mechanistic model of gene expression using metastability and temporal dynamics”. In: *In Silico Biology* 14 (2021), pp. 89–113.

[2] Elias Ventre, Ulysse Herbach, Thibault Espinasse, Gérard Benoit, and Olivier Gandrillon. “One model fits all: combining inference and simulation of gene regulatory networks”. In: *bioRxiv* (2022).

[3] U. Herbach. “Gene regulatory network inference from single-cell data using a self-consistent proteomic field”. In: *arXiv* 2109.14888 (2021), pp. 1–21.

[4] U. Herbach, A. Bonnaffoux, T. Espinasse, and O. Gandrillon. “Inferring gene regulatory networks from single-cell data: a mechanistic approach”. In: *BMC Systems Biology* 11 (2017), pp. 1–15.

[5] D. M Suter, N. Molina, D. Gatfield, K. Schneider, U. Schibler, and F. Naef. “Mammalian genes are transcribed with widely different bursting kinetics”. In: *Science* 332.6028 (2011), pp. 472–474.

[6] D. Nicolas, N. E. Phillips, and F. Naef. “What shapes eukaryotic transcriptional bursting?” In: *Mol Biosyst* 13 (2017), pp. 1280–1290.

[7] J. Rodriguez and D. R. Larson. “Transcription in Living Cells: Molecular Mechanisms of Bursting”. In: *Annu Rev Biochem* 89 (2020), pp. 189–212.

[8] E. Tunnacliffe and J. R. Chubb. “What Is a Transcriptional Burst?” In: *Trends Genet* 36 (2020), pp. 288–297.

[9] C. Albayrak, C. A. Jordi, C. Zechner, J. Lin, C. A. Bichsel, M. Khammash, and S. Tay. “Digital Quantification of Proteins and mRNA in Single Mammalian Cells”. In: *Molecular Cell* 61 (2016), pp. 914–924.

[10] C. Li, F. Cesbron, M. Oehler, M. Brunner, and T. Hofer. “Frequency Modulation of Transcriptional Bursting Enables Sensitive and Rapid Gene Regulation”. In: *Cell Syst* 6.4 (2018), pp. 409–423.

[11] A. Bonnaffoux, U. Herbach, A. Richard, A. Guillemin, S. Gonin-Giraud, P.-A. Gros, and O. Gandrillon. “WASABI: a dynamic iterative framework for gene regulatory network inference”. In: *BMC Bioinformatics* 20 (2019), pp. 1–19.