Parallelized single-cell perturbation experiments: improving parameter inference for stochastic reaction networks using iterative likelihood evaluation

Davidović Anđela[1] and Ruess Jakob[2]

[1] Hub of bioinformatics and biostatistics, Institut Pasteur, Paris, France

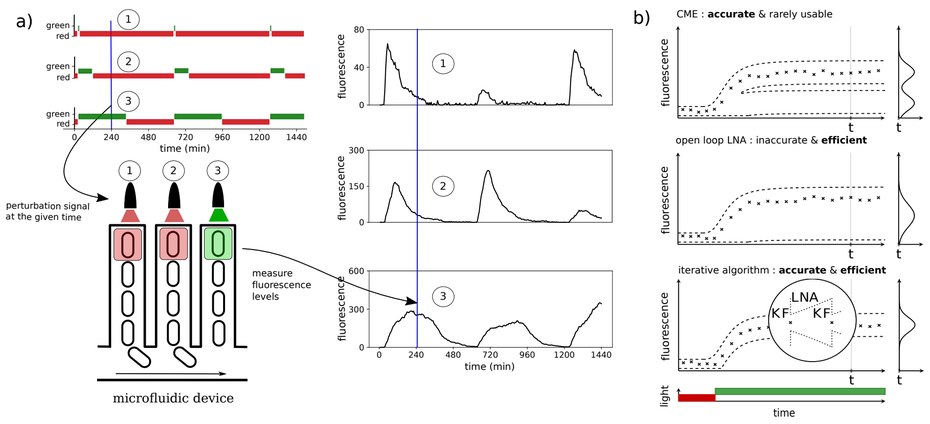
[2] Lifeware team, Inria Saclay, France

andela.davidovic@pasteur.fr jakob.ruess@inria.fr

Background

The biochemical reactions inside single cells are inherently stochastic. They are modelled with stochastic kinetic models governed by the chemical master equation (CME). Learning parameters of these models requires solving CME which is rarely possible, hence we need to deploy methods to approximate its solution. The traditional methods to infer parameters of stochastic kinetic models from single-cell longitudinal data have generally been developed under the assumption that experimental data is sparse. Using them on datasets with many measurement time points for each cell may lead to a large computational cost.

Modern microscopy and new questions raised



1. What is more informative: to diversify input perturbations or not? Do we need many cells per input or many inputs?
2. Which method to use to calculate likelihoods and infer parameters of stochastic kinetic models from data sets in which each cell receives a different input perturbation?
3. Computational efficiency of parameter inference methods: do they scale with the number of inputs and the number of measurement times?
4. Are there approaches that are particularly well-suited for novel data sets?

Algorithm

The observed data for a single cell is given by $\boldsymbol{y} = \{y_i \mid i = 1, ..., M\}$.
The measurement model,

$$y_i = \boldsymbol{C}\boldsymbol{X}(t_i) + \xi_i, \quad i = 1, \ldots, M,$$

where the $\xi_i \sim \mathcal{N}(0, \sigma^2)$ are independent technical measurement errors, the matrix $\boldsymbol{C}$ maps the full system state to the measured output species.

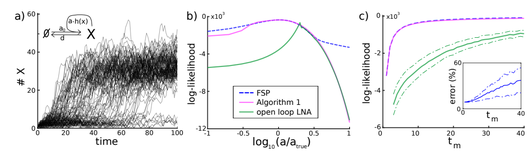
Likelihood in terms of transition probabilities:

$$\begin{aligned} p(y_1, \ldots, y_M) &= p(y_1) \cdot p(y_2 \mid y_1) \cdots p(y_M \mid y_{M-1}, \ldots, y_1), \\ p(y_i \mid y_{i-1}, \ldots y_1) &= \int\int \underbrace{p(y_i \mid \boldsymbol{x}_i)}_{\text{tech. noise}} \cdot \underbrace{p(\boldsymbol{x}_i \mid \boldsymbol{x}_{i-1})}_{\text{trans. prob.}} \cdot \underbrace{p(\boldsymbol{x}_{i-1} \mid y_{i-1}, \ldots, y_1)}_{\text{state posterior}} d\boldsymbol{x}_i d\boldsymbol{x}_{i-1}. \end{aligned}$$

1. Calculate approximate moments, $\eta^1_{\boldsymbol{x}_1}, \eta^2_{\boldsymbol{x}_1}$, of $p(\boldsymbol{x}_1)$ by moment closure.
2. Approximate the true $p(\boldsymbol{x}_1)$ by a Gaussian distribution that has $\eta^1_{\boldsymbol{x}_1}, \eta^2_{\boldsymbol{x}_1}$ as moments.
3. Marginal likelihood $p(y_1) = \int p(y_1 \mid \boldsymbol{x}_1) p(\boldsymbol{x}_1) d\boldsymbol{x}_1$ is Gaussian and it can be calculated from $\eta^1_{\boldsymbol{x}_1}, \eta^2_{\boldsymbol{x}_1}$ and σ. Evaluate $p(y_1)$ and store it for the likelihood calculation.
4. The state posterior $p(\boldsymbol{x}_1 \mid y_1)$ is also a Gaussian distribution that can be calculated from $p(\boldsymbol{x}_1)$ and $p(y_1 \mid \boldsymbol{x}_1)$ thanks to Bayes' theorem, as in Kalman filtering.
5. Extract the moments, $\eta^1_{\boldsymbol{x}_1|y_1}, \eta^2_{\boldsymbol{x}_1|y_1}$, of $p(\boldsymbol{x}_1 \mid y_1)$.
6. Solve moment equations over t_m time units (i.e. over $[t_1, t_2]$) using $\eta^1_{\boldsymbol{x}_1|y_1}, \eta^2_{\boldsymbol{x}_1|y_1}$ as initial conditions in order to obtain moments $\eta^1_{\boldsymbol{x}_2|y_1}, \eta^2_{\boldsymbol{x}_2|y_1}$ that approximate the moments of the distribution $p(\boldsymbol{x}_2 \mid y_1)$.
7. Iterate: $p(\boldsymbol{x}_2 \mid y_1)$ is approximated by a Gaussian equivalently to $p(\boldsymbol{x}_1)$ in the step 2. and so forth.

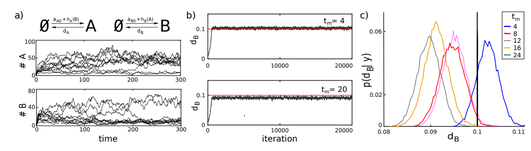
Toy models

$$\emptyset \xrightarrow{a_0 + a \cdot h(X)} X \xrightarrow{d} \emptyset$$



Accurate approximations: the approximation of the CME only needs to be accurate over the time that passes in between subsequent measurement steps.

$$\emptyset \xrightarrow{a_{A0} + h_A(B)} A \qquad A \xrightarrow{d_A} \emptyset$$
$$\emptyset \xrightarrow{a_{B0} + h_B(A)} B \qquad B \xrightarrow{d_B} \emptyset$$



Genetic toggle switch: MCMC algorithm generally converges very quickly to the vicinity of the true value of d_B.

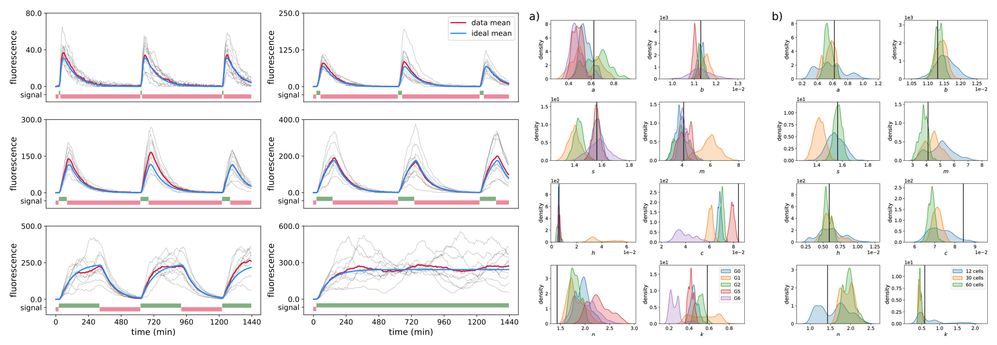
Experiments parallelised at the cell scale

Optogenetic system:

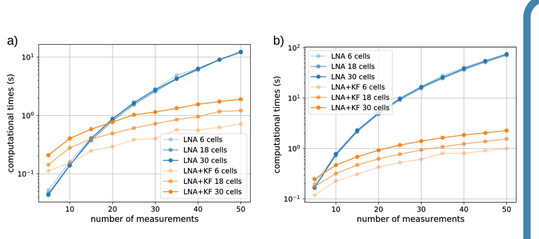
$$\emptyset \xrightarrow{h_0} E \xrightarrow{h} \emptyset, \quad \emptyset \xrightarrow{aE(t)L(t)} F \xrightarrow{b} \emptyset, \quad L(t) = \frac{(c \cdot l(t))^n}{(c \cdot l(t))^n + k^n}, \quad \frac{dl(t)}{dt} = u(t) - c \cdot l(t),$$

the input light signal $u(t) = 1$ for green light, and $u(t) = 0$ for red light.

Parameters to infer: $\boldsymbol{\theta} = \{a, b, s, m, h, c, n, k\}$, where $m := \frac{h_0}{h}$ is the mean of E.



- The quality of parameter estimates depends strongly on the group of cells that is used for inference.
- Diversified light signals in cells lead to tight posterior distributions for all parameters and MAP estimates that are close to the true values of the parameters.

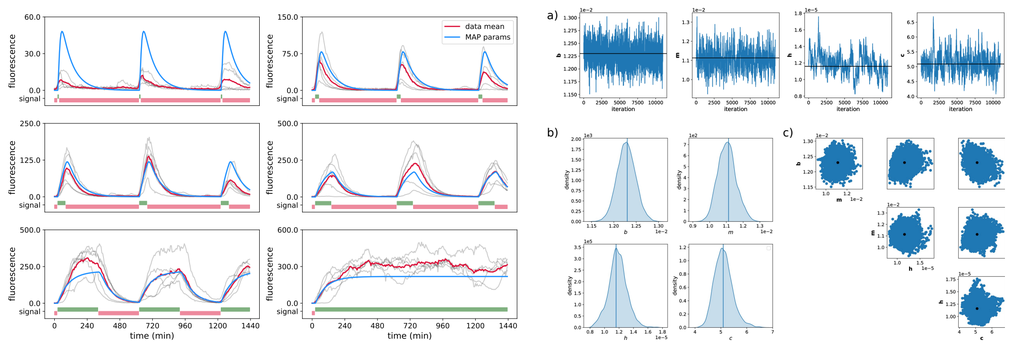


a) all cells received the same light input
b) cells have received diversified light inputs.

Computational efficiency:

- The algorithm outperforms the open loop likelihood computation starting from a small number of measurements.
- Linear vs exponential growth in computational cost.

Inference of parameters from experimental data.



Reference Davidović et al. PLoS Computational Biology. 2022.