A Statistical Learning Algorithm for Inferring Reaction Systems from Data Time Series

Julien Martinelli, Jeremy Grignard, François Fages, Sylvain Soliman

May 27th, 2019

Mechanistic Model Learning for explainable AI

The Machine Learning field provides tools to analyze time series data and yield predictions. Classical examples are Recurrent Neural Networks.

- While these predictions can be accurate, they do not come with an interpretation
- We say that the model is **Black Box**

On the contrary, Mechanistic Model Learning aims at achieving the same predictive results while being **explainable**

(XAI : Explainable Artificial Intelligence)

Focus : chemical reaction network (CRN) inference

Input : time series data from multiple traces describing evolution of molecular species

Output :

- CRN structure
- CRN kinetics

The learned model provides an understanding of the underlying processes involving the species while allowing predictions

Some attempts at Mechanistic Model Learning

- DREAM3 (2008) - Network Inference Challenge
- Logic programming
  - Prior knowledge on network's structure
  - Learn boolean function acting on species

  Boolean Network Identification from Perturbation Time Series Data combining Dynamics Abstraction and Logic Programming. L. Pauleve et al.
- Evolutionary Algorithms
  - Given number of reactions
  - Fitness to observed transitions

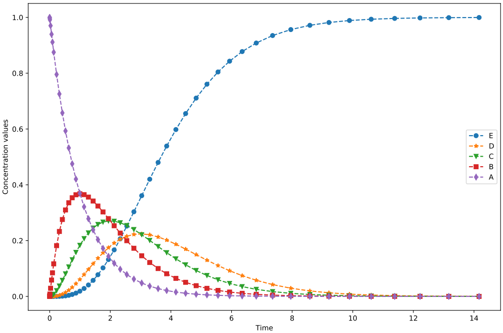
  Inferring Reaction Networks using Perturbation Data. H. Sauro et al.
- TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach
  - Information theory framework
  - Detect dependencies between genes at different time delays

  P. Zoppoli et al.

Learning parameters : well-understood
Learning the structure : remains difficult without prior knowledge.

Chain CRN learning example

On a chain of 4 reactions with mass action law kinetics, our algorithm is able to reconstruct the CRN from a single simulation trace.



Hidden CRN	Learned CRN
$A \xrightarrow{1} B$	$A \xrightarrow{1.07} B$
$B \xrightarrow{1} C$	$B \xrightarrow{1.09} C$
$C \xrightarrow{1} D$	$C \xrightarrow{1.04} D$
$D \xrightarrow{1} E$	$D \xrightarrow{0.99} E$

Application on Real data

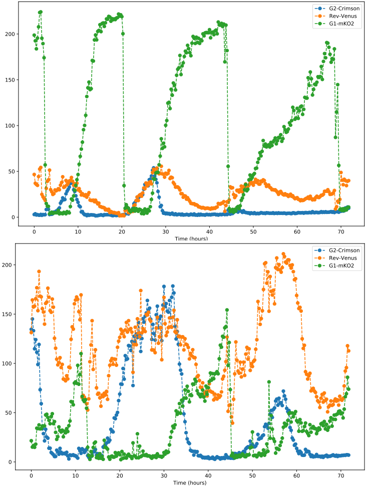
We apply the algorithm to real data and search for mechanistic models.

- NIH3T3 embryonic mouse fibroblasts left to proliferate in regular medium supplemented with 20% FBS concentration
- Time lapse videomicroscopy, one image taken every 15 minutes during 72 hours
- *Cell tracking* using three different fluorescent markers of the circadian clock and the cell cycle :
  - *Reverb-*α clock gene reporter
  - Fluorescence Ubiquitination Cell Cycle Indicators, *Cdt1* and *Geminin*, two cell cycle proteins which accumulate during the G1 and S/G2/M phases respectively.

Plot of two traces from the dataset

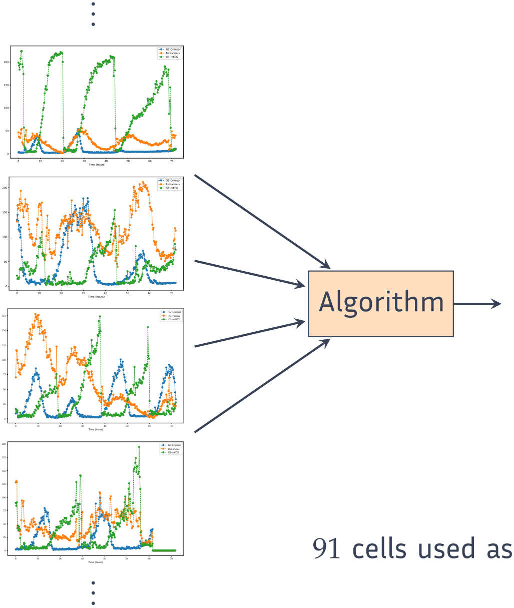
Videomicroscopy

Feillet Delaunay INSERM 2013



The cells display a high variability

Results



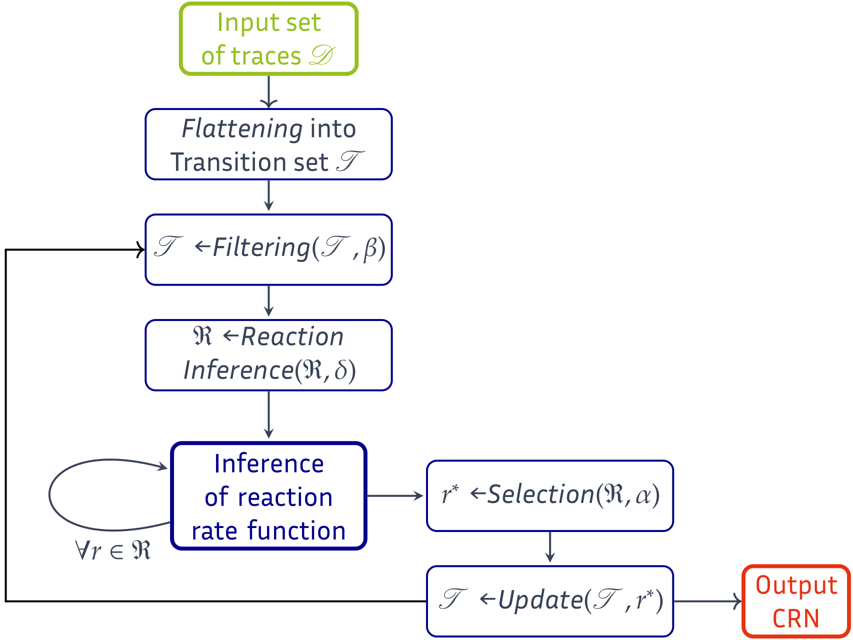
Learned CRN	Rate Functions
$G1 \implies G2$	$7.1 \dfrac{G1}{G1 + 3.68}$
$RevErb\alpha \implies G1$	$22.56 \dfrac{RevErb\alpha}{RevErb\alpha + 71.45}$
$G1 \implies \varnothing$	$5.96 \dfrac{G1}{G1 + 5.0}$
$G2 \implies \varnothing$	$54.84 \dfrac{G2}{G2 + 176.23}$

91 cells used as input $\implies \approx 18000$ data points

Contributions

- A statistical learning algorithm to iteratively infer reactions from time series data
- Infer reaction structures that maximise the pairing between reactant consumption and product formation
- Infer reaction rates that minimize the standard deviation between the observed kinetics and the inferred kinetics

Flowchart of the algorithm



Input set of traces $\mathscr{D}$
Flattening into Transition set $\mathscr{T}$
$\mathscr{T} \leftarrow$ *Filtering*$(\mathscr{T}, \beta)$
$\Re \leftarrow$ *Reaction Inference*$(\Re, \delta)$
Inference of reaction rate function
$\forall r \in \Re$
$r^* \leftarrow$ *Selection*$(\Re, \alpha)$
$\mathscr{T} \leftarrow$ *Update*$(\mathscr{T}, r^*)$
Output CRN

Table of Contents

1. Introduction
2. Application to videomicroscopy data
3. Preliminaries on CRN and time-series data
4. Algorithm and results on synthetic data

Chemical Reaction Networks

Let $S = \{1, \dots, n\}$ be the set of n molecular species. Species can also be noted with simple capital letters like A, B, C instead of by their index.

Definition

A *reaction* over S is a triple (R, P, f), where R is a multiset of *reactants* in S, P is a multiset of *products* in S and $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a *rate function* over molecular concentrations.

A *catalyst* of the reaction is a species $i \in R \bigcap P$.

A *Chemical Reaction Network* (CRN) is a finite set of reactions.

Example

$(\{A\}, \{B\}, k \cdot [A])$ also written $A \overset{k}{\Longrightarrow} B$ is the case of mass action law.

CRN classification

Definition

A *reactant-parallel* CRN is a CRN in which any two reactions do not share the same reactant (catalysts aside).

A *product-parallel* CRN is a CRN in which any two reactions do not share the same product (catalysts aside).

A *parallel* CRN is a CRN in which any two reactions do not share the same species (catalysts aside).

The chain CRN $A \implies B \implies C \implies D \implies E$ is both reactant-parallel and product-parallel but not parallel.

Differential semantics

A CRN can be interpreted in different manners, in a hierarchy of continuous differential, stochastic, discrete and Boolean semantics.

Here we consider the continuous interpretation by ordinary differential equations

$$\forall s \in S, \frac{ds}{dt} = \sum_{(R,P,f)\in\mathscr{R}} f.(P(s) - R(s))$$

Traces

Definition

A state vector is a vector $x \in \mathbb{R}^{n+1}$ where x_0 represents the real time, and x_i the concentration of species i

A *trace*, or time-series data is a finite sequence $(x(1), \dots, x(d))$ of state vectors at increasing times, i.e. $x_0(1) < \cdots < x_0(d)$.

Such traces can be :

- simulation traces, e.g. numerical integration or stochastic simulation
- experimental traces, e.g. time lapse videomicroscopy

Hypotheses

- We only study reactions with stoichiometry at most 1 : the multisets R and P are actually sets of $\mathscr{P}(S)$.

- We also restrict ourselves to the following common rate functions
  - mass action law kinetics
  - Michaelis-Menten kinetics
  - Hill of order 4 kinetics.

Flattening and Filtering

The set of traces $\mathscr{D}$ is flattened into a set of transitions $\mathscr{T}$:

$$\mathscr{T} \leftarrow \{(x^j(t), x^j(t+1) - x^j(t), j) \mid 1 \leqslant j \leqslant l,\ 1 \leqslant t \leqslant d_j - 1\}$$

A filterering step is applied on $\mathscr{T}$. For each species $i \in S$ and $\forall (x, d, j) \in \mathscr{T}$

$$\text{if } \left|\frac{d_i}{d_0}\right| < \beta . \max_{1 \leqslant t < d_j} \left|\frac{x_i^j(t+1) - x_i^j(t)}{x_0^j(t+1) - x_0^j(t)}\right| \text{ then } d_i \leftarrow 0$$

Collecting possible reactions

For $(x, d, j) \in \mathscr{T}$, let $s^* = \underset{s}{\operatorname{argmax}} |d_s|$.

Let $I := \{i \in S \text{ s.t. } |d_{s^*}| \leqslant (1+\delta)|d_i|\}$. $\forall j \in I$

- $d_j < 0 \implies j$ is a reactant of the reaction.
- $d_j > 0 \implies j$ is a product of the reaction.

Example

$I = \{u, v, s^*\}$ with $d_{s^*} > 0, d_u > 0$ and $d_v < 0$ gives $\mathbf{v} \implies \mathbf{s}^* + \mathbf{u}$

This is done $\forall (x, d, j) \in \mathscr{T}$ and leads to a set of reactions $\mathscr{R}$

Mass action law kinetics rate function inference

Once a reaction $r = (R, P)$ has been inferred :

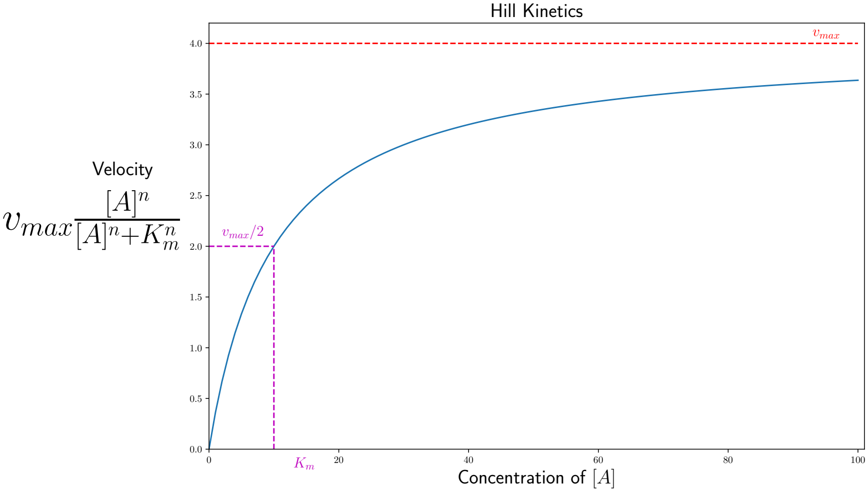
- Mass action law kinetics is computed as inferred kinetics
- the ratio between inferred kinetics and observed kinetics is measured $\forall (x, d, j) \in \mathscr{T}$ s.t. $x_i > 0, \forall i \in R$

For a reaction $A \overset{k}{\Longrightarrow} B$, this ratio reads $\epsilon = k\frac{[A]}{\frac{dA}{dt}} = k\frac{[A]}{\frac{dB}{dt}}$

$K = \left\{ \frac{x_A}{\frac{d_A}{d_0}}, (x, d, j) \in \mathscr{T} \text{ s.t. } x_A > 0 \right\}$ and $k = \left| \frac{1}{mean(K)} \right|$ so that $\epsilon = 1$.

Moreover, we set $\sigma = std(K)$ to be the error criterion on the reaction

Hill rate function computation



$$\frac{dA}{dt} = v_{max}\frac{[A]^n}{[A]^n + K_m^n} \xrightarrow{[A]\to+\infty} v_{max}$$

Then, setting $K_m = [A]^n$ yields $\frac{dA}{dt} = \frac{v_{max}}{2}$

Search for a catalyst molecule

Let's assume reaction $\mathbf{A} \implies \mathbf{B}$ produced an error $\sigma > \alpha$ for any of the rate function described above.

$\forall C \in S \backslash \{A, B\}$

- Reaction $\mathbf{A} + \mathbf{C} \implies \mathbf{B} + \mathbf{C}$ is considered
- Its inferred dynamics $k[A][C]$ are compared to $k[A]$

$$C^* = \underset{C \in S \backslash \{A,B\}}{\operatorname{argmin}} \sigma_C$$

If $\sigma_{C^*} < \sigma$, reaction $\mathbf{A} + \mathbf{C}^* \implies \mathbf{B} + \mathbf{C}^*$ is selected.

Selection and Update

Reaction $(R, P, f) = r^* = \underset{r \in \mathscr{R}}{\operatorname{argmin}}\, \sigma_r$ is selected.

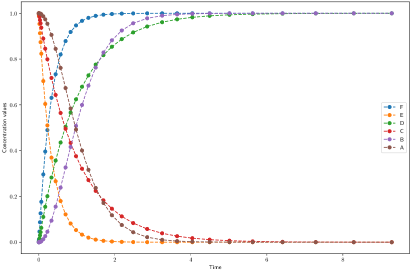
Its effect on the transitions removed.

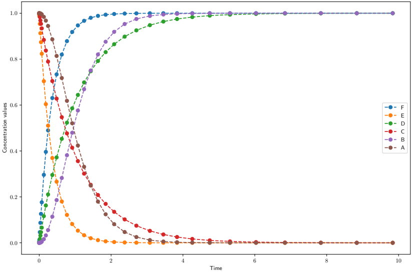
$\forall (x, d, j) \in \mathscr{T}$ s.t. $x_i > 0 \ \forall i \in R$

- $\forall i \in P, d_i \leftarrow d_i - d_0 f(x)$
- $\forall i \in R, d_i \leftarrow d_i + d_0 f(x)$

This update is followed by a new iteration of the algorithm. The main loop goes on while $\sigma_{r^*} < \alpha$

Results on parallel CRN



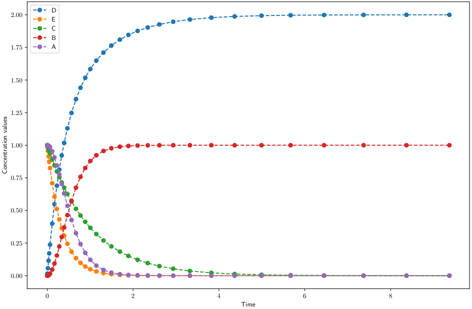


Hidden CRN
$A + D \xRightarrow{2} B + D$
$C \xRightarrow{1} D$
$E \xRightarrow{3} F$

Learned CRN
$A + D \xRightarrow{2.12} B + D$
$C \xRightarrow{0.96} D$
$E \xRightarrow{2.73} F$

It should be noticed that in this case, we have exactly $\forall (x, d, j) \in \mathscr{T}, |d_A| = |d_B|, |d_C| = |d_D|$ and $|d_E| = |d_F|$.

Reactant parallel CRN



Hidden CRN	Learned CRN
$A + D \xrightarrow{2} B + D$	$A + D \xrightarrow{2.2} B + D$
$C \xrightarrow{1} D$	$C \xrightarrow{0.98} D$
$E \xrightarrow{3} D$	$E \xrightarrow{2.78} D$

Here, $\nexists\, i \in S \backslash \{D\}$ s.t. $|d_D| \approx |d_i|$

- solution : find (i_1, i_2) s.t. $|d_D| \approx |d_{i_1}| + |d_{i_2}|$
- ensure $d_{i_1} d_{i_2} > 0$ and $d_{i_1} d_D < 0$

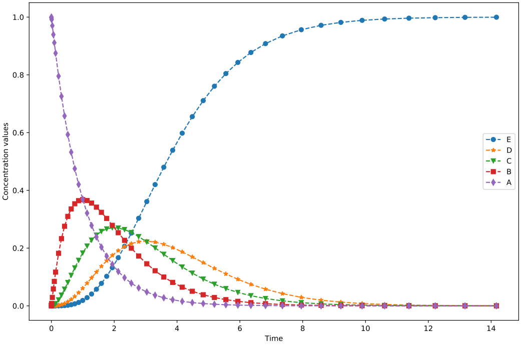
Promote reactions inferred on *sparse* transitions

Let $\mathscr{T}_r := \{(x, d, j) \in \mathscr{T} \mid r \in \text{ reaction\_inference}(d)\}$ be the *support* of reaction r :

Few species present $\implies$ more **informative** transitions $\implies$ inferred reaction more reliable

- Species s is considered *absent* of transition (x, d, j) if $x_s < \gamma \max_t x_s^j(t)$
- Let $m = mean(\#\{\text{absent species in } x \; \forall (x, d, j) \in \mathscr{T}_r\})$. $\sigma \leftarrow \frac{\sigma}{1+m}$

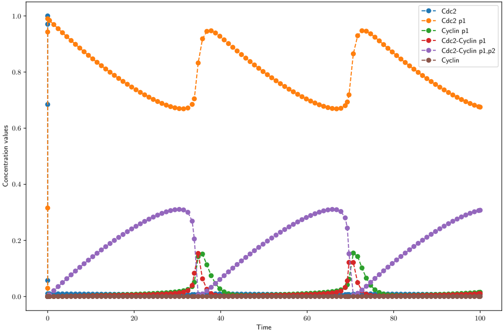
Back to the chain CRN



Hidden CRN	Learned CRN
$A \xrightarrow{1} B$	$A \xrightarrow{1.07} B$
$B \xrightarrow{1} C$	$B \xrightarrow{1.09} C$
$C \xrightarrow{1} D$	$C \xrightarrow{1.04} D$
$D \xrightarrow{1} E$	$D \xrightarrow{0.99} E$

- At the end of the simulation only species D and E are showing non negligible concentrations values.
- Reaction $D \implies E$ will then benefit of this sparsity criterion.

Results on the Cell Cycle of Tyson (1991)



Hidden CRN	Learned CRN
$\emptyset \xRightarrow{0.015} cy$	$\emptyset \xRightarrow{0.66} cy1 + cdcy2$
$cy + cd1 \xRightarrow{200} cdcy2$	$\emptyset \xRightarrow{0.01} cdcy2$
$cdcy2 \xRightarrow{0.018} cdcy1$	$cdcy2 \xRightarrow{0.1152} cdcy1$
$cdcy2 + 2 * cdcy1 \xRightarrow{180} 3 * cdcy1$	$cdcy2 \xRightarrow{0.05} cy1$
$cdcy1 \xRightarrow{1} cy1 + cd$	$cdcy1 \xRightarrow{1.62} \emptyset$
$cy1 \xRightarrow{0.6} \emptyset$	$cy1 \xRightarrow{0.4} cdcy1$
$cd1 \xRightarrow{100} cd$	$cd1 \xRightarrow{11259} cd$
$cd \xRightarrow{10000} cd1$	$cd \xRightarrow{5912} cd1$

F-score on simulations traces from a hidden model

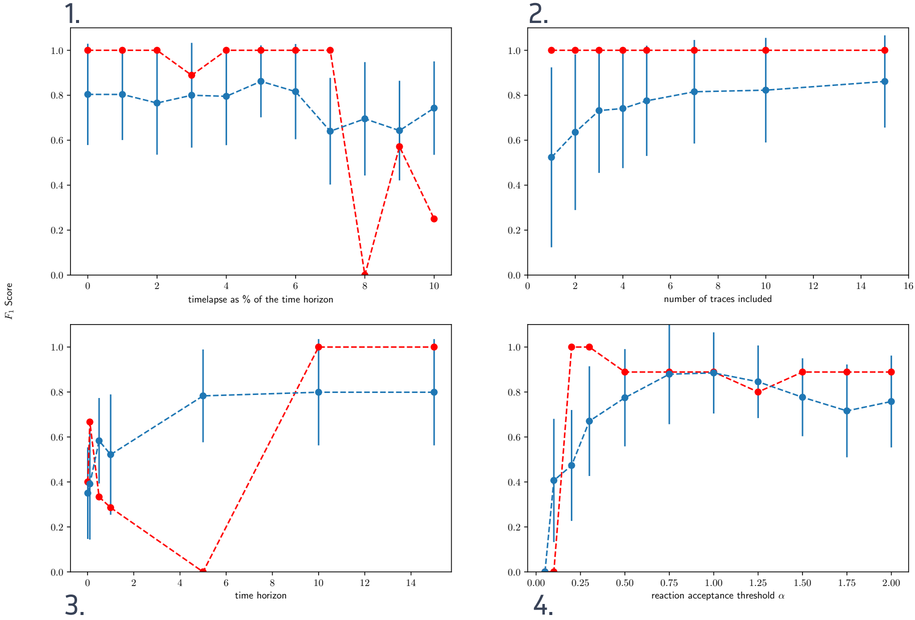
$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \text{ where precision } = \frac{\text{tp}}{\text{tp+fp}} \text{ recall } = \frac{\text{tp}}{\text{tp+fn}}$$

(tp : true positive ; fp : false positive ; fn : false negatives)

Allows to assess sensibility of the algorithm to :

1. Level of trace subsampling
2. Number of traces with random initial conditions
3. Length of the traces
4. Reaction acceptance threshold α

Evaluation of the algorithm on the Chain CRN



Complexity

Proposition

The time complexity of the CRN learning algorithm for inferring one reaction is $\mathscr{O}(t.n^2)$ where t is the number of transitions in the traces and n the number of variables.

$\Longrightarrow$ 5 minutes on real data (91 cells and ≈ 18000 transitions, 3 variables)

Conclusion and Perspectives

- An unsupervised greedy algorithm able to infer meaningful reaction networks from time-series data.
- Reaction selection is driven by the analysis of the ratio between observed dynamics and inferred dynamics for each reaction
- Linear complexity in the number of data points and quadratic in the number of species

Perspectives :

- Relax the stoichiometry bounded to 1 constraint
- Add the kinetics in the F score of the learned model w.r.t. hidden model
- Infer hidden variables