Unsupervised learning of chemical reaction networks from Time Series Data

Julien Martinelli

December, 19^{th} 2018

François Fages

Annabelle Ballesta

Inria
inventeurs du monde numérique

Inserm
La science pour la santé
From science to health

Mechanistic Model Learning

The Machine Learning area 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)

Some attempts at Mechanistic Model Learning

- DREAM3 (2008) - Network Inference Challenge
- Logic programming combined with prior knowledge on the network's structure allows to learn the boolean function responsible for each species

  Boolean Network Identification from Perturbation Time Series Data combining Dynamics Abstraction and Logic Programming. L. Pauleve et al.

- Evolutionnary Algorithms : based on the minimization of a fitness criterion measuring the difference between the observed data and the proposed mechanistic models

  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

  P. Zoppoli et al.

Toward Personalized Chronotherapies at INSERM U935

- In the biomedical case, predictions are required for instance to determine the optimal hour of drug delivery.
- Moreover, in the case of Personalized Medicine, we want to learn a model of the patient
- Learning a mechanistic model would give these predictions consistency through the understanding of the biological processes underneath
- We aim at learning not only parameters but also model structure

The Problem of Reaction Network Inference

Input : observed time-series data from biological experiments such as proteomic data.

Output : a set of reactions defining a model $\mathcal{M}$ reproducing similar time-series data

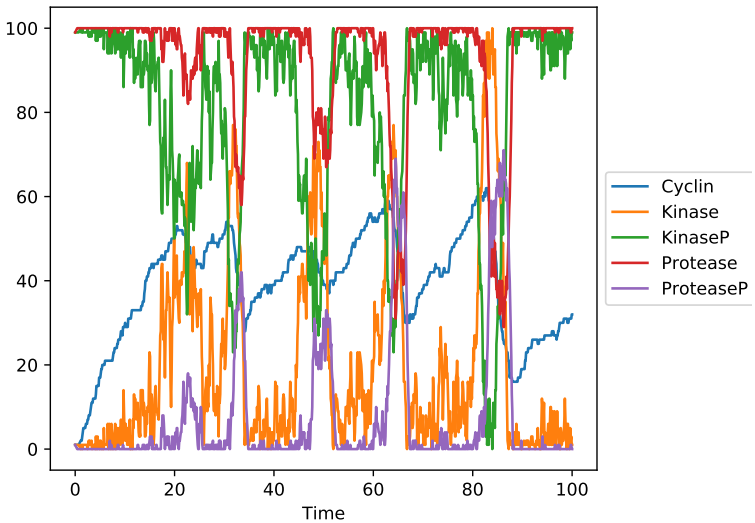
Molecular Species

$$W_i = \begin{pmatrix} A & B & C & D \\ A(t_0) & B(t_0) & C(t_0) & D(t_0) \\ A(t_1) & B(t_1) & C(t_1) & D(t_1) \\ \vdots & \vdots & \vdots & \vdots \\ A(T) & B(T) & C(T) & D(T) \end{pmatrix}$$

$$\longrightarrow \begin{cases} A \Longrightarrow B \\ C \Longrightarrow A \\ \vdots \quad \vdots \quad \vdots \\ D + B \Longrightarrow C \end{cases}$$

Inferred set of reactions

Time series Data



Chemical Reaction Network (CRN)

Hypothesis : Stoichiometry coefficients are less or equal to 1.

Definition

A reaction j is a triplet (R_j, P_j, h_j)
R_j is the set of reactants
P_j the set of products
h_j is the rate function

A CRN is a set of reactions $\mathcal{M} = (R_j, P_j, h_j)_{1 \leqslant j \leqslant J}$

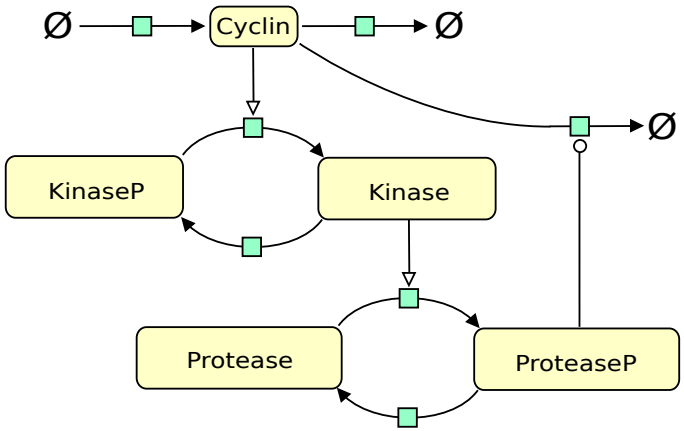
A catalyst is a species $B \in R_j \bigcap P_j$

Example

$R = \{A\} \quad P = \{B\} \quad h : x \longmapsto k \cdot x$

$k * A \ for \ A \implies B$

Simulated Data from Minimal Mitotic Oscillator



Goldbeter, 1991 - Biomodels Repository

$$0.02\,Cyclin \quad for \quad Cyclin \Longrightarrow \_$$

$$0.025\,Cyclin \quad for \quad \_ \Longrightarrow Cyclin$$

$$3\frac{KinaseP}{0.005 + KinaseP}\frac{Cyclin}{0.5 + Cyclin} \quad for \quad KinaseP + Cyclin \Longrightarrow Kinase + Cyclin$$

$$1.5\frac{Kinase}{0.005 + Kinase} \quad for \quad Kinase \Longrightarrow KinaseP$$

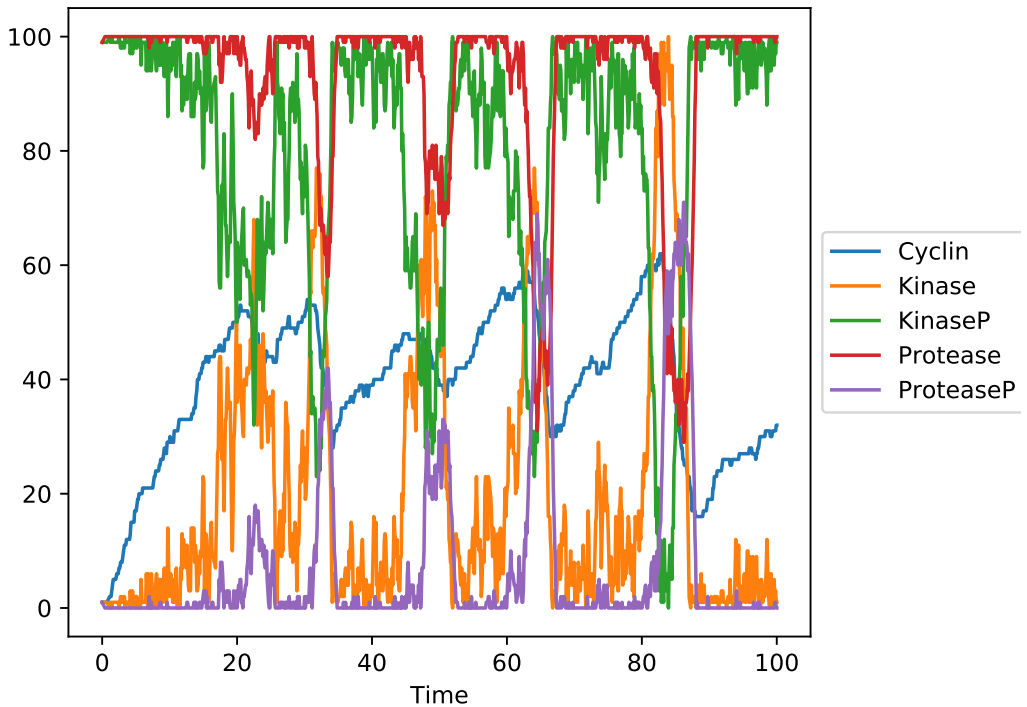
$$\frac{Kinase \cdot Protease}{0.0051 + Protease} \quad for \quad Protease + Kinase \Longrightarrow ProteaseP + Kinase$$

$$0.5\frac{ProteaseP}{ProteaseP + 0.005} \quad for \quad ProteaseP \Longrightarrow Protease$$

$$0.25\frac{Cyclin}{0.02 + Cyclin} \quad for \quad Cyclin + ProteaseP \Longrightarrow ProteaseP$$

Stochastic Simulation Algorithm

We consider stochastic simulation traces from an hidden model (Continuous time Markov chain)

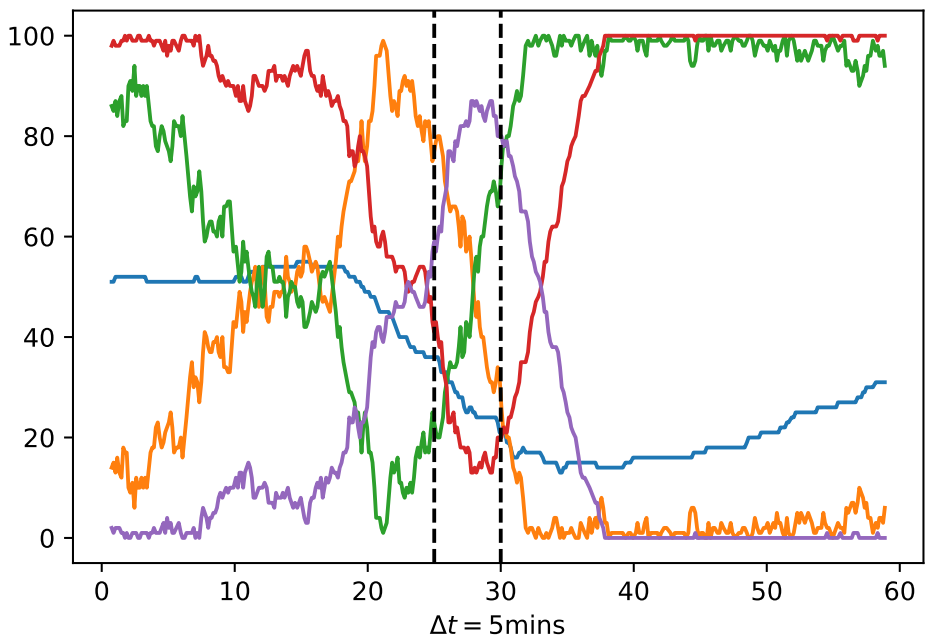


Numerical simulation using the *Gillespie Algorithm* - Minimal Mitotic Oscillator

Adding Subsampling to traces

Subsampling hypothesis : We do not observe every transition from the Markov chain simulation, only a sample of them every $\Delta t = 5 mins$

$\rightarrow$ Therefore we do not observe reactions one by one but *macro transitions*.



Adding Noise to traces

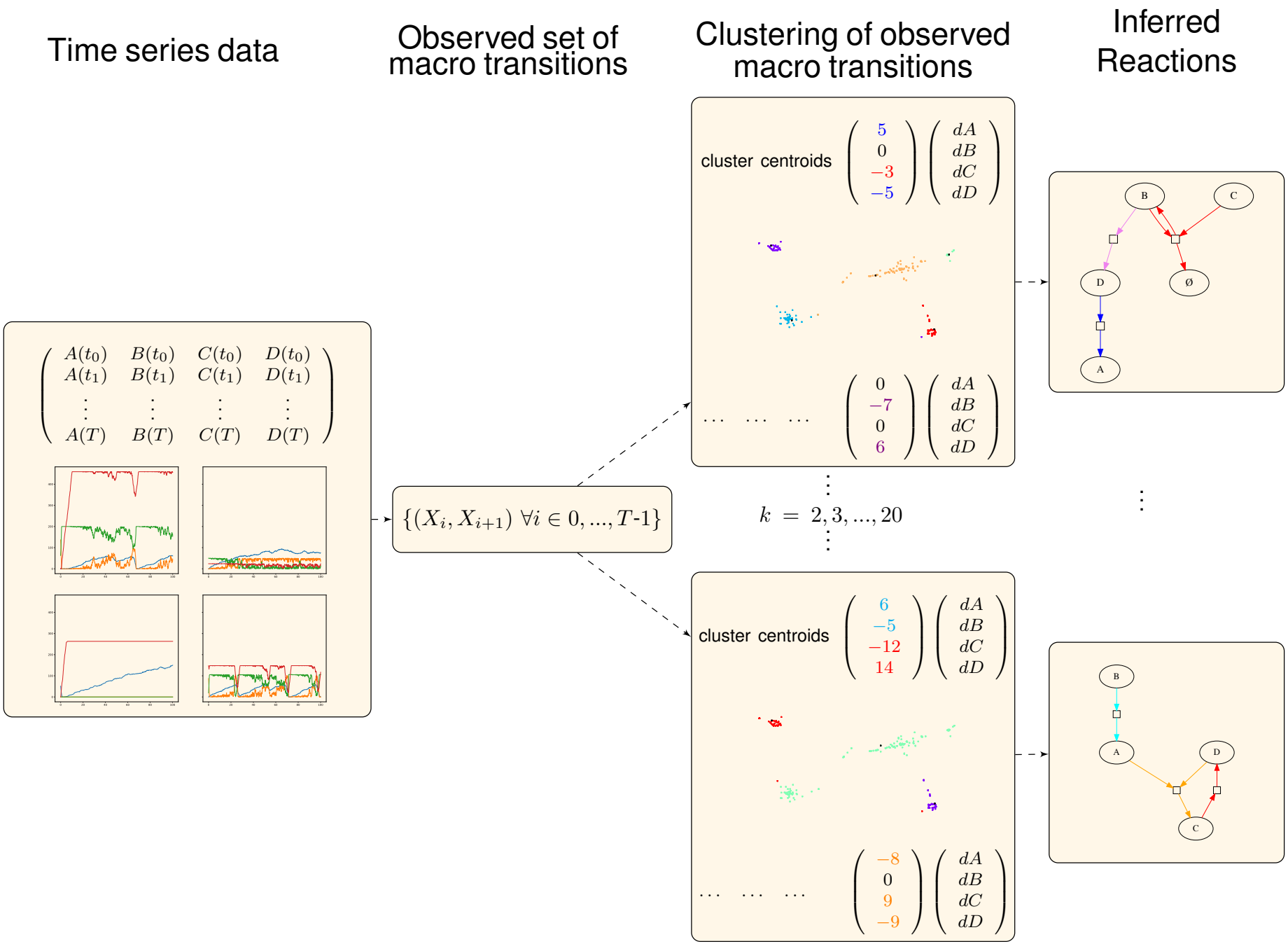
Multiplicative Gaussian noise is added to the predecessor state and the successor state.

$$X_{meas} = X_{sim} * e^{w} \text{ where } w \sim \mathcal{N}(0, \sigma) \text{ and } \sigma = 0.003$$

$\rightarrow$ A species more present than another will then be more noisy.

Noise is then suppressed by rounding to the closest integer.

Workflow of the learning algorithm



Clustering of the Observed Macro Transitions

Finite differences between the successor state and the predecessor state are computed while predecessor and successor state are stored.

Example

$$\begin{pmatrix} 32 \\ 19 \\ 88 \\ 57 \\ 6 \end{pmatrix} \rightarrow \begin{pmatrix} 32 \\ 24 \\ 82 \\ 49 \\ 14 \end{pmatrix} \begin{pmatrix} Cyclin \\ Kinase \\ KinaseP \\ ProteaseP \\ Protease \end{pmatrix} \qquad \begin{pmatrix} 0 \\ 5 \\ -6 \\ -8 \\ 8 \end{pmatrix} \begin{pmatrix} Cyclin \\ Kinase \\ KinaseP \\ ProteaseP \\ Protease \end{pmatrix}$$

Macro transition (P_i, S_i) and associated difference vector $\delta_i = S_i - P_i$

Macro Transitions Clustering based on difference vectors

Clustering as a way to extract information from the dataset
$\rightarrow$ We choose the *K-medoids* algorithm with the squared euclidean distance

- Start with randomly chosen centroids and update the clusters :

$$C_k = \{\delta_i \text{ s.t } \underset{\delta \in M}{\text{argmin}} \, ||\delta - \delta_i||_2^2 = M_k\}$$

- Then the centroids are updated

$$M_k = \underset{\delta \in C_k}{\text{argmin}} \frac{1}{|C_k|} \sum_{\delta_i \in C_k} ||(\delta - \delta_i)||_2^2$$

- Repeat until the partitioning reaches a stable state

Centroids are actual members of the dataset

Inferring reactions from the centroids

$$\begin{pmatrix} 0 \\ -7 \\ 6 \\ 2 \\ -3 \end{pmatrix} \qquad \begin{pmatrix} \mathit{Cyclin} \\ \mathit{Kinase} \\ \mathit{KinaseP} \\ \mathit{ProteaseP} \\ \mathit{Protease} \end{pmatrix}$$

$$x^* = \underset{\mathit{species}}{\operatorname{argmin}}\, x^i$$

Inferring reactions from the centroids

$$\begin{pmatrix} 0 \\ -7 \\ 6 \\ 2 \\ -3 \end{pmatrix} \quad \begin{pmatrix} Cyclin \\ Kinase \\ KinaseP \\ ProteaseP \\ Protease \end{pmatrix}$$

$$x^* = \underset{species}{\text{argmin}}\, x^i$$

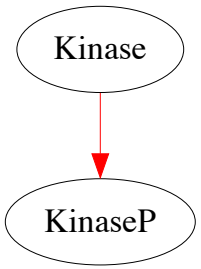
$$\begin{pmatrix} 0 \\ -7 \\ 6 \\ 2 \\ -3 \end{pmatrix} \quad \begin{pmatrix} Cyclin \\ Kinase \\ KinaseP \\ ProteaseP \\ Protease \end{pmatrix}$$

Collect every x^i $s.t. |\frac{x^*}{x^i}| \leqslant \alpha$

Inferring reactions from the centroids

$$\begin{pmatrix} 0 \\ -7 \\ 6 \\ 2 \\ -3 \end{pmatrix} \quad \begin{pmatrix} Cyclin \\ Kinase \\ KinaseP \\ ProteaseP \\ Protease \end{pmatrix}$$

$$x^* = \underset{species}{\text{argmin}}\, x^i$$



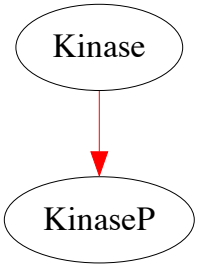
$$\begin{pmatrix} 0 \\ -7 \\ 6 \\ 2 \\ -3 \end{pmatrix} \quad \begin{pmatrix} Cyclin \\ Kinase \\ KinaseP \\ ProteaseP \\ Protease \end{pmatrix}$$

Collect every x^i $s.t. |\frac{x^*}{x^i}| \leqslant \alpha$

Inferring reactions from the centroids

$$\begin{pmatrix} 0 \\ -7 \\ 6 \\ 2 \\ -3 \end{pmatrix} \quad \begin{pmatrix} Cyclin \\ Kinase \\ KinaseP \\ ProteaseP \\ Protease \end{pmatrix}$$

$$x^* = \underset{species}{\operatorname{argmin}}\, x^i$$



$$\begin{pmatrix} 0 \\ -7 \\ 6 \\ 2 \\ -3 \end{pmatrix} \quad \begin{pmatrix} Cyclin \\ Kinase \\ KinaseP \\ ProteaseP \\ Protease \end{pmatrix}$$

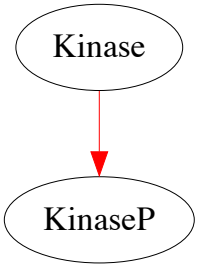
Collect every x^i $s.t. |\frac{x^*}{x^i}| \leqslant \alpha$

$$\begin{pmatrix} 0 \\ -7 \\ 6 \\ 2 \\ -3 \end{pmatrix} \quad \begin{pmatrix} Cyclin \\ Kinase \\ KinaseP \\ ProteaseP \\ Protease \end{pmatrix}$$

Inferring reactions from the centroids

$$\begin{pmatrix} 0 \\ -7 \\ 6 \\ 2 \\ -3 \end{pmatrix} \quad \begin{pmatrix} Cyclin \\ Kinase \\ KinaseP \\ ProteaseP \\ Protease \end{pmatrix}$$

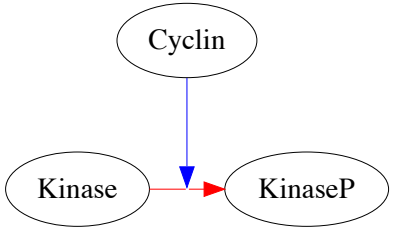
$$x^* = \underset{species}{\operatorname{argmin}}\, x^i$$



$$\begin{pmatrix} 0 \\ -7 \\ 6 \\ 2 \\ -3 \end{pmatrix} \quad \begin{pmatrix} Cyclin \\ Kinase \\ KinaseP \\ ProteaseP \\ Protease \end{pmatrix}$$

Collect every x^i $s.t. |\frac{x^*}{x^i}| \leqslant \alpha$

$$\begin{pmatrix} 0 \\ -7 \\ 6 \\ 2 \\ -3 \end{pmatrix} \quad \begin{pmatrix} Cyclin \\ Kinase \\ KinaseP \\ ProteaseP \\ Protease \end{pmatrix}$$



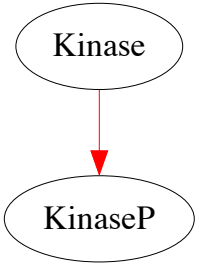
Catalyst Candidate

with 0 variation

Inferring reactions from the centroids

$$\begin{pmatrix} 0 \\ -7 \\ 6 \\ 2 \\ -3 \end{pmatrix} \quad \begin{pmatrix} Cyclin \\ Kinase \\ KinaseP \\ ProteaseP \\ Protease \end{pmatrix}$$

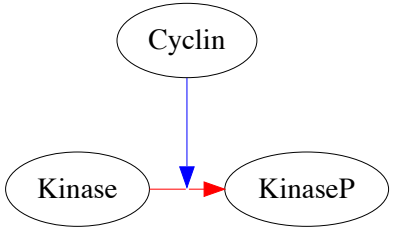
$$x^* = \underset{species}{\operatorname{argmin}}\, x^i$$



$$\begin{pmatrix} 0 \\ -7 \\ 6 \\ 2 \\ -3 \end{pmatrix} \quad \begin{pmatrix} Cyclin \\ Kinase \\ KinaseP \\ ProteaseP \\ Protease \end{pmatrix}$$

Collect every x^i $s.t. |\frac{x^*}{x^i}| \leqslant \alpha$

$$\begin{pmatrix} 0 \\ -7 \\ 6 \\ 2 \\ -3 \end{pmatrix} \quad \begin{pmatrix} Cyclin \\ Kinase \\ KinaseP \\ ProteaseP \\ Protease \end{pmatrix}$$



Catalyst Candidate

with 0 variation

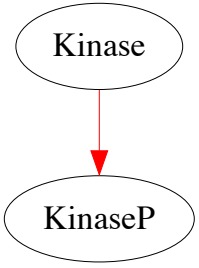
$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 2 \\ -3 \end{pmatrix} \quad \begin{pmatrix} Cyclin \\ Kinase \\ KinaseP \\ ProteaseP \\ Protease \end{pmatrix}$$

Set species to 0 and start over.

Inferring reactions from the centroids

$$\begin{pmatrix} 0 \\ -7 \\ 6 \\ 2 \\ -3 \end{pmatrix} \quad \begin{pmatrix} Cyclin \\ Kinase \\ KinaseP \\ ProteaseP \\ Protease \end{pmatrix}$$

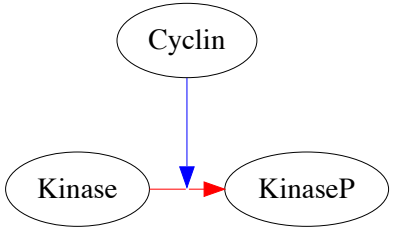
$$x^* = \underset{species}{\operatorname{argmin}}\, x^i$$



$$\begin{pmatrix} 0 \\ -7 \\ 6 \\ 2 \\ -3 \end{pmatrix} \quad \begin{pmatrix} Cyclin \\ Kinase \\ KinaseP \\ ProteaseP \\ Protease \end{pmatrix}$$

Collect every x^i $s.t. |\frac{x^*}{x^i}| \leqslant \alpha$

$$\begin{pmatrix} 0 \\ -7 \\ 6 \\ 2 \\ -3 \end{pmatrix} \quad \begin{pmatrix} Cyclin \\ Kinase \\ KinaseP \\ ProteaseP \\ Protease \end{pmatrix}$$



Catalyst Candidate

with 0 variation

$$\begin{pmatrix} 0 \\ 0 \\ 0 \\ 2 \\ -3 \end{pmatrix} \quad \begin{pmatrix} Cyclin \\ Kinase \\ KinaseP \\ ProteaseP \\ Protease \end{pmatrix}$$

Set species to 0 and start over.

Results on Minimal Mitotic Oscillator (Goldbeter, 1991)

R_1 : ProteaseP $\Longrightarrow$ Protease

R_2 : Kinase + Protease $\Longrightarrow$ ProteaseP + Kinase

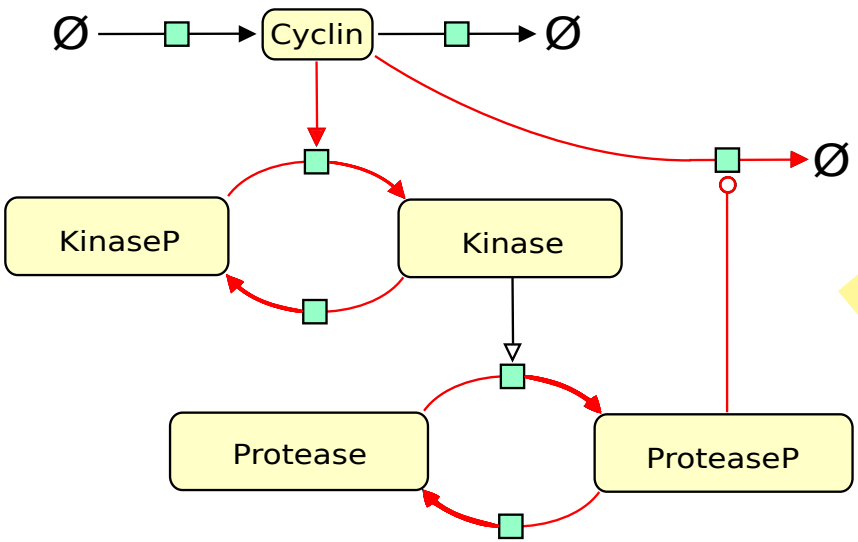
R_3 : Cyclin + KinaseP $\Longrightarrow$ Kinase + Cyclin

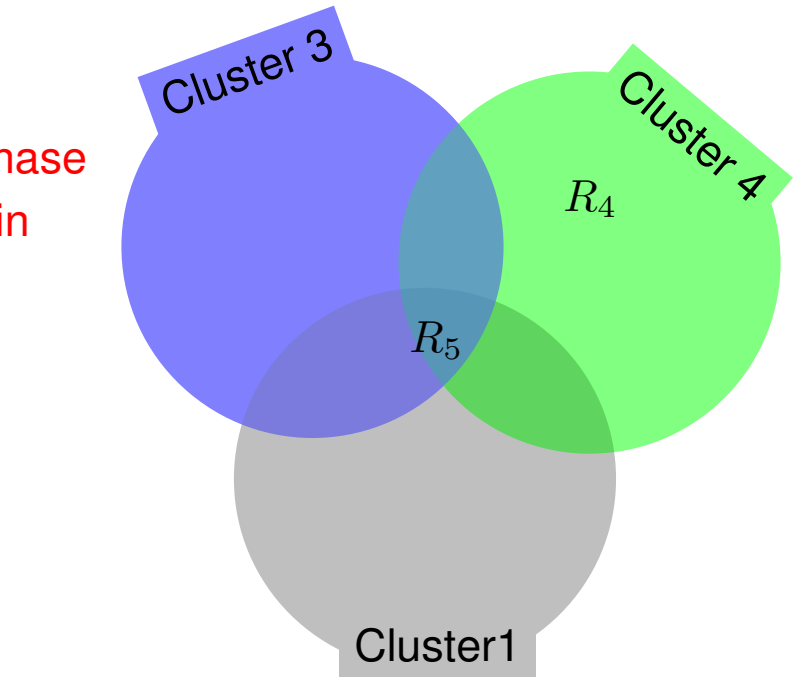
R_4 : ProteaseP + Cyclin $\Longrightarrow$ ProteaseP

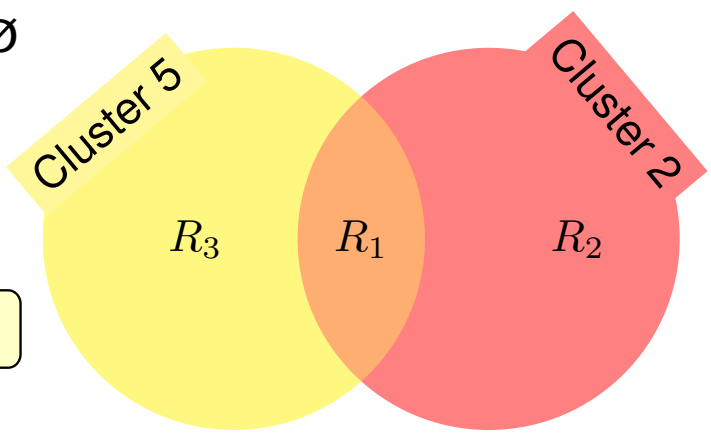
R_5 : Kinase $\Longrightarrow$ KinaseP

R_6 : _ $\Longrightarrow$ Cyclin

R_7 : Cyclin $\Longrightarrow$ _







Model Selection Step

- Choosing the right aggregated network amounts to choosing the optimal number of clusters k
- The reaction inference algorithm outputs a set of reactions, defining a generative model $\hat{\mathcal{M}}$
- Model quality can be assessed by comparing the distribution of $\mathcal{M}$ to the one described by $\hat{\mathcal{M}}$

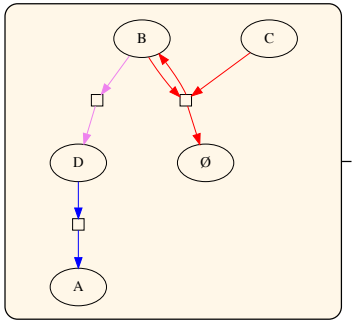
Model selection protocol

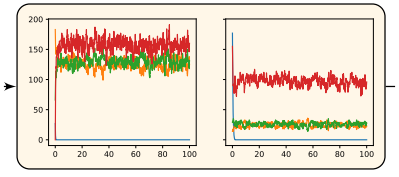
Learned Reactions Networks

New Simulations

Compute Distribution of difference vectors

Comparison with observed data's distribution





$$Y_i^1 = X_{i+1}^1 - X_i^1$$

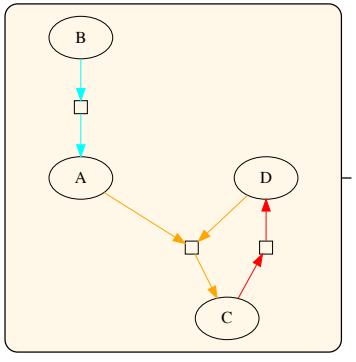
$$\mu_1 = \frac{1}{n\text{-}1} \sum_{i=1}^{n-1} \mathbb{1}_{\{Y_i \in \cdot\}}$$

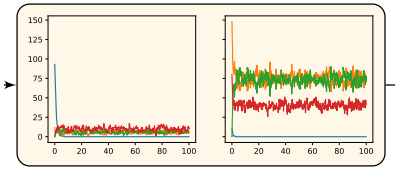
ν

$||\mu_1 - \nu||_{TV}$

$k = 2, 3, ..., 20$

$\mu^* = argmin_\mu ||\mu - \nu||_{TV}$



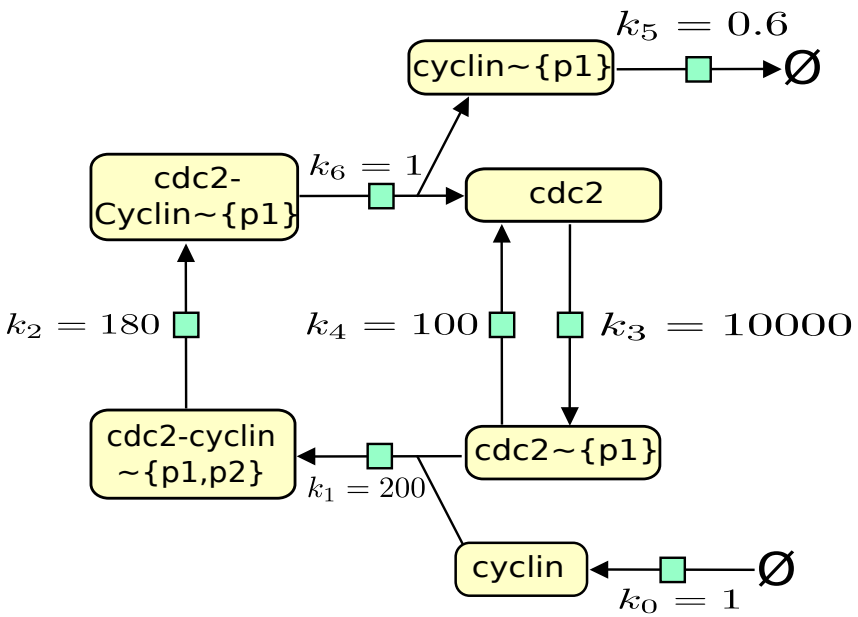


$$Y_i^k = X_{i+1}^k - X_i^k$$

$$\mu_k = \frac{1}{n\text{-}1} \sum_{i=1}^{n-1} \mathbb{1}_{\{Y_i \in \cdot\}}$$

$||\mu_k - \nu||_{TV}$

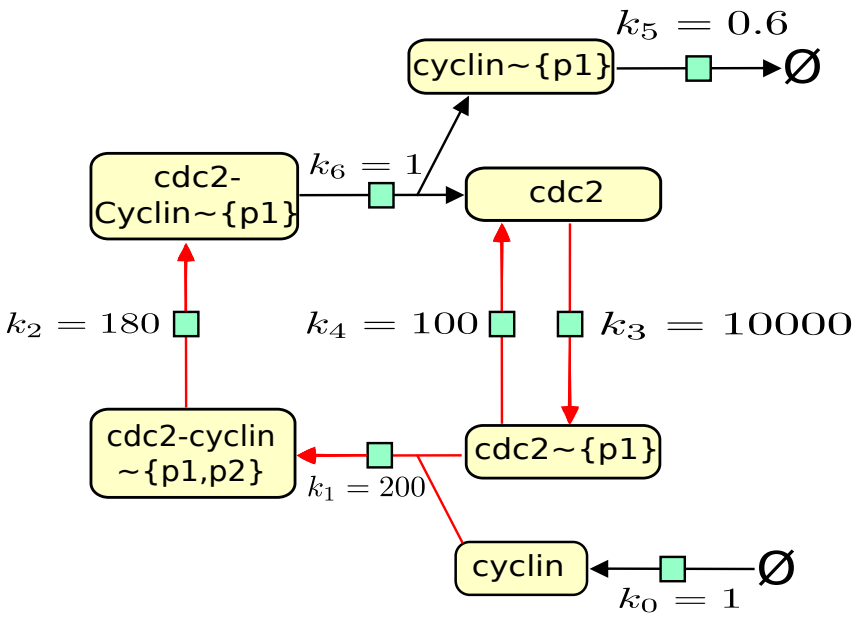
6 variables Cell Cycle Model (Tyson, 1991)



$$Cdc2 \implies Cdc2 \sim \{p1\}$$
$$Cdc2 \sim \{p1\} \implies Cdc2$$
$$Cdc2\text{-}Cyclin \sim \{p1, p2\} \implies Cdc2\text{-}Cyclin \sim \{p1\}$$
$$Cdc2 \sim \{p1\} + Cyclin \implies Cdc2\text{-}Cyclin \sim \{p1, p2\}$$
$$\_ \implies Cyclin$$
$$Cyclin \sim \{p1\} \implies \_$$
$$Cdc2\text{-}Cyclin \sim \{p1\} \implies Cyclin \sim \{p1\} + Cdc2$$
$$Cdc2\text{-}Cyclin \sim \{p1, p2\} + 2 * Cdc2\text{-}Cyclin \sim \{p1\} \implies 3 * Cdc2\text{-}Cyclin \sim \{p1\}$$

	mean transition difference vector	max transition difference vector
Cyclin	3.3	45
Cyclin~{p1}	1.02	2
Cdc2	53.54	853
Cdc2~{p1}	50.3	840
Cdc2-Cyclin~{p1}	6.43	123
Cdc2-Cyclin~{p1,p2}	4.7	122

6 variables Cell Cycle Model (Tyson, 1991)



$Cdc2 \Longrightarrow Cdc2 \sim \{p1\}$

$Cdc2 \sim \{p1\} \Longrightarrow Cdc2$

$Cdc2\text{-}Cyclin \sim \{p1, p2\} \Longrightarrow Cdc2\text{-}Cyclin \sim \{p1\}$

$Cdc2 \sim \{p1\} + Cyclin \Longrightarrow Cdc2\text{-}Cyclin \sim \{p1, p2\}$

$\_ \Longrightarrow Cyclin$

$Cyclin \sim \{p1\} \Longrightarrow \_$

$Cdc2\text{-}Cyclin \sim \{p1\} \Longrightarrow Cyclin \sim \{p1\} + Cdc2$

$Cdc2\text{-}Cyclin \sim \{p1, p2\} + 2 * Cdc2\text{-}Cyclin \sim \{p1\}$

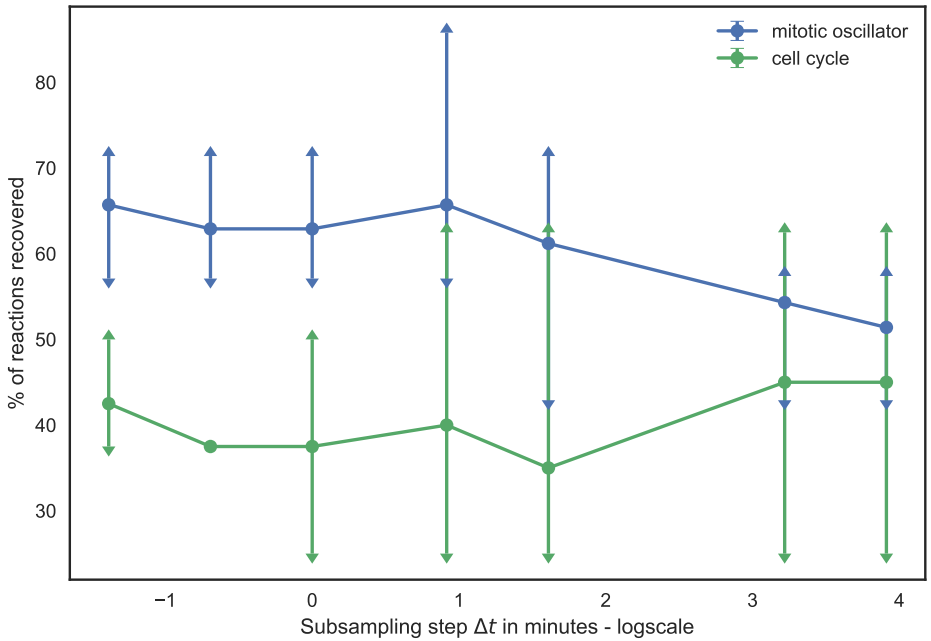
$\Longrightarrow 3 * Cdc2\text{-}Cyclin \sim \{p1\}$

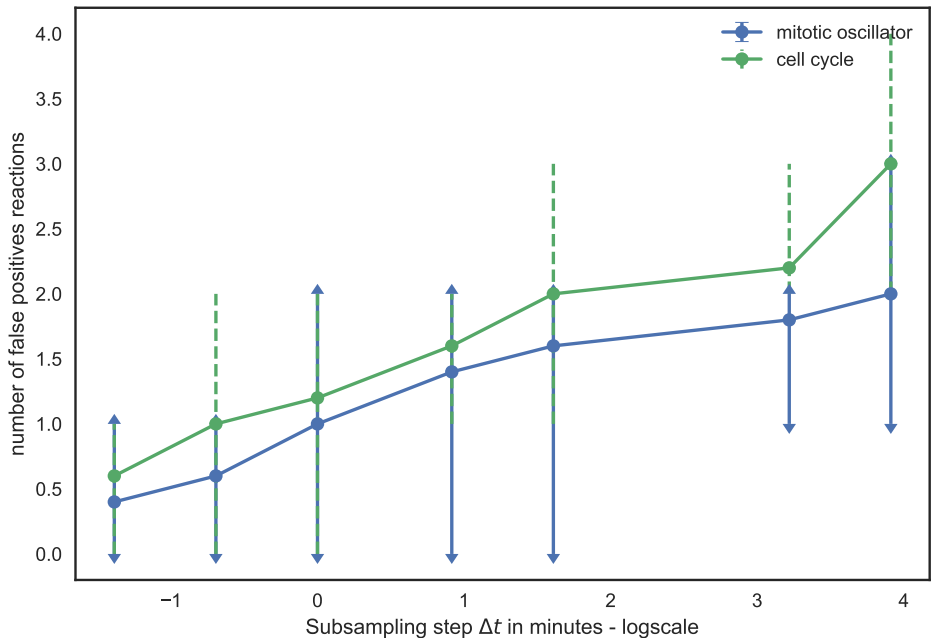
- Reaction recovered are precisely the four fastest ones hence those with the highest probability to occur when possible

  (False Positive : 0%, False Negative : 50%)

- The gap between kinetic parameters values results in a slow/fast dynamic, a limit of the stochastic approach.

Subsampling effect on learning





- Rules including catalysts are inferred without the latter : more false positives
- Scarce reactions such as $\_ \Longrightarrow A$ are inferred : less false negatives
- As the subsampling step grows, more false positives appear.

Conclusion and Perspectives

- Unsupervised reaction inference algorithm dealing with subsampled and noisy time-series data
- The algorithm finds original reactions
- But also misses other original reactions *(false negatives)*

  $\rightarrow$ high precision but low recall

Perspectives:

- Case where not all species are observed (Elisabeth Degrand's Master Thesis - Evolving CRN)