# A Statistical Learning Algorithm for Inferring Reaction Networks from Time Series Data

Julien Martinelli, Jeremy Grignard, Sylvain Soliman, François Fages
EP Lifeware, INRIA Saclay Ile de France

## Explainable AI by Learning Mechanistic Models

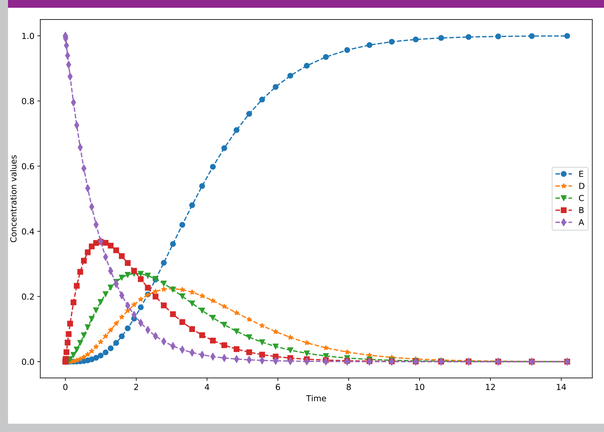
- The Machine Learning field provides tools to analyze time series data and yield predictions.
  - Classical algorithms are Recurrent Neural Networks
  - While predictions can be accurate, they do not come with an explanation
  - Black box model
- **Mechanistic Model Learning** aims at achieving the same predictive power with an explainable learned model

## Focus: Chemical Reaction Networks (CRN) Inference

- Input : **time series data on molecular concentrations**
  - single trace (wild type)
  - multiple traces with perturbed conditions (gene knock outs)
- Output :
  - **CRN structure**: reactions with -1/0/1 stoichiometry
  - **CRN kinetics**: mass action law, Michaelis-Menten or Hill functions

The learned CRN provides a mechanistic explanation of the observations and allows predictions
Learning parameters: well-understood
Learning structure: hard without prior knowledge (see DREAM challenge)

## Chain CRN Example

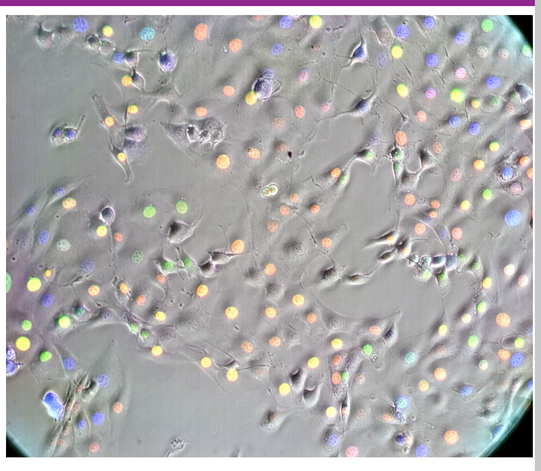



Learning from a single simulation trace

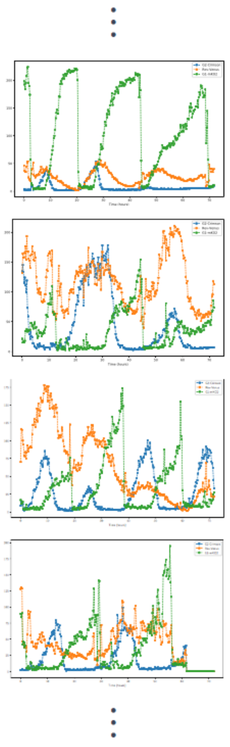
| Hidden CRN | Learned CRN |
|---|---|
| $A \xrightarrow{1} B$ | $A \xrightarrow{1.07} B$ |
| $B \xrightarrow{1} C$ | $B \xrightarrow{1.09} C$ |
| $C \xrightarrow{1} D$ | $C \xrightarrow{1.04} D$ |
| $D \xrightarrow{1} E$ | $D \xrightarrow{0.99} E$ |

## Application to Time Lapse Videomicroscopy Data

- NIH3T3 embryonic mouse fibroblasts
- Time lapse of 15 min during 72 hours
- Cell tracking (through cell divisions)
- 3 fluorescent markers of
  - cell cycle ($G1$ and $S$-$G2$-$M$)
  - circadian clock (*Reverb*-$\alpha$)



Feillet Delaunay INSERM 2013



Algorithm

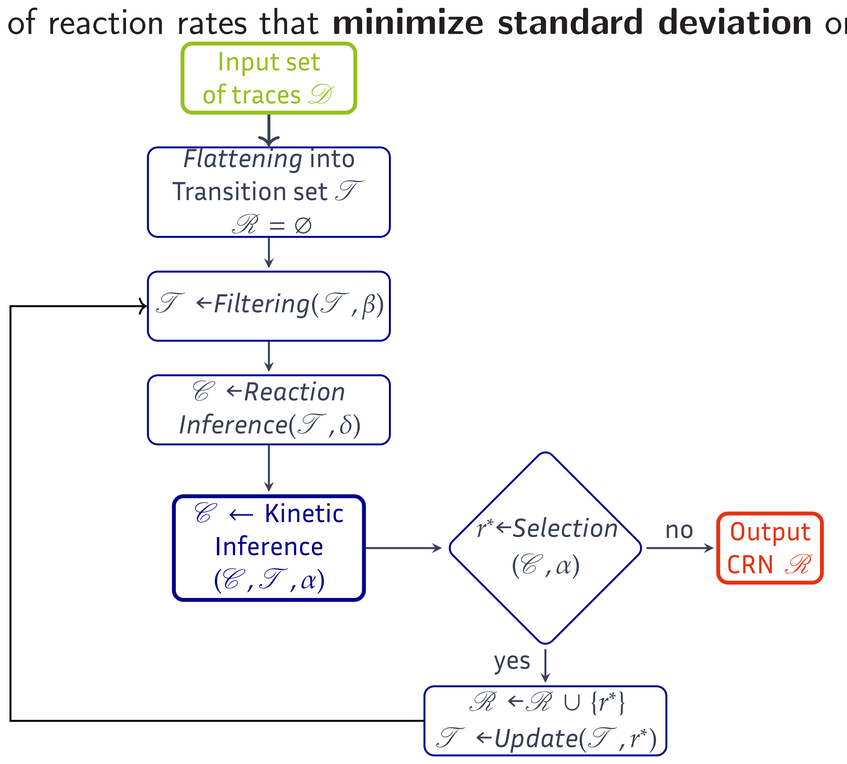
| Learned CRN | Rate Functions |
|---|---|
| $G1 \Longrightarrow G2$ | $7.1\frac{G1}{G1+3.68}$ |
| $RevErb\alpha \Longrightarrow G1$ | $22.56\frac{RevErb\alpha}{RevErb\alpha+71.45}$ |
| $G1 \Longrightarrow \varnothing$ | $5.96\frac{G1}{G1+5.0}$ |
| $G2 \Longrightarrow \varnothing$ | $54.84\frac{G2}{G2+176.23}$ |

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

$\Longrightarrow$ CRN learned in 5 minutes CPU time

## Statistical Learning Algorithm

- **Greedy algorithm** that iteratively infers reactions
- Reaction structures that maximise the pairing between reactant consumption and product formation in the observed transitions $\mathcal{F}$
- Choice of reaction rates that **minimize standard deviation** on $\mathcal{F}$




## Proposition

Time complexity in $\mathcal{O}(t.n^2)$ where

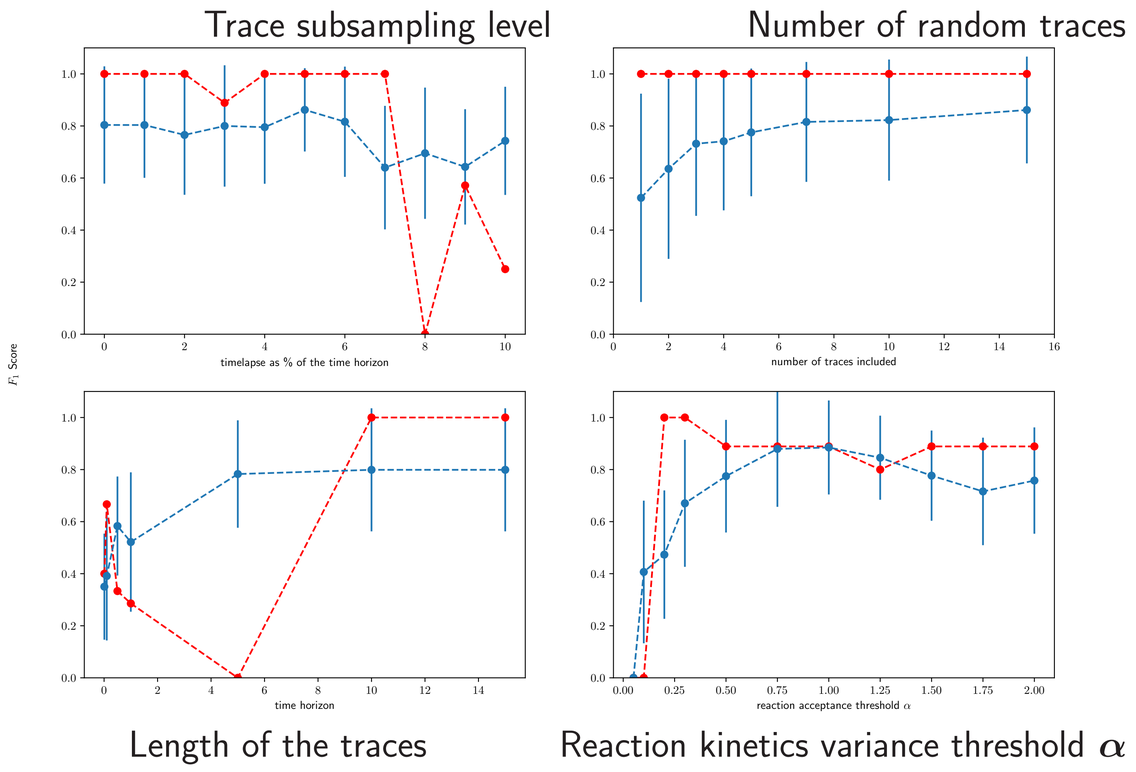
- $t$ is the number of observed transitions in the traces
- and $n$ the number of observed molecular species

## $F$-score on Simulation Traces from a Hidden Model

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

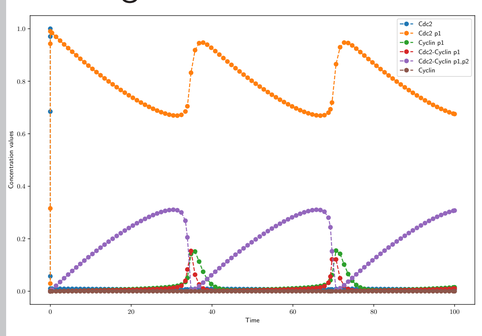
$\Longrightarrow$ **Sensitivity of the F-score** w.r.t. algorithmic parameters

Chain CRN example red: single trace (wild type)
blue: multiple traces (from perturbed initial states with random zeroes)




## Results on the Yeast Cell Cycle Model [Tyson 1991]

Single simulation trace




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

Slow/fast multiple time scales make fast reactions unobservable.
The slow dynamics is inferred.