On Inferring Reactions from Data Time Series by a Statistical Learning Greedy Heuristics

Julien Martinelli, Jeremy Grignard, Sylvain Soliman, François Fages

On Inferring Reactions from Data Time Series by a Statistical Learning Greedy Heuristics

Julien Martinelli[1,2], Jeremy Grignard[1,3], Sylvain Soliman[1], and François Fages[1]

[1] Inria Saclay-Île de France, Palaiseau, France
[2] INSERM U935, Villejuif, France
[3] Institut de Recherches Servier, Croissy sur Seine, France

Abstract. With the automation of biological experiments and the increase of quality of single cell data that can now be obtained by phosphoproteomic and time lapse videomicroscopy, automating the building of mechanistic models from these data time series becomes conceivable and a necessity for many new applications. While learning numerical parameters to fit a given model structure to observed data is now a quite well understood subject, learning the structure of the model is a more challenging problem that previous attempts failed to solve without relying quite heavily on prior knowledge about that structure. In this paper, we consider mechanistic models based on chemical reaction networks (CRN) with their continuous dynamics based on ordinary differential equations, and finite time series about the time evolution of concentration of molecular species for a given time horizon and a finite set of perturbed initial conditions. We present a greedy heuristics unsupervised statistical learning algorithm to infer reactions with a time complexity for inferring one reaction in $\mathcal{O}(t.n^2)$ where n is the number of species and t the number of observed transitions in the traces. We evaluate this algorithm both on simulated data from hidden CRNs, and on real videomicroscopy single cell data about the circadian clock and cell cycle progression of NIH3T3 embryonic fibroblasts. In all cases, our algorithm is able to infer meaningful reactions, though generally not a complete set for instance in presence of multiple time scales or highly variable traces.

1 Introduction

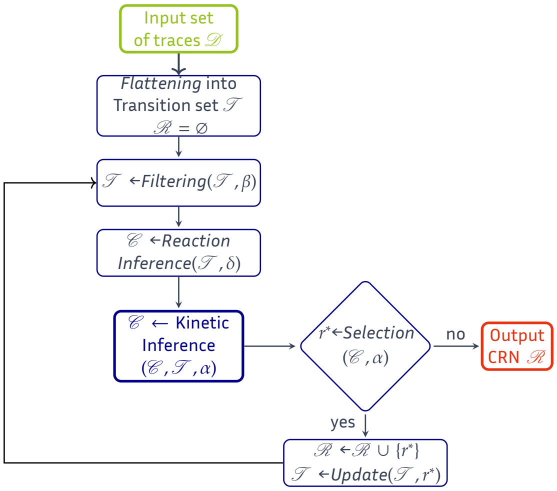
Recent breakthroughs in Machine Learning are paving the way for new kinds of algorithms for analysing data and making diagnosis and predictions in biology and medicine. While capable of making accurate predictions, the direct application of machine learning methods do not provide however a biological understanding of the underlying processes nor explanation for the predictions, and may be not accepted in the biomedical domain. For these reasons, a lot of work aims at providing explanations for the predictions made as output of neural networks or other machine learning algorithms trained on data.

Another approach is to try to learn mechanistic models that will make predictions instead of learning directly the predictions from the data. Building mechanistic models of cell processes is however a hard work which necessitates to determine the biochemical mechanisms that are responsible for the high level functions of the cell and its behaviors in normal and perturbed conditions. Automating this process would enable new applications such as automated experiment design or patient-tailored therapeutics.

The main difficulty is to be able to discriminate between causality and correlations in data time series [4]. Most of the work on network inference concerns either undirected interaction graph models, or influence models such as gene regulatory network but then requiring prior knowledge on the structure of the network. The DREAM challenges are used to measure progress in this field.

Less work is devoted to the learning of reaction models and chemical reaction networks (CRNs). In [1], this problem is defined as the minimization of a fitness criterion based on the compatibility of the learned mechanistic model with the observed traces. An evolutionary algorithm is proposed via a two-step iterative procedure: first a set of reactions is inferred, then mass action law kinetic parameters are estimated.

In this paper, we present a greedy heuristics with low computational complexity for inferring reactions from data time series. This unsupervised statistical learning algorithm does not require prior knowledge nor training. We consider at most binary reactions with mass-action, Michaelian or order 4 Hill kinetics. Based on a pairing of the variations of molecular species in each observed transition, the algorithm repeatedly infers the reaction that minimizes the standard deviation of the inferred rate function among all the observed transitions where the reaction can occur. Once inferred, the contributions of that reaction to state change in the set of observed transitions are subtracted before inferring the next reaction. Fig. 1 shows the flowchart and low complexity of this algorithm [3].

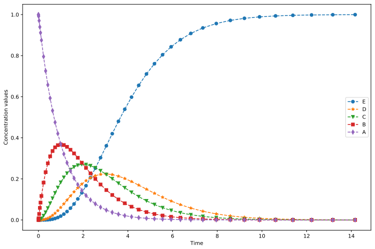

Input set of traces $\mathscr{D}$
Flattening into Transition set $\mathscr{T}$
$\mathscr{R} = \varnothing$
$\mathscr{T}$ ←*Filtering*($\mathscr{T}$, β)
$\mathscr{C}$ ←*Reaction Inference*($\mathscr{T}$, δ)
$\mathscr{C}$ ← Kinetic Inference ($\mathscr{C}$, $\mathscr{T}$, α)
r^*←*Selection* ($\mathscr{C}$, α)
no
Output CRN $\mathscr{R}$
yes
$\mathscr{R}$ ←$\mathscr{R} \cup \{r^*\}$
$\mathscr{T}$ ←*Update*($\mathscr{T}$, r^*)

Proposition 1. *The time complexity for inferring one reaction is $\mathcal{O}(t.n^2)$ where t is the number of observed transitions in the traces and n is the number of variables.*

Fig. 1. Flowchart of our CRN learning algorithm and complexity.

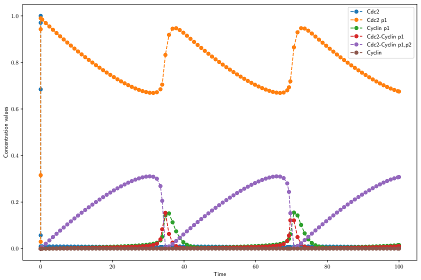
2 Evaluation on Simulation Traces

In the context of evaluating the learning algorithm on simulation traces, the hidden CRN used to generate the traces can be used to compare the learned CRN in terms of correct reactions (true positives), wrong reactions (false positives) and missing reactions (false negatives). On a simple chain of 4 reactions with mass action law kinetics over 5 molecular species, our algorithm is able to reconstruct the CRN from a single simulation trace (Fig. 2) with a low sensitivity to statistical learning thresholds.



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

Fig. 2. Chain example: simulation trace and learned CRN.

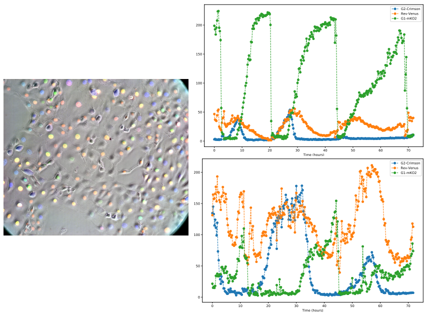


Hidden CRN	Learned CRN
$\emptyset \xRightarrow{0.015} cy$	$\emptyset \xRightarrow{0.66} cy1 + cdcy2$
$cy + cd1 \xRightarrow{200} cdcy2$	$\emptyset \xRightarrow{0.01} cdcy2$
$cdcy2 \xRightarrow{0.018} cdcy1$	$cdcy2 \xRightarrow{0.1152} cdcy1$
$cdcy2 + 2 * cdcy1 \xRightarrow{180} 3 * cdcy1$	$cdcy2 \xRightarrow{0.05}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

Fig. 3. Cell cycle model of Tyson [5] and learned reactions from the canonical trace where *cd* is present.

On the simulation trace depicted in Fig. 3 of the yeast cell cycle model of Tyson [5], our algorithm infers reactions corresponding to the observable slow dynamics of that system. In particular, the discrepancies concerning the synthesis reaction of the cyclin can be very well explained by the existence of multiple time scales in this model. When it is produced, the Cyclin is indeed immediately complexed with Cdc and phosphorylated by very fast reactions. Therefore the free state of the Cyclin cannot be observed and what is inferred is the synthesis of the fast equilibrium state where the Cyclin is in complex form. On the other hand, the autocatalysis reaction cannot be recovered since our algorithm does not consider stoichiometric coefficients other than $\{-1, 0, 1\}$.

3 Evaluation on Videomicroscopy Data



Learned reactions	Rate functions
$G1 \implies G2$	$7.1\frac{G1}{G1 + 3.68}$
$RevErb\alpha \implies G1$	$22.56\frac{RevErb\alpha}{RevErb\alpha + 71.45}$
$G1 \implies \emptyset$	$5.96\frac{G1}{G1 + 5.0}$
$G2 \implies \emptyset$	$54.84\frac{G2}{G2 + 176.23}$

Fig. 4. Inferred reactions on videomicroscopy data of embryonic NIH3T3 fibroblasts [2].

Fig. 4 shows the results on videomicroscopy data obtained over 3 days of NIH3T3 embryonic mice fibroblasts, using FUCCI markers of the cell cycle, and Rev-erb-α marker of the circadian clock, with a total of 91 tracked cell traces and 26000 observed state transitions [2]. It is remarkable that, despite the very high variability of cell behaviors, the stochasticity of the cell cycle, and the noise of measurements, meaningful reactions coupling the cell cycle progression with the circadian clock marker could be inferred from this dataset, in just 5 mn CPU time on a laptop.

On-going work concerns strategies to automatically adapt the threshold parameters of this statistical algorithm to the quality of the trace dataset.

References

1. Choi, K., Hellerstein, J., Wiley, H.S., Sauro, H.M.: Inferring reaction networks using perturbation data. Bio arXiv (2018)
2. Feillet, C., Krusche, P., Tamanini, F., Janssens, R.C., Downey, M.J., Martin, P., Teboul, M., Saito, S., Lévi, F., Bretschneider, T., van der Horst, G.T.J., Delaunay, F., Rand, D.A.: Phase locking and multiple oscillating attractors for the coupled mammalian clock and cell cycle. Proceedings of the National Academy of Sciences of the United States of America 111(27), 9928–9833 (2014)
3. Martinelli, J., Grignard, J., Soliman, S., Fages, F.: A statistical unsupervised learning algorithm for inferring reaction networks from time series data. In: ICML Workshop on Computational Biology. Long Beach, CA, USA (Jun 2019)
4. Pearl, J.: Causality: Models, Reasoning and Inference. Cambridge University Press, New York, NY, USA, 2nd edn. (2009)
5. Tyson, J.J.: Modeling the cell division cycle: cdc2 and cyclin interactions. Proceedings of the National Academy of Sciences 88(16), 7328–7332 (Aug 1991)