

Inferring Update Sequences in Boolean Gene Regulatory Networks

Fabien Tarissan^a Camilo La Rota^b

^a*Complex System Institute (ISC) & CNRS, Palaiseau, France*

^b*Complex System Institute (IXXI), Lyon, France*

Key words: Mathematical programming, Inverse problems, Gene regulatory networks reconstruction

1 Introduction

This paper employs mathematical programming and mixed integer linear programming techniques for solving a problem arising in the study of genetic regulatory networks. More precisely, we solve the inverse problem consisting in the determination of the sequence of updates in the digraph representing the gene regulatory network (GRN) of *Arabidopsis thaliana* in such a way that the generated gene activity is as close as possible to the observed data.

Differences among cells of different tissues depend on the specific set of genes that are active in each tissue. Therefore, one usually assumes that the different steady states of a GRN dynamics correspond to the different possible cell fates ([7]). This leads to explain the changes observed during the development of the organisms by the fact that perturbations on specific elements of the network make the system switch from one steady state to another. Some hypothesis can be made about these perturbations, which are then treated as initial conditions for the new tissue being formed. However, an important unknown is (are) the update sequence(s) of the gene activity that let the system evolve from a given set of initial conditions to the set of steady states. Indeed, different update sequences determine different sets of basins of attraction of the GRNs. However, the steady states remain the same under any sequence.

Usually, a specific update sequence is assumed to rule the dynamics of the GRNs [1,3]. The present study differs from this approach in that we sought to *infer* the update sequence from the biological observations. It also differs from our previous paper as we focus here on asynchronous sequences whereas in [6] the updates were synchronous.

Email addresses: tarissan@lix.polytechnique.fr (Fabien Tarissan), camilo.larota@ens-lyon.fr (Camilo La Rota).

2 The problem

Given a directed graph $G = (V, E)$, a discrete set T of time instants (which we suppose to be an initial contiguous proper subset of \mathbb{N}) and the following functions:

- a function $\alpha : E \rightarrow \{+1, -1\}$ called the *arc sign function*;
- a function $\omega : E \mapsto \mathbb{R}_+$ called the *arc weight function*;
- a function $\chi : V \times T \mapsto \{0, 1\}$ called the *gene state function*;
- a function $\iota : V \mapsto \{0, 1\}$ called the *initial configuration*;
- a function $\theta : V \mapsto \mathbb{R}$ called the *threshold function*;
- a function $\gamma : V \times T \mapsto \{0, 1\}$ called the *updating function*.

A *gene regulatory network* (GRN) is a 8-tuple $(G, T, \alpha, \omega, \theta, \chi, \iota, \gamma)$ such that:

$$\forall v \in V \quad \chi(v, 0) = \iota(v) \quad (1)$$

$$\forall v \in V, t \in T \setminus \{0\} \quad \chi(v, t) = \begin{cases} H(v, t-1) & \text{if } \gamma(v, t) = 1 \\ \chi(v, t-1) & \text{otherwise} \end{cases} \quad (2)$$

where H is the *Heaviside* function defined for $v \in V$ and $t \in T$ by

$$H(v, t) = \begin{cases} 1 & \text{if } \sum_{u \in \delta^-(v)} \alpha(u, v) \omega(u, v) \chi(u, t) \geq \theta(v) \\ 0 & \text{otherwise,} \end{cases} \quad (3)$$

with $\delta^-(v) = \{u \in V \mid (u, v) \in E\}$ for all $v \in V$. Eqns. (1)-(2)-(3) together are called the *evolution rules* of the GRN. For any particular $t \in T$, $\chi(\cdot, t) : V \rightarrow \{0, 1\}$ is called a *configuration*. Since the evolution rules relate a configuration at time t with a configuration at time $t-1$, $\chi(\cdot, t)$ is called a *fixed configuration* (or fixed point) if it remains invariant under the application of one complete cycle of updates encoded by γ . Furthermore, as long as the evolution rules are purely deterministic (as is modelled above), a fixed point of a GRN is determined by its initial configuration.

In this paper we deal with an inverse problem related to the estimation of update sequence in GRNs. More precisely, we address the following.

UPDATE SEQUENCE ESTIMATION IN GRNs (USEGRN). Given a digraph G , a time instant set T , an arc sign function α , an arc weight function ω , a threshold function θ and a set I of initial configurations, find an update function γ with the property that for all $\iota \in I$ there exists a gene activation function χ such that $(G, T, \alpha, \omega, \theta, \chi, \iota, \gamma)$ are GRNs whose fixed points are at a minimum distance to observed data.

In other words, we attempt to estimate the sequence of updates in a GRN from the knowledge of the digraph topology in such a way that (a) the GRN

evolution rules are consistent with respect to a certain set of initial configurations and (b) the fixed points induced by the estimated values are as close as possible to the observed ones.

As the reader might notice, the problem strongly depends on the modelling of the update sequence encoded by γ . In [3], the authors proposed to describe such a sequence by means of *periods* and *delays* parameters for each gene. Assuming p_v and d_v to be such values for gene v , we can reformulate Equation 2 in the previous modellisation according to the following relation:

$$\forall t \in T \quad \gamma(v, t) = 1 \iff \exists n \in \mathbb{N} \text{ s.t. } t = np_v + d_v$$

3 The mathematical programming formulation

The methodology we shall follow is that of modelling the USEGRN by means of a mathematical programming formulation:

$$\left. \begin{array}{l} \min_x f(x) \\ \text{subject to } g(x) \leq 0, \end{array} \right\}$$

where $x \in \mathbb{R}^n$ are the *decision variables* and $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is the *objective function* to be minimized subject to a set of constraints $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ which may also include variable ranges or integrality constraints on the variables.

The primary concern in solving the USEGRN is thus modellistic rather than algorithmic. One of the foremost difficulties is that of employing a static modelling paradigm — such as mathematical programming — in order to describe a problem whose very definition depends on time. Another important difficulty resides in describing the necessary and sufficient conditions for a configuration to be a fixed point in a mathematical form. We solve this difficulty by introducing two decision variables: a binary variable s stating that the network has been stable for at least two successive time steps; a binary variable y that will indicate the first time the network is stable. The last difficulty concerns the proper modelling of the update sequence as proposed in [3]. The solution relies on the use of two binary variables π and δ for each gene and indexed over the possible values for the periods and the delays. Then, $\pi_v(p)$ (resp. $\delta_v(d)$) is set to 1 if the period (resp. delay) of v is p (resp. d). We provide above such a formulation:

- *Sets*: V of genes in the network, E of edges in the network, T of time instants, P of periods values, D of delay values and R of regions.
- *Parameters*:
 - $\iota : R \times V \mapsto \{0, 1\}$ is the initial configuration of the network (vector of boolean values affected to the genes) for each region.
 - $\alpha : A \rightarrow \{+1, -1\}$ is the sign of the arc weights;
 - $w : V \mapsto \mathbb{R}_+$ is the arc weight function;
 - $\theta : V \mapsto \mathbb{R}$ is the threshold function;

- $\phi : V \times R \mapsto \{0, 1\}$ is the targeted fixed configuration for region r .
- *Variables:*
 - for all $r \in R, v \in V, t \in T, x_{r,v}^t \in \{0, 1\}$ is the activation state of gene v at time t in region r ;
 - for all $r \in R, v \in V, t \in T, h_{r,v}^t \in \{0, 1\}$ is the projection of state of gene v at time t in region r according to Heaviside function;
 - $s : R \times T \mapsto \{0, 1\}$ is a decision variable indicating that the network is stable during at least two successive time steps in region r .
 - $y : R \times T \mapsto \{0, 1\}$ is a decision variable that indicates the first time the network reaches a stable state in region r .
 - for all $v \in V, p \in P, \pi_{v,p} \in \{0, 1\}$ is a decision variable that indicates that the periodicity of gene v is p .
 - for all $v \in V, d \in D, \delta_{v,d} \in \{0, 1\}$ is a decision variable that indicates that the delay of gene v is d .
- *Objective function:*

$$\min \sum_{r \in R} \sum_{t \in T \setminus \{1\}} \left((y_r^{t-1} - y_r^t) \sum_{v \in V} |x_{r,v}^t - \phi_{r,v}| \right).$$

- *Constraints:*
 - Heaviside function computation rule (for all $t \in T \setminus \{1\}, v \in V, r \in R$):

$$\theta_v h_{r,v}^t - |V|(1 - h_{r,v}^t) \leq \sum_{u \in \delta^-(v)} \alpha_{uv} w_{uv} x_{r,u}^{t-1} \leq (\theta_v - 1)(1 - h_{r,v}^t) + |V|h_{r,v}^t$$

- state transition rules (for all $r \in R, v \in V, p \in P, d \in D$):

$$\begin{aligned} x_{r,v}^0 &= \iota_{r,v} \\ \forall t \in T \setminus \{1\} \text{ s.t. } t \neq np + d & \quad \pi_{v,p} \delta_{v,d} x_{r,v}^t = \pi_{v,p} \delta_{v,d} x_{r,v}^{t-1} \\ \forall t \in T \setminus \{1\} \text{ s.t. } t = np + d & \quad \pi_{v,p} \delta_{v,d} x_{r,v}^t = \pi_{v,p} \delta_{v,d} h_{r,v}^{t-1} \\ & \quad \pi_{v,p} \delta_{v,d} d \leq p \end{aligned}$$

- fixed point conditions (for all $r \in R, t \in T \setminus \{1\}$):

$$\begin{aligned} \sum_{v \in V} |x_{r,v}^t - x_{r,v}^{t-1}| &\leq \|V\| s_r^t & y_r^t f_r^t &= 0 & (1 - y_r^t) &\leq f_r^t \\ \sum_{v \in V} |x_{r,v}^t - x_{r,v}^{t-1}| &\geq s_r^t & \sum_{u > t} s_r^u &= f_r^t & (|P| + |D|)^2 &\leq \sum_{\tau \in T} y_r^\tau \end{aligned}$$

4 Reformulations and solutions

The above problem is a nonconvex Mixed-Integer Non-Linear Problem that can be reformulated exactly to a Mixed-Integer Linear Problem using the techniques proposed in [5]. After standard mathematical manipulations, all the nonlinearities reduce to product terms of binary and/or integer variables, which can be reformulated by adding new auxiliary variables and constraints

as follows:

xy terms (x, y : binary)	xz terms (x : binary, z : integer)
$\eta \geq 0$	$\zeta \geq z^L x$
$\eta \leq y$	$\zeta \leq z + (z^L + z^U)(1 - x)$
$\eta \leq x$	$\zeta \leq z^U x$
$\eta \geq x + y - 1$	$\zeta \geq z - (z^L + z^U)(1 - x)$

where z^L and z^U stand for the boundaries of z and η and ζ are the new variables that replace the products in the equations.

We solved to optimality a few real-life instances from the GRN of *Arabidopsis thaliana* using AMPL [2] to model the problem and CPLEX [4] to solve it. The size of the GRNs involved were such that CPLEX obtained the optimal solution in a matter of minutes.

5 Acknowledgements

This work was supported by EU FP6 FET Project MORPHEX. We are deeply grateful to L. Liberti (LIX) for his feedback on the mathematical formulation.

References

- [1] J. Demongeot, A. Elena, and S. Sené. Robustness in Regulatory Networks: a Multi-Disciplinary Approach. *Acta Biotheoretica*, 56(1-2):27–49, 2008.
- [2] R. Fourer and D. Gay. *The AMPL Book*. Duxbury Press, Pacific Grove, 2002.
- [3] Carlos Gershenson. Classification of random boolean networks. In Abbass Standish and Bedau, editors, *Artificial Life VIII, Proceedings of the Eighth International Conference on Artificial Life*, pages 1–8. MIT Press, 2002.
- [4] ILOG. *ILOG CPLEX 10.1 User's Manual*. ILOG S.A., Gentilly, France, 2006.
- [5] L. Liberti, S. Cafieri, and F. Tarissan. Reformulations in mathematical programming: A computational approach. In A. Abraham, A.-E. Hassanien, and P. Suarry, editors, *Global Optimization: Theoretical Foundations and Application*, Studies in Computational Intelligence. Springer, New York, to appear.
- [6] F. Tarissan, C. La Rota, and L. Liberti. Network reconstruction: a mathematical programming approach. In *Proceedings of the European Conference of Complex Systems (ECCS'08)*, to appear.
- [7] R Thomas and M Kaufman. Multistationarity, the basis of cell differentiation and memory. i. structural conditions of multistationarity and other nontrivial behavior. *Chaos*, 11(1):170–179, Mar 2001.