INFERRING PARAMETERS IN GENETIC REGULATORY NETWORKS

Camilo La Rota*, Fabien Tarissan†, Leo Liberti†

*Complex System Institute, Lyon, France
e-mail: camilo.larota@ens-lyon.fr

†École Polytechnique, Palaiseau, France
e-mail: {tarissan,liberti}@lix.polytechnique.fr

1 Introduction

We formulate and solve the problem of determining the arc weights in a Genetic Regulatory Network (GRN) so that it presents expected transitory stable states.

Complex biological regulatory networks determine the function and some observed emergent phenomena in most biological systems. In particular, molecular regulatory networks (proteins, mRNAs, hormones, etc) underlie morphogenesis and physiological phenomena. It is therefore very important to understand the dynamics of these networks in order to infer the molecular mechanisms regulating plant and animal development. In the present work, we use experimental data to reconstruct various aspects of the GRN acting in each cell. More precisely, we attempt to determine some possible scenarios of development together with candidate gene sub-networks that we assume can partially explain the expression patterns observed at each development stage. The estimated values of the kinetic parameters (arc weights) must be such that the stable states (fixed points with respect to the network evolution dynamics) of each sub-network reflect the spatial patterns of gene expression (molecule presence or absence) observed at each development stage.

In particular, we apply this approach to the inference of the regulatory network controlling the early stages of carpel development in the model plant *Arabidopsis thaliana* and we show some results of this work. This network is still poorly understood, even if data concerning genes and morphogenes involved, single regulatory interactions and small circuits exist in the literature [1].

2 Modelling the problem

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

- a function $\alpha : A \to \{+1, -1\}$ called the *arc sign function*;
- a function $\omega : A \to \mathbb{R}_+$ called the *arc weight function*;
- a function $\gamma : V \times T \to \{0, 1\}$ called the *gene activation function*;
- a function $\iota : V \to \{0, 1\}$ called the *initial configuration*;
- a function $\theta : V \to \mathbb{R}_+$ called the *threshold function*,

A *gene regulatory network* (GRN) is a 7-tuple $(G, T, \alpha, \omega, \gamma, \iota, \theta)$ such that:

$$\forall v \in V \quad \gamma(v, 1) = \iota(v) \tag{1}$$

$$\forall v \in V, t \in T \smallsetminus \{1\} \quad \gamma(v, t) = \begin{cases} 1 & \text{if } \sum_{u \in \delta^-(v)} \alpha(u, v)\omega(u, v)\gamma(u, t-1) \geq \theta(v) \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $\delta^-(v) = \{u \in V \mid (u, v) \in A\}$ for all $v \in V$. Eqns. (1)-(2) together are called the *evolution rules* of the GRN. For any particular $t \in T$, $\gamma(\cdot, t) : V \to \{0, 1\}$ is called a *configuration*. Since the evolution rules relate a configuration at time t with a configuration at time $t - 1$, if $\gamma(\cdot, t) = \gamma(\cdot, t - 1)$ then $\gamma(\cdot, t') = \gamma(\cdot, t)$ for all $t' > t$: such configurations are called *fixed points* of the GRN. 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 relating to the reconstruction of a stable subnetwork in GRNs from observed data. More precisely, we address the following.

> Stable Subnetwork Reconstruction in GRNs (SSRGRN). Given a digraph $G = (V, A)$, a time instant set T, an arc sign function α, an initial configuration $C_0 : V \to \{0, 1\}$, a set $U \subseteq V$ determining the nodes of the (induced) subnetwork of G to be reconstructed and the stable configuration $S : U \to \{0, 1\}$ sought for the subnetwork, find an arc weight function ω and a threshold function θ with the property that for all $\iota \in C$ there exists a gene activation function γ such that $(G, T, \alpha, \omega, \gamma, \iota, \theta)$ is a GRN whose fixed points are at a minimum distance to observed data.

In other words, we attempt to estimate the arc weights and threshold functions of a GRN from the knowledge of the digraph topology G, the induced subnetwork topology U and the arc sign function α 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 in the subnetwork induced by the estimated values are as close as possible to the given ones.

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

$$\left.\begin{array}{rll} \min_x & f(x) & \\ \text{subject to} & g(x) & \leq \quad 0, \end{array}\right\} \tag{3}$$

where $x \in \mathbb{R}^n$ are the *decision variables* and $f : \mathbb{R}^n \to \mathbb{R}$ is the *objective function* to be minimized subject to a set of constraints $g : \mathbb{R}^n \to \mathbb{R}^m$ which may also include variable ranges or integrality constraints on the variables. As shown below, the SSRGRN is a Mixed-Integer Nonlinear Program (MINLP) which can be reformulated to a Mixed-Integer Linear Program (MILP).

Our primary concern in solving the SSRGRN 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 expressing the necessary and sufficient conditions for a configuration to be a fixed point as constraints suitable for use in a formulation like (3).

3 Mathematical programming formulation

- *Sets*:

  1. set V of genes in the network;
  2. set A of arcs in the network;
  3. set T of time instants.
  4. set $U \in V$ of genes involved in the stable sub-network.

- *Parameters*:

  1. $C_0 : V \to \{0, 1\}$ is the initial configuration of the network (vector of boolean values assigned to the genes).
  2. $\alpha : A \to \{+1, -1\}$ is the sign of the arc weights;
  3. θ^L, θ^U are the bounds on the threshold values;
  4. $S : U \to \{0, 1\}$ is the targeted configuration of the sub-network composed by U.

- *Variables*:

1. for all $i \in V$, $t \in T$, $x_i^t \in \{0,1\}$ is the activation status of gene i at time t;
2. $s : T \to \{0,1\}$ is a decision variable indicating that the sub-network is stable during at least two successive time steps.
3. $y : T \to \{0,1\}$ is a decision variable that indicates the first time the sub-network reaches a stable state.
4. $\theta : V \to \mathbb{R}$ is the threshold function;
5. $w : A \to \mathbb{R}_+$ is the arc weight function.

- *Objective function*:

$$\min \sum_{t \in T \setminus \{1\}} \big((y^{t-1} - y^t) \sum_{u \in U} |x_u^t - S_u|\big).$$

- *Constraints*:

1. evolution rule:

$$\begin{aligned} &\forall t \in T \setminus \{1\}, v \in V \quad \theta_v \\ &\quad x_v^t - |V|(1 - x_v^t) \quad \leq \quad \sum_{u \in \delta-(v)} \alpha_{uv} w_{uv} x_u^{t-1} \leq \theta_v (1 - x_v^t) - |V| x_v^t \end{aligned} \tag{4}$$

2. fixed point conditions:

$$\forall t \in T \setminus \{1\} \quad \sum_{u \in U} |x_u^t - x_u^{t-1}| \quad \leq \quad |U| s^t \tag{5}$$

$$\forall t \in T \setminus \{1\} \quad s^t \quad \leq \quad \sum_{u \in U} |x_u^t - x_u^{t-1}| \tag{6}$$

$$\forall t \in T \setminus \{1, |T|\} \quad y^t \quad = \quad s^t y^{t-1} \tag{7}$$

$$\forall t \in T \setminus \{1, |T|\} \quad \sum_{r > t} y^r \quad \leq \quad (|T| - t) y^t; \tag{8}$$

3. boundary conditions:

$$\forall v \in V \quad x_v^0 = C_0(v).$$

4 Reformulation and solution

The above problem is a nonconvex MINLP that can be reformulated exactly to a MILP using the techniques proposed in [4]. We solved to optimality a few real-life instances from the GRN of *Arabidopsis thaliana* using AMPL [2] to model the problem and CPLEX [3] to solve it. The size of the GRNs involved were such that CPLEX obtained the optimal solution in a matter of seconds and/or minutes.

Acknowledgments

We are deeply grateful to Françoise Monéger and Jan Traas from the RDP-ENS-Lyon institute for having provided the data on which we tested our approach.

REFERENCES

[1] V. Balanzá, M. Navarrete, M. Trigueros, and C. Ferrándiz. Patterning the female side of arabidopsis: the importance of hormones. *Journal of Experimental Botany*, 57:3457–3469, 2006.

[2] R. Fourer and D. Gay. *The AMPL Book*. Duxbury Press, Pacific Grove, 2002.

[3] ILOG. *ILOG CPLEX 10.1 User's Manual*. ILOG S.A., Gentilly, France, 2006.

[4] L. Liberti. Reformulation techniques in mathematical programming, November 2007. Thèse d'Habilitation à Diriger des Recherches.