Identification of Qualitative Genetic Regulatory Network Models by a Mathematical Programming Approach

Camilo La Rota & Fabien Tarissan

IXXI – LIP CNRS UMR5668 - ENS-Lyon
Lyon, France
camilo.larota@ens-lyon.fr

ISC – CNRS LIX - École Polytechnique
Palaiseau, France
Fabien.Tarissan@lip6.fr

Abstract

We propose a method of Genetic Regulatory Network (GRN) model identification using mathematical programming and global optimization techniques. The problem consists in the estimation of the unknown parameters of a GRN model such that the asymptotic dynamics of the model closely match a set of experimental observations. This problem can be naturally cast as an optimization problem that minimizes a given distance between a set of observed expression patterns and estimated values of the parameters, subject to constraints derived from the algebraic equations that describe the dynamics of the biological system. We apply this approach to the inference of GRNs controlling early flower organ development in the model plant Arabidopsis thaliana.

1. Introduction

GRNs in each cell control morphogenesis, their **reconstruction** is needed in order to understand the molecular mechanisms underlying development.

Spatial gene expression data are available for many genes at several stages of flower development (*in-situ* hybridization); they provide important information about gene activities but suffer from several limitations:

- Noisy, unreliable
- Poor precision
- Poor temporal resolution

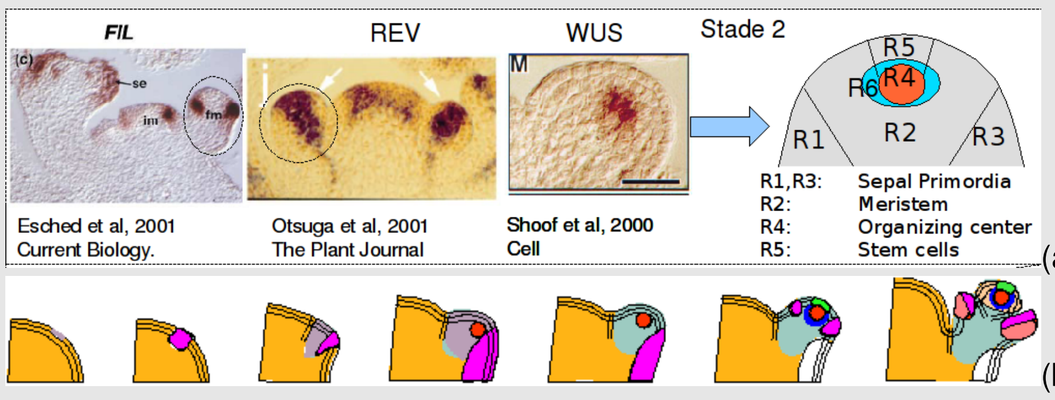

Figure 1: *(a) Spatial* in-situ *gene expression maps at different stages are interpreted into qualitative expression images, superposition reveals regions with different expression patterns;(b) different regions of gene expression appear, disappear and evolve during development.*

Expression data interpreted as (multi-level or binary activities) **qualitative** activities, superposition gives tentative **regions** of homogeneous gene expression

Gene interaction data listed from diverse experimental evidences are used to propose candidate network interaction structures. Suffer from:

- Uncertain, unreliable
- Contradictory
- Unrelated sources

A qualitative GRN model is used to describe gene activation mechanisms; its **steady states** must reflect homogeneous regions of gene expression.

The GRN model identification problem (GRNMI) consists in the estimation of the unknown parameters of the GRN model such that the asymptotic dynamics of the model closely match the interpreted qualitative expression data, subject to structural and dynamical constraints. At the same time, identification results are used to test different GRN interaction structure hypotheses against observed expression patterns.

2. The method

The GRN conceptual model: Let the directed graph $G = (V, A)$ be the interaction structure, V being the set of genes and A the set of pair-wise interactions, $w : A \mapsto \mathbb{R}_+$, $\alpha : A \mapsto \{-1, 1\}$ and $\theta : A \mapsto \mathbb{R}$, denote the interaction strengths, types and thresholds, respectively, and $\delta^-(i) = \{j \in V | (j, i) \in A\}$ and $\delta^+(i) = \{j \in V | (i, j) \in A\}$ the input and output neighborhoods. The qualitative activity state of gene i is denoted by $q_i \in P_i$ where $P_i = \{1 \dots Q_i\}$ and $Q_i \in \mathbb{N}$ is the number of interactions with different threshold values having gene i as source. The GRN verifies the state transition condition:

$$q_i^{n+1} = \sum_{p_i \in P_i} H\left(\sum_{j \in \delta^-(i)} \alpha_{ji} w_{ji} H(q_j^n, \pi_j(i)), \theta_{i,\pi^{-1}(p_i)} \right), \quad (1)$$

$H(x, \theta) = \{1 \text{ if } x \geq \theta \text{ , } 0 \text{ if } x < \theta\}$ and $\pi_i : \delta^+(i) \mapsto P_i$ relates gene target indexation with threshold magnitude ordering such that if $\pi_i(j) < \pi_i(k)$ then $\theta_{ij} < \theta_{ik}$.

The GRN Model identification problem (GRNMIP):

$$\min_x f(x), \quad \text{subject to } g(x) \leq 0, \quad (2)$$

where $x \in \mathbb{R}^n$ are *decision variables*, $f : \mathbb{R}^n \mapsto \mathbb{R}$ is the *objective function* and $g : \mathbb{R}^n \mapsto \mathbb{R}^m$ is a set of constraints including variable ranges and integrality constraints.

Given G, α and the sought-for steady configurations $\phi = \{\phi_i\}_{i \in V}, \phi_i \in P_i$, given by the expression data, the GRNMIP consists in finding w and θ such that $(G, \alpha, w, q, \theta)$ collectively minimize the hamming distance ($D_H(q, \phi) = \sum_{i \in V} |q_i - \phi_i|$) from *steady states* to data and satisfy the set of *steadiness constraints*:

$$\begin{aligned} \forall i \in V \text{ and } j \in \delta^-(i): \quad q_i &= \sum_{l \in P_i} f_i(l) \\ \pi_j(i) h_{ji} - |V|(1 - h_{ji}) &\leq q_j \leq (\pi_j(i) - 1)(1 - h_{ji}) + |V| h_{ji} \\ \theta_{i,\pi_i^{-1}(i)} f_i(l) - |V|(1 - f_i(l)) &\leq \sum_{j \in \delta^-(i)} \alpha_{ji} w_{ji} h_{ji} \leq (\theta_{i,\pi_i^{-1}(i)} - \epsilon)(1 - f_i(l)) + |V| f_i(l) \end{aligned} \quad (3)$$

Reformulation and Solution: the GRNMIP is a non-convex Mixed Integer Nonlinear Program (MINLP); after standard mathematical manipulations non-linearities reduce to product terms of binary and/or real variables that can be reformulated exactly to a linear one (MILP) [**?**]. We use the modelling language AMPL [**?**] to describe the reformulated problem and solve it with CPLEX [**?**].

The same methodology is used to model other GRNMIP using different reconstruction criteria and/or gene activity conceptual models; in the particular case where the same conceptual model is used but genes states take binary qualitative values, a similar formalization applies but the constraints are simplified.

3. Results

Early stages of plant lateral organ development The method is used to reconstruct GRNs controlling early flower organ development in *Arabidopsis*. A qualitative boolean model for lateral organ polarity has been identified for which the steady states match precisely observed expression data.

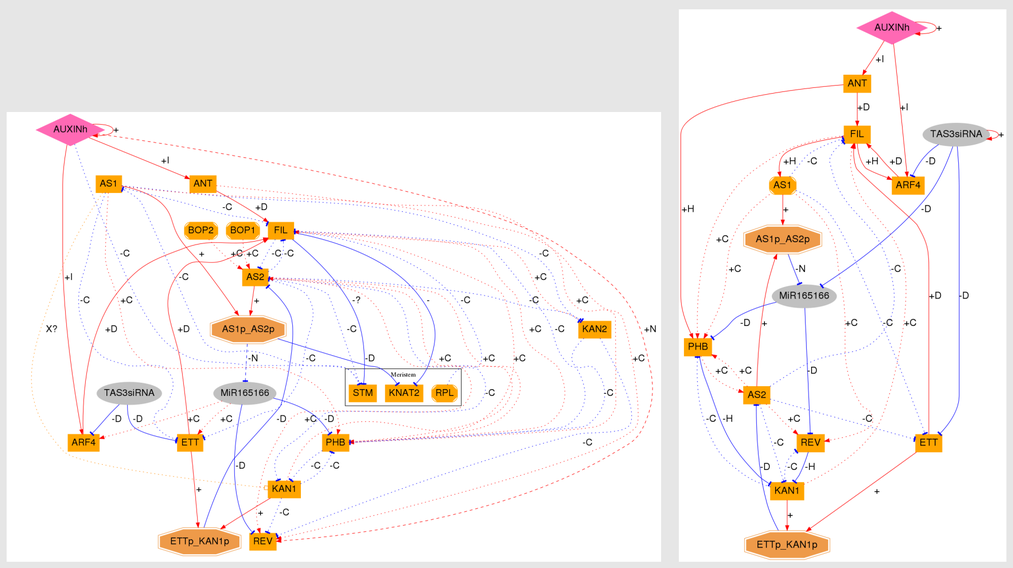

Figure 2: *(a) Interaction data for lateral organ primordia; (b) Proposal of an identified GRN model for polarity of lateral organ primordia whose steady states match expression data*

4. Discussion

The approach we propose focuses on problem modelling rather than in algorithmics and has the advantage of generality. The identification method presented doesn't depend on the particular form of the GRN model used here and can be extended to more intricate gene activation mechanisms. Different inverse problems can be modelled, reformulated and solved with reliable and fast general-purpose algorithms.

ECCS 2009, European Conference on Complex Systems, 21-25 September 2009, University of Warwick, Coventry, UK