Mathematical programming techniques applied to biology

Fabien Tarissan[1]

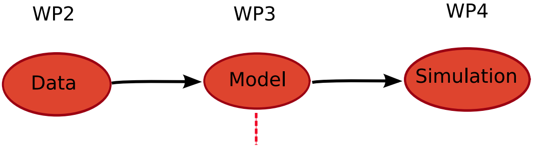
Leo Liberti[2] Camilo La Rota[3]

[1] ISC-PIF (Paris, France)
[2] École Polytechnique (Paris, France)
[3] IXXI (Lyon, France)

October 31, 2008

Context of work

Pre-simulation tool for the Morphex european project:

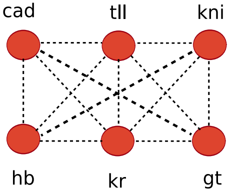


How to propose a **first** possible model

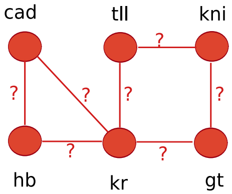
Heterogeity at many levels:

- organisms
- data
- reliability
- level of details
- . . .

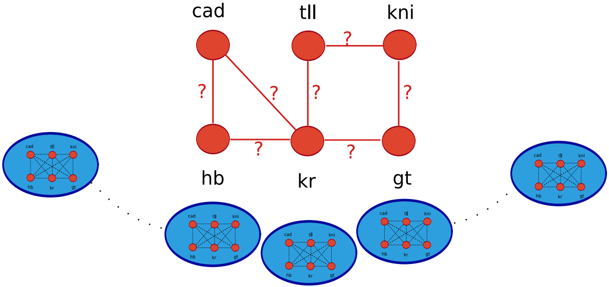
Network reconstruction



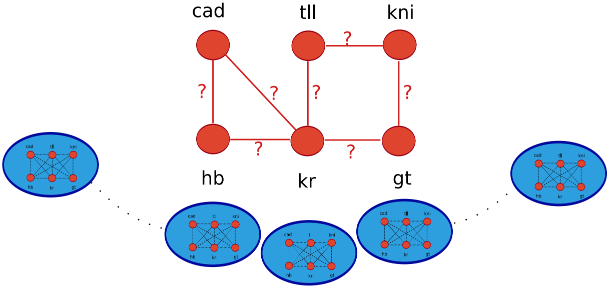
Network reconstruction



Network reconstruction



Network reconstruction



Our approach:

- Modelisation by means of mathematical programming techniques (constraints)
- Reformulation of the models in order to ease the solving

Contributions :

- Reconstruction of gene regulatory networks:
  - with continuous dynamics (drosophila)
  - with discrete dynamics (arabidopsis)

Mathematical Programming

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

- $x \in \mathbb{R}^n$ are the decision variables
- $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is the objective function
- $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is the set of constraints

$+$ distinction between integer and continuous variables.
Let $Z \in \{1, \ldots, n\}$ such that $\forall i \in Z$, $x_i \in \mathbb{Z}$.

Classes of problems

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

AMPL: A Mathematical Programming Language.

Class	f, g	Z	**Best solver**	**Best free solver**	**Complexity**
LP	linear	$Z = \emptyset$	CPLEX	CLP	$\Theta(10^6)$
cNLP	convex	$Z = \emptyset$	SNOPT/FILTER	IPOPT	$\Theta(10^4)$
MILP	linear	$Z \neq \emptyset$	CPLEX	BCP/SYMPHONY	$\Theta(10^3)$
NLP	non linear	$Z = \emptyset$	BARON	?	$\Theta(10^2)$
cMINLP	convex	$Z \neq \emptyset$	MINLP_bb/FILMINT	BONMIN/FILMINT	$\Theta(10^3)$
MINLP	non linear	$Z \neq \emptyset$	BARON	?	$\Theta(10^2)$

Application to the drosophila model

Continuous regulation of gene products concentrations:

$$\frac{dg_{ia}(t)}{dt} = R_a\Phi(u_{ia}(t)) - \lambda_a g_{ia}(t) + D_a(g_{i+1,a}(t) - 2g_{ia}(t) + g_{i-1,a}(t))$$

- $g_{ia}(t)$ is the concentration of gene a in nucleus i at time t
- R_a is the production rate for gene a
- Φ is the sigmoid regulation function
- λ_a is the decay rate
- D_a is the diffusion coefficient for gene a

Regulation term

The sigmoid definition:

$$\Phi(u) = \frac{1}{2}\left(\frac{u}{\sqrt{u^2+1}} + 1\right)$$

Relies on:

$$u_{ia}(t) = \sum_{b \in N^\gamma} W_{ba} g_{ib}(t) + m_a g_i^{\mathtt{bcd}} + h_a$$

- W_{ba} is the weight on the arc (b, a) in the GRN
- m_a is the regulatory influence of the maternal gene `bcd`
- h_a is the activation threshold for Φ

The problem

Size of the problem:

- Network of 6 genes
- but missing values for W, R, D, m, λ, h : 66 variables.

Confronting the estimation to the observed data:

$$\min \sum_{i \in N^{\iota}} \sum_{t} (g_{ia}(t) - g_{ia}^{\text{data}}(t))^2 + \Pi_R + \Pi_\lambda + \Pi_D + \Pi_u$$

Penalty function:

$$\Pi_u = e^{\Theta} - 1$$

$$\Theta = \Lambda(\sum_{(b,a) \in A} (W_{ba} v_b^{\text{max}})^2 + (m_a v_{\text{bcd}}^{\text{max}})^2 + h_a^2)$$

Modelling in AMPL

1. Translating the model into AMPL:
   - Objective function:

$$\min \sum_{\substack{a\in N^{\gamma}\\ i\in N^{\iota}\\ t\in T^{\text{data}}}} (\mathrm{g}_i^a(t) - \mathrm{g_{data}}{}_i^a(t))^2 + \sum_{\substack{a\in N^{\gamma}\\ b\in N^{\gamma}}} (\mathrm{W}_b^a v_{\max}^b)^2 + \sum_{a\in N^{\gamma}} ((m_a v_{\max}^{\text{bcd}})^2 + h_a^2)$$

   - Some penalty functions as constraints:

$$\forall a \in N^{\gamma} \left\{ \begin{array}{l} R^L \leq R_a \leq R^U \\ \lambda^L \leq \lambda_a \leq \lambda^U \\ D^L \leq D_a \leq D^U \end{array} \right.$$

   - PDE as a constraint (discretization):

$$\mathrm{g}_i^a(t) - \mathrm{g}_i^a(t-1) = \Delta t \left(\frac{R_a}{2}\left(\frac{\mathrm{u}_i^a(t)}{\sqrt{\mathrm{u}_i^a(t)^2+1}} + 1\right) - \lambda_a \mathrm{g}_i^a(t) + D_a(\mathrm{g}_{i+1}^a(t) - 2\mathrm{g}_i^a(t) + \mathrm{g}_{i-1}^a(t)) \right)$$

2. Other issues:
   - Mitosis time
   - Modelling cell division
   - Updating diffusion coefficient
   - ...

Simplifying the model

- Driven by biological knowledge: (e.g. boundaries on W, m and h)
- Mathematical reformulating of terms:
  - exact reformulation: e.g. for $\frac{u}{\sqrt{u^2+1}}$
    1. $z = \frac{1}{\sqrt{u^2+1}} \implies z^2(u^2+1) = 1 \implies (zu)^2 + z^2 = 1$
    2. Let u', u'' and z' be respectively the uz, u'^2 and z^2.
    3. Substitute $\frac{u}{\sqrt{u^2+1}}$ with u' and add constraints:

$$\begin{cases} u' = uz \\ u'' + z' = 1 \\ z' = z^2 \\ u'' = u'^2 \end{cases}$$

  - approximative reformulation of z^2

Work achieved so far

What is done:

1. the raw model (without any reformulation)
2. various reformulations:
   - sigmoid (exact): too many variables.
   - sigmoid (approx): ok.
   - convex products (approx): ok but feasability issues.
3. run on small data set: good results

What will be done:

- run on large data set: too heavy for now (need to split the model).
- trying other modellisations ($g_{ia}(t) = g_{ia}^{\text{data}}(t)$?)

Other case of study: Arabidopsis

Same approach:

- Gene regulatory network
- Some knowledge of the network topology
- Don't know the weight on edges

Different dynamics:

- Descretization of the time
- Qualitative activity of gene i: $x_i^{t+1} = H\left(\sum_{j=1}^{n} \alpha_{ij} w_{ij} x_j^t - \theta_i\right)$
  - θ_i: threshold of activation.
  - w_{ij}: interaction strength $\left(\frac{(induced \quad production)}{decay}\right)$.
  - α_{ij} : Kind of the interaction
    $(repression = -1, \quad activation = +1)$

Similar problem: Find w_{ij} and θ_i

Modelling: defining the GRN

Gene Regulatory Network (GRN): $(G, T, \alpha, w, x, \iota, \theta)$

- Sets and Graph:
  V: vertexes (genes)
  A: arcs (interactions)
  $T \, : =\{1, 2, ..\} \subset \mathbb{N}$
  $G = (V, A)$

- Functions:

$\alpha : A \to \{+1, -1\}$	*arc sign*;
$w : A \to \mathbb{R}_+$	*arc weight*;
$x : V \times T \to \{0, 1\}$	*gene activation*;
$\iota : V \to \{0, 1\}$	*initial configuration*;
$\theta : V \to \mathbb{R}$	*threshold*,

- Evolution rules

$$\begin{aligned} x(v, 1) &= \iota(v) \\ x(v, t) &= \begin{cases} 1 & \text{if } \displaystyle\sum_{u \in \delta^-(v)} \alpha(u, v) w(u, v) x(u, t-1) \geq \theta(v) \\ 0 & \text{otherwise,} \end{cases} \end{aligned}$$

where $\delta^-(v) = \{u \in V \mid (u, v) \in A\}$ for all $v \in V$.

Modelling: defining the problem

Given

- (G, T, α)
- $S := \{1..Smax\}$: set of stages.
- $U = \{U_s\}_{s \in S}; U_s \subseteq V$: nodes of G_s (induced subnetworks of G).
- $I = \{\iota_{s,u}\}_{s \in S, u \in U_s}; \iota_{s,u} : V \to \{0,1\}$: initial conditions.
- $\Phi = \{\phi_{s,u}\}_{s \in S, u \in U_s}; \phi_{s,u} : V \to \{0,1\}$: expression data.

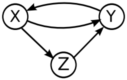
Find

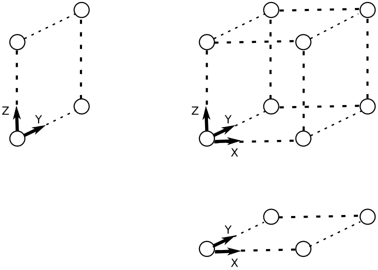
w, θ with the property that $\forall \ \vec{\iota}_s$, $(G_s, T, \alpha, w, \vec{x}_s, \vec{\iota}_s, \theta)$ satisfies the evolution rules and has fixed points that collectively minimize the total $D_H(\rho, \phi)$.

D_H : hamming distance from model fixed points to data.

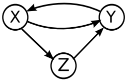
fixed points $(\vec{\rho})$: If $\vec{x}_t = \vec{x}_{t-1} = \vec{\rho}$ then $\vec{x}_{t'} = \vec{x}_t$ for all $t' > t$.

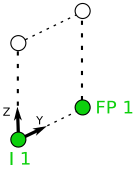
Finding fixed points

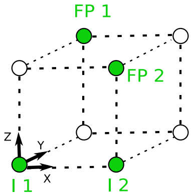


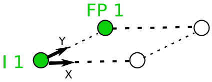


Finding fixed points

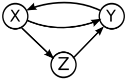


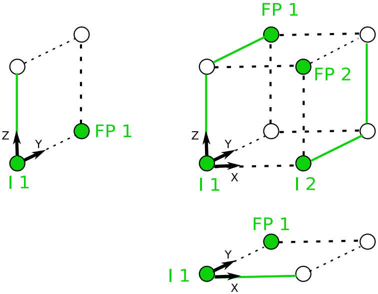






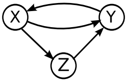
Finding fixed points

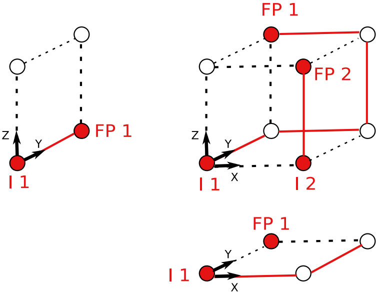




d1 = 5

Finding fixed points





d1 = 5

d2 = 1

Mathematical programming formulation

- Objective function

$$\sum_{s\in S}\sum_{t\in T\setminus 1}(y_{s,t-1}-y_{s,t})\sum_{u\in U_s}|x_{s,u,t}-\rho_{s,u}|$$

- Fixed point conditions

$$\begin{array}{rclcrcl}
\sum_{u\in U_s}|x^t_{s,u}-x^{t-1}_{s,u}| & \leq & \|U_s\|\,\sigma^t_s & \quad & 1-y^t_s & \leq & \sum_{r\geq t}\sigma^t_r \\
\sum_{u\in U_s}|x^t_{s,u}-x^{t-1}_{s,u}| & \geq & \sigma^t_s & \quad & y^t_s\sum_{r\geq t}\sigma^t_r & = & 0
\end{array}$$

- Evolution rules

$$\begin{array}{rcl}
\sum_{u\in U_s:(u,v)\in A}\alpha_{u,v}w_{u,v}x^{t-1}_{s,u} & \geq & \theta_v x^t_{s,v}-\|V\|\,(1-x^t_{s,v}) \\
\sum_{u\in U_s:(u,v)\in A}\alpha_{u,v}w_{u,v}x^{t-1}_{s,u} & \leq & (\theta_v-\epsilon)(1-x^t_{s,v})+\|V\|\,x^t_{s,v}
\end{array}$$

Conclusion on the modelling approach

Static modelling of a dynamic system

A framework for reconstructing regulatory networks:

- of different biological organisms
- with different dynamics

Drawbacks:

- loose of efficiency
- might require to introduce new elements

Perspectives:

- automatization of the reformulations
- study more complex qualitative models of GRN
- integrating different kind of knowledge (experimental, theoretical, . . .)

Automatic (re)formulation

For the modelling part: E.g. 4 “virtual” constraints to express the *fixed point* (should have been generated!)

For the simplification part:

Name	Nonlinear feasible set	Linear feasible set
PowBin exact	$(x_1, x_2) \in \{0,1\} \times \mathbb{R} : x_2 = x_1^n$	$(x_1, x_2) \in \{0,1\} \times \mathbb{R} : x_2 = x_1$
ProdBin exact	$(x, x_{n+1}) \in \{0,1\}^n \times \mathbb{R} : x_{n+1} = \prod\limits_{i \leq n} x_i$	$(x, x_{n+1}) \in \{0,1\}^n \times [0,1]$: $x_{n+1} \leq x_i \quad \forall i \leq n$ $x_{n+1} \geq 1 - n + \sum\limits_{i \leq n} x_i$
ProdBin-Cont exact	$(x_1, x_2, x_3) \in \{0,1\} \times [x_2^L, x_2^U] \times \mathbb{R}$: $x_3 = x_1 x_2$	$(x_1, x_2, x_3) \in \{0,1\} \times [x_2^L, x_2^U]^2$: $x_3 \leq x_2^U x_1$ $x_3 \geq x_2^L x_1$ $x_3 \leq x_2 + x_2^L(1 - x_1)$ $x_3 \geq x_2 - x_2^U(1 - x_1)$

Leads to *Term Rewriting Systems (TRS)* properties:

- termination
- confluence
- optimality?