Identification of genetic network dynamics: A model invalidation approach

Eugenio Cinquemani, IBIS

Workshop on Identification and Control of Biological Interaction Networks
INRIA Grenoble – Rhône-Alpes, February 8, 2011

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
GRENOBLE - RHÔNE-ALPES

Preamble

- Fact: Experimental techniques for quantitative monitoring of gene expression over time enable dynamic modelling
- Goal: Automated procedure for inferring structure and parameters of gene regulatory network models from experimental data
- What models from what data ? And for what purpose ?
  - Kinetic models ? Linear models ? Power laws ? Interaction networks ?
  - Analysis ? Control ? Re-engineering ?
    - ► Define relevant class of models well suited for identification
    - ► Preserve information on gene activation logics
- Structure and parameter estimation are interrelated
  - Parameter estimation is challenging in models of realistic size, cannot do it for every possible model structure
    - ► Exploit *a priori* information on the model structure
    - ► Problem decoupling: eliminate hypotheses without fitting parameters

Outline

- Boolean-like gene regulatory network models
- Models with unate structure
- Identification of ODE models with unate structure
  - From protein concentration and synthesis rate profiles in cell colonies
  - Focus on an interesting subclass of unate models
  - Performance assessment on a test case
- Results on IRMA (yeast synthetic network, Cantone *et al*, *Cell* 2009)
- Conclusions

- With Riccardo Porreca (ETH), John Lygeros (ETH), Giancarlo Ferrari-Trecate (UniPv) (Porreca *et al*, *Bioinformatics* 2010)

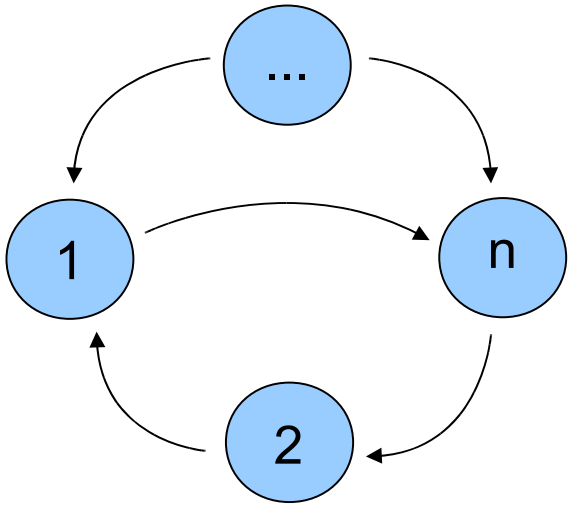
Boolean models

- N Boolean variables representing n genes

$$(X_1, X_2 \ldots, X_n) \in \{0,1\}^n$$

$X_i = 0$ gene not expressed

$X_i = 1$ gene expressed



- Boolean regulation function

X_i expressed iff $b_i(X) = 1$

- Dynamic Boolean networks (discrete time):

$$X_i(t+1) = b_i\big(X(t)\big) \qquad i = 1, \ldots, n \qquad t = 0, 1, 2, \ldots$$

- Can associate regulatory interaction graph
  - n nodes (genes), arcs (incoming arcs of node i = effective inputs of b_i)

Boolean-like ODE models

- Recall Boolean update map:

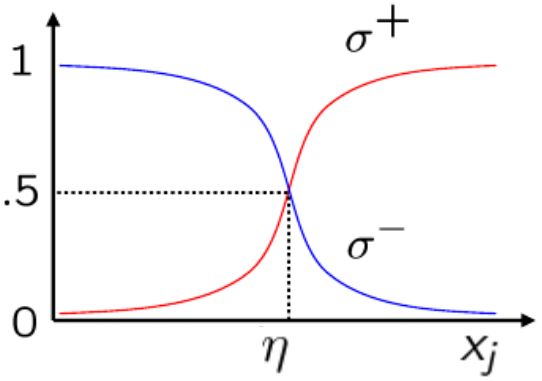
$$X_i^+ = b_i(X), \quad \text{where } b_i = \bigvee_l \bigwedge_j X'_{l,j}, \quad X'_{l,j} \in \{X_j, \neg X_j\}$$

- Algebraic equivalent (Plahte *et al*, *J Math Biol* 1998): apply the transformation

$$\begin{aligned} X_j &\rightarrow \sigma^+(x_j) \\ \neg\, expr(X) &\rightarrow 1 - expr(x) \\ expr(X) \wedge expr'(X) &\rightarrow expr(x) \cdot expr'(x) \end{aligned}$$

$$\sigma^+(x_j) = \frac{x_j^d}{x_j^d + \eta^d}$$

$$\sigma^-(x_j) = 1 - \sigma^+(x_j)$$



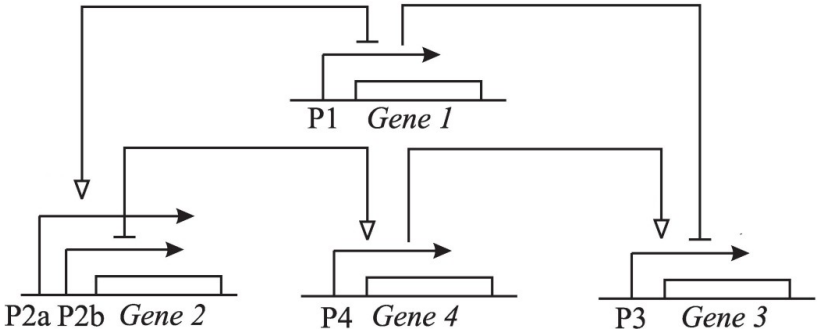
- Boolean-like model: define ODE

$$\dot{x}_i = \kappa_i^1 + \kappa_i^2 b_i(x) - \gamma_i x_i$$

$b_i(x)$ algebraic equivalent of $b_i(X)$

Plausibility ?

- Experimental evidence that *often*
  - Relationships among transcription factor concentrations and transcription rates, as well as post-transcriptional, transport, (and reaction) processes at equilibrium, can be described by sigmoidal functions
  - Nonlinear regulatory effects associated to multiple regulators combine into algebraic expressions (sums and products)



Gene	Expressed when
1	G2 not expressed
2	G1 expressed or G4 not expressed
3	G4 expressed and G1 not expressed
4	G2 expressed

Boolean model

$$b_1(X) = \neg X_2$$
$$b_2(X) = X_1 \vee \neg X_4$$
$$b_3(X) = X_4 \wedge \neg X_1$$
$$b_4(X) = X_2$$

Boolean-like model

$$b_1(x) = \sigma^-(x_2)$$
$$b_2(x) = 1 - \sigma^-(x_1) \cdot \sigma^+(x_4)$$
$$b_3(x) = \sigma^+(x_4) \cdot \sigma^-(x_1)$$
$$b_4(x) = \sigma^+(x_2)$$

- Starting point for biologically relevant and structured quantitative models

Tractability ?

- General Boolean-like model:

$$\dot{x}_i = \kappa_i^1 + \kappa_i^2 b_i(x) - \gamma_i x_i, \qquad \text{where } b_i = \sum_l \prod_j \sigma^{\pm}(x_j|\theta_{l,j})$$

- Structure identification: based on data, decide
  - The number of summands
  - The sigmoids in each product and their sign
- Parameter identification: paramaters of each sigmoid, rates
- Intractable problem: cannot enumerate and fit all model structures!
  - Combinatorial explosion of model alternatives
  - Heavy nonlinear parameter estimation, identifiability issues
- But, good starting point
  - Reduction to specific families of Boolean-like functions
  - Use for approximation of more general nonlinear models

Boolean-like models with unate structure

- Unate functions: Boolean rules monotone in each input variable
  - Transcription factors with unambiguous role (activator XOR repressor)
  - Arguably, the only distinguishable rules (Grefenstette *et al*, *Biosystems* 2006)
  - Includes most known gene activation rules (Nikolajewa *et al*, *Biosystems* 2007)
- Boolean-like formulation: preserves monotonicity properties
  - Model:

$$b_i(x) = \prod_{l=1}^{n_i} \tau_l, \quad \tau_l = 1 - \prod_{j \in J_l} (1 - \sigma^{\pm}(x_j)) \quad \text{where} \quad \sigma^{\pm}(x_j) = \begin{cases} \sigma^+(x_j), & \text{or} \\ \sigma^-(x_j), \end{cases}$$

  - Sign pattern:

$$p = (p_1, \ldots p_n), \qquad p_j = \begin{cases} 1, & \text{if } \sigma^{\pm}(x_j) = \sigma^+(x_j), \\ -1 & \text{if } \sigma^{\pm}(x_j) = \sigma^-(x_j), \quad j = 1, \ldots, n \\ 0 & \text{if } j \notin J_l \ \forall l \end{cases}$$

Example, $p = (-1, 1)$: $\sigma^-(x_1)\sigma^+(x_2)$, $1 - \sigma^+(x_1)\sigma^-(x_2)$, $\sigma^-(x_1)\sigma^+(x_2) + \frac{1}{2}\sigma^+(x_2)$, ...

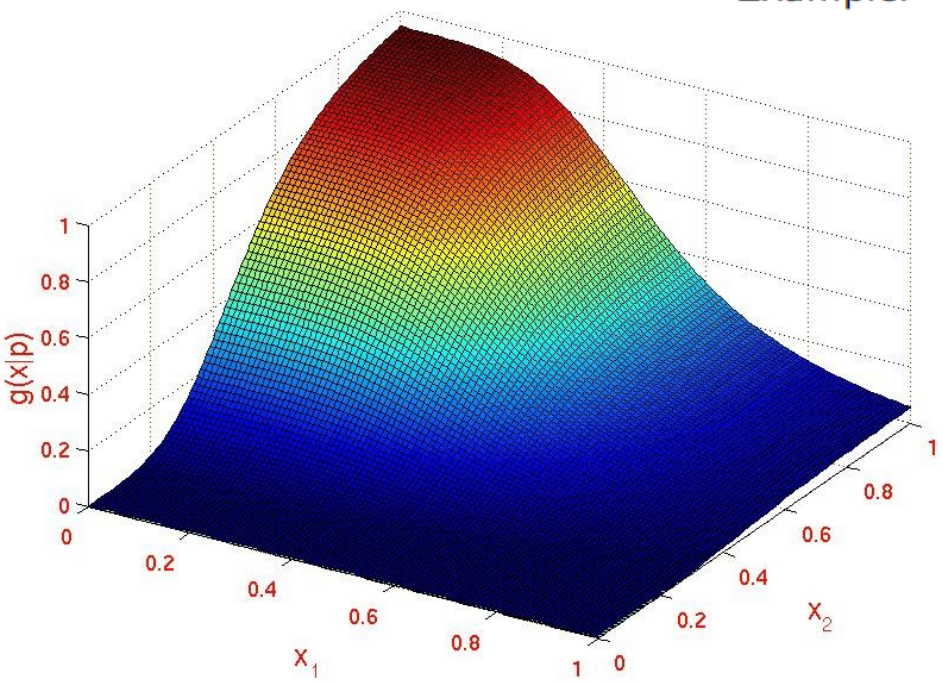
$b_i(x)$ is nondecreasing (resp. nonincreasing) in x_j if $p_j = 1$ (resp. $p_j = -1$)

... and so is any synthesis rate $g_i(x) = \kappa_i^1 + \kappa_i^2 b_i(x)$, provided $\kappa_i^1, \kappa_i^2 \geq 0$

Identification of unate-like models

- Two-step strategy: eliminate bad hypotheses before fitting parameters
- Given: protein concentrations & synthesis rates (recall $\dot{x}_i = g_i(x) - \gamma_i(x)$)
- Step 1: Exploit monotonicity properties to invalidate sign patterns

Example:



Consider $g(x|p)$, $x = (x_1, x_2)$.

Let $p = (p_1, p_2)$ be unknown.

Given data (x, g_i), (x', g_i'), assume

$x_1 > x_1'$, $\quad x_2 < x_2'$, $\quad g_i > g_i'$.

Can exclude:

$p = (-1, 1) = (\text{sign}(x_1' - x_1), \text{sign}(x_2' - x_2))$

and also

$p = (0, 1)$, $\quad p = (-1, 0)$, $\quad p = (0, 0)$

independently of parameter values!

Sign patterns: definitions and properties

- Given data pairs: $(x^1, g^1), \ldots, (x^m, g^m)$, with $g^k = g(x^k|p)$
- Definition: p is *inconsistent* if the property

$$p_j(x_j^k - x_j^l) \geq 0,\ j = 1, \ldots, n \implies g(x^k|p) - g(x^l|p) \geq 0$$

  is falsified for some k,l
- Definition: subpattern and superpattern

		Complexity
Superpatterns	1 1 -1 1 1 1 -1 -1 1 -1 -1 1 1 -1 -1 -1 1 1 -1 0 1 -1 -1 0	4 3
Pattern	1 0 -1 0	2
Subpatterns	1 0 0 0 0 0 -1 0 0 0 0 0	1 0

- Subpatterns of inconsistent patterns are also inconsistent
- Superpatterns of consistent patterns are also consistent
- Minimal consistent and maximal inconsistent patterns exist

Identification of unate-like models cont'd

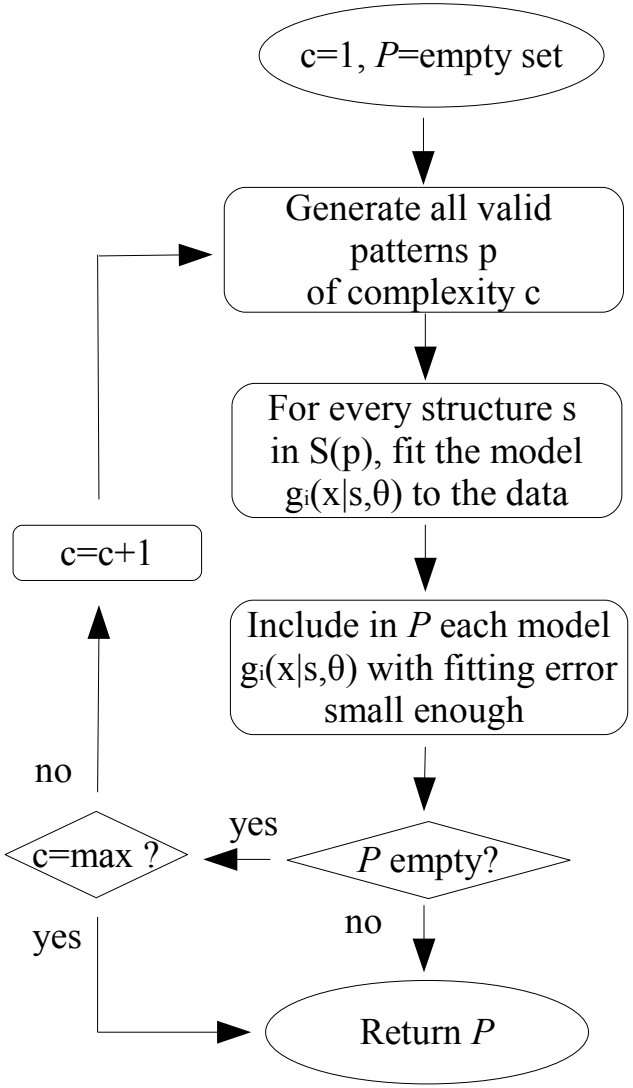
- Step 2: Search best fitting model structure with valid sign pattern
  - Define model structures S(p) of interest
  - Enumerate model structures with valid sign patterns of increasing levels of complexity
  - Stop at the level of complexity where at least one model fits the data "well enough"

Hierarchical search favors simpler models

Example with products of sigmoids only:

$(1\ 0\ -1\ 0)$
$\sigma_1^+ \sigma_3^-$

$(1\ 1\ -1\ 0)$ $\sigma_1^+ \sigma_2^+ \sigma_3^-$ $(1\ -1\ -1\ 0)$ $\sigma_1^+ \sigma_2^- \sigma_3^-$ $(1\ 0\ -1\ 1)$ $\sigma_1^+ \sigma_3^- \sigma_4^+$ $(1\ 0\ -1\ -1)$ $\sigma_1^+ \sigma_3^- \sigma_4^-$



Comments

- Separate identification of regulation function of each gene
- Based on nonconvex regression with noisy data:

$$\delta = \min_{\theta} \sum_{k} w_k \left(g_i^k - g_i(x^k | s, \theta) \right)^2$$

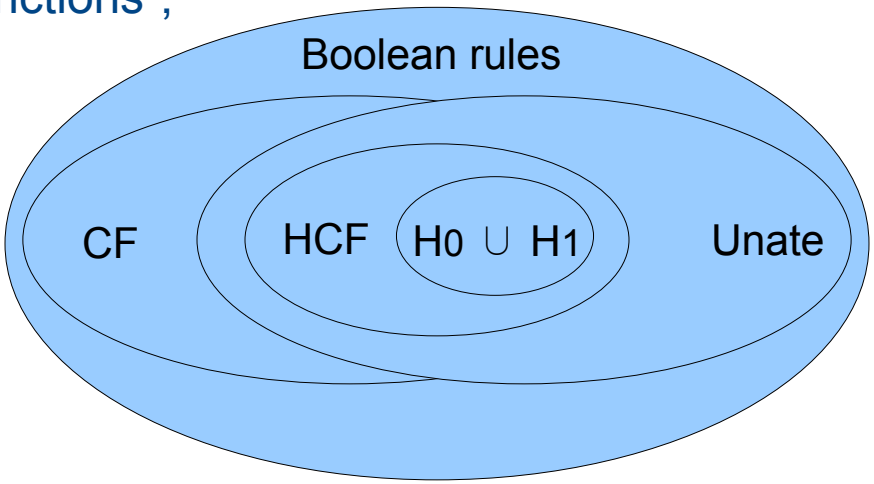
  - Weights normalize for the variance of measurement errors
  - Standard statistical tests for data ordering
- What is a statistically good model?
  - Under the null hypothesis that the estimated model is correct, the fitting residual is approximately distributed as $\chi^2(m)$
  - We accept a model structure with confidence level α if

$$\delta < \tau(\alpha), \qquad \tau(\alpha) \text{ computed from } \chi^2(m)$$

- Key question: how to define the model structures of interest ?

Some interesting model classes

- Goal: use a priori knowledge to reduce the family of network structures
- Intuition: many Boolean expression rules are unlikely/uncommon
- Evidence: (Szallasi *et al*, *Proc Pac Symp Bioc* 98, Kauffman *et al*, *PNAS* 04, ...)
  out of 139 gene activation rules analyzed, 99% are “Canalizing Functions”,
  95% are “Hierarchically Canalizing Functions”,
  90% are “$H_0 \cup H_1$”
  - CFs: at least one (canalizing) value of at least one (canalizing) variable determines the value of the function
  - HCFs: when the canalizing variable takes its non-canalizing value, a second variable is canalizing, etc.



We focus on H_0 U H_1

Identification of models in S = H_0 ∪ H_1

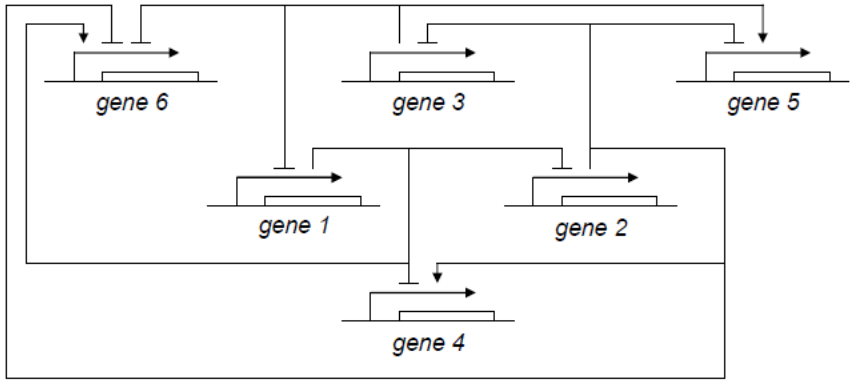
- Models in S: Only “and” (H_0) or at most one “or” (H_1)

$$\dot{x}_i = \kappa_i^1 + \kappa_i^2 b_i(x) - \gamma_i x_i \qquad \text{where}$$

$$b_i(x) = \begin{cases} \sigma^{\pm}(x_{j_1}) \cdot \sigma^{\pm}(x_{j_2}) \cdots \sigma^{\pm}(x_{j_\ell}) \\ \sigma^{\pm}(x_{j_1}) \cdot \sigma^{\pm}(x_{j_2}) \cdots \sigma^{\pm}(x_{j_{\ell-2}}) \left(1 - \sigma^{\mp}(x_{j_{\ell-1}}) \sigma^{\mp}(x_{j_\ell})\right) \end{cases}$$

- Given concentration and synthesis rates
  - Additive or multiplicative noise with known variance
- Estimate
  - Structure: $\ell,\ (j_1, j_2, \ldots, j_\ell),\ H_0$ vs. H_1
  - Rates and sigmoid parameters: $\kappa_i^1,\ \kappa_i^2,\ \theta_j$ (possibly depending on i)

Test on a repressilator-based system



$$\dot{x}_1 = \kappa_{0,1} + \kappa_{1,1}\sigma^-(x_3) - \gamma_1 x_1,$$

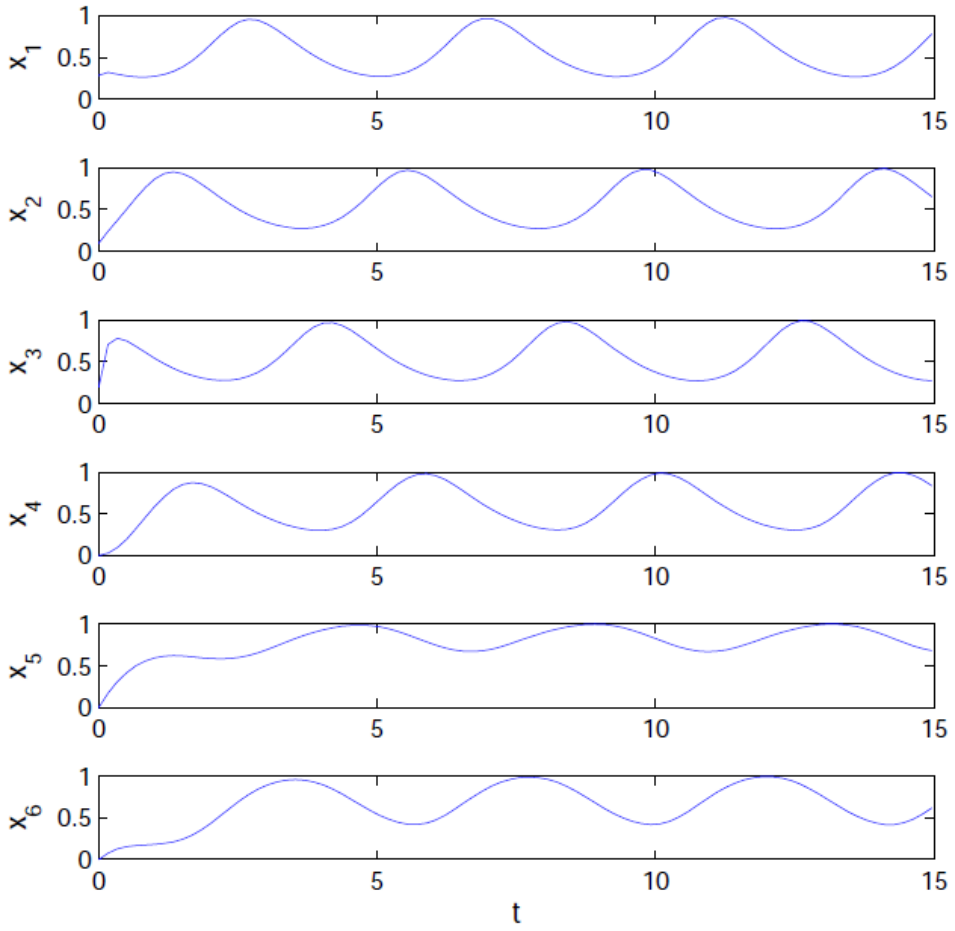
$$\dot{x}_2 = \kappa_{0,2} + \kappa_{1,2}\sigma^-(x_1) - \gamma_2 x_2,$$

$$\dot{x}_3 = \kappa_{0,3} + \kappa_{1,3}\sigma^-(x_2) - \gamma_3 x_3,$$

$$\dot{x}_4 = \kappa_{0,4} + \kappa_{1,4}\sigma^-(x_1)\sigma^+(x_2) - \gamma_4 x_4,$$

$$\dot{x}_5 = \kappa_{0,5} + \kappa_{1,5}[1 - \sigma^+(x_2)\sigma^-(x_3)] - \gamma_5 x_5,$$

$$\dot{x}_6 = \kappa_{0,6} + \kappa_{1,6}[1 - \sigma^+(x_2)\sigma^+(x_3)]\sigma^+(x_1) - \gamma_6 x_6.$$



Performance results

We attempted identification of this system with 90 equally spaced data points over a time interval such that the product concentrations of the core genes complete three full oscillations. Measurements $\tilde{x}_i^k$ and $\tilde{g}_i^k$ were artificially corrupted by Gaussian noise samples according to the observation model (7), with $v_e(x_i^k) = (\sigma_e x_i^k)^2$ and $v_\epsilon(g_i^k) = (\sigma_\epsilon g_i^k)^2$, for the different noise levels $\sigma_e = \sigma_\epsilon = 0.01, 0.03, 0.05, 0.07$. This corresponds to noise roughly within 3%, 10%, 15% and 20% of the actual values of x_i^k and g_i^k. The performance of Algorithm 1 (with $N=6$ and $\alpha=0.95$) for the various noise levels and all genes is conveyed by the scores on the performance indices R, S, A and D (Table 1). These were computed as described in Section 2.3.4 on the basis of $M=100$ identification runs with the same system evolution, but with different random outcomes of the noise. Each run (MATLAB V.7 R.14) took on an average roughly 5 min on a Windows XP workstation with Pentium 3.20 GHz processor and 2.00 GB RAM. Computational time ranged from ~2 s for the identification of g_3 to ~4 min for the identification of g_6. Step 1 always performs very reliably, i.e. index R is constantly

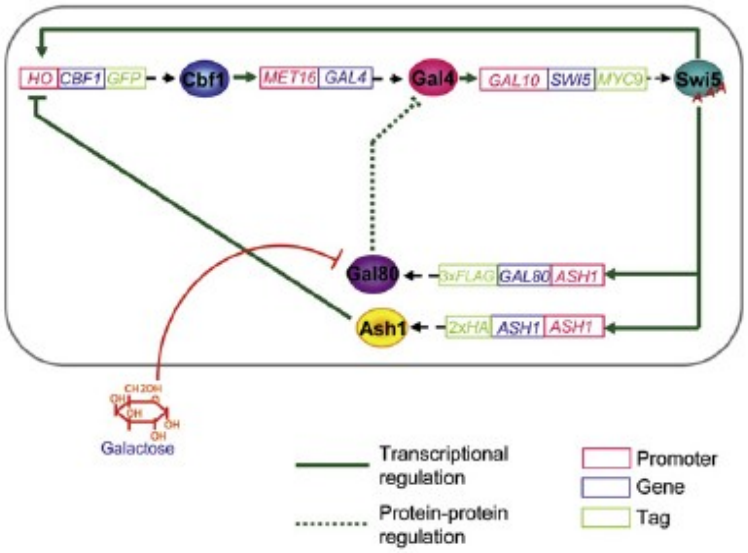
		$\sigma_e, \sigma_\epsilon$	0.01	0.03	0.05	0.07
Gene 1	Step 1	R	1	1	1	1
		S	0.92	0.92	0.92	0.91
	Step 2	A	0.90	0.92	0.91	0.89
		D	1	1	1	1
Gene 2	Step 1	R	1	1	1	1
		S	0.92	0.92	0.92	0.91
	Step 2	A	0.93	0.92	0.89	0.89
		D	1	1	1	1
Gene 3	Step 1	R	1	1	1	1
		S	0.92	0.92	0.92	0.92
	Step 2	A	0.93	0.93	0.93	0.92
		D	1	1	1	1
Gene 4	Step 1	R	1	1	1	1
		S	0.94	0.92	0.87	0.65
	Step 2	A	0.94	0.94	0.93	0.89
		D	1	1	1.02	1.44
Gene 5	Step 1	R	1	1	1	1
		S	0.94	0.74	0.53	0.48
	Step 2	A	0.95	0.94	0.91	0.83
		D	1	1	1.79	4
Gene 6	Step 1	R	1	1	1	1
		S	0.79	0.65	0.57	0.43
	Step 2	A	0.89	0.92	0.85	0.42
		D	1	1.02	2.76	2.74

(Porreca *et al*, *Bioinformatics* 2010)

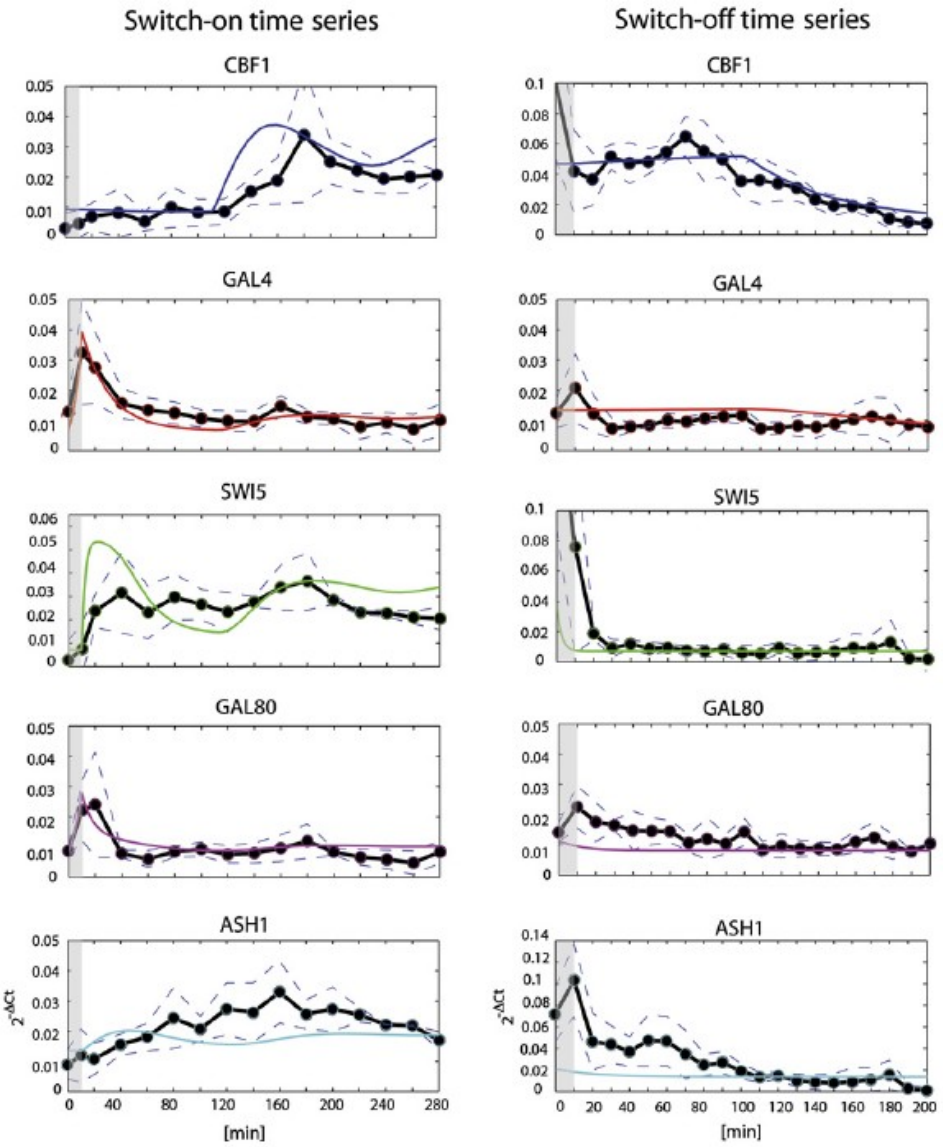
	Index	Range	Description
Step 1	R eliability	[0,1]	Probability that the true p is deemed consistent
	S electivity	[0,1]	Percentage of sign patterns eliminated from the search in Step 2
Step 2	A ccuracy	[0,1]	Probability that the true structure is In the pool of identified models
	D ispersion	≥1	Average number of models in the pool

Experiment on IRMA

- Synthetic gene network engineered in Yeast (Cantone *et al.*, *Cell* 2009)
- 5 genes with fluorescent reporters
- “Switch on” from glucose to galactose
- “Switch off” from galactose to glucose



(Cantone *et al.*, *Cell* 2009)



Mathematical model

Letting $[CBF1] = x_1$; $[GAL4] = x_2$; $[SWI5] = x_3$; $[GAL80] = x_4$; $[ASH1] = x_5$, the evolution of the mRNAs concentrations were modelled as follows:

$$\frac{dx_1}{dt} = \alpha_1 + v_1 \left(\frac{x_3^{h_1}(t-\tau)}{(k_1^{h_1} + x_3^{h_1}(t-\tau)) \cdot \left(1 + \frac{x_5^{h_2}}{k_2^{h_2}}\right)} \right) - d_1 x_1, \quad (1)$$

$$\frac{dx_2}{dt} = \alpha_2 + v_2 \left(\frac{x_1^{h_3}}{k_3^{h_3} + x_1^{h_3}} \right) - (d_2 - \Delta(\beta_1)) x_2, \quad (2)$$

$$\frac{dx_3}{dt} = \alpha_3 + \widehat{v_3} \left(\frac{x_2^{h_4}}{\widehat{k_4}^{h_4} + x_2^{h_4}(1 + \frac{x_4^4}{\widehat{\gamma}^4})} \right) - d_3 x_3, \quad (3)$$

$$\frac{dx_4}{dt} = \alpha_4 + v_4 \left(\frac{x_3^{h_5}}{k_5^{h_5} + x_3^{h_5}} \right) - (d_4 - \Delta(\beta_2)) x_4, \quad (4)$$

$$\frac{dx_5}{dt} = \alpha_5 + v_5 \left(\frac{x_3^{h_6}}{k_6^{h_6} + x_3^{h_6}} \right) - d_5 x_5, \quad (5)$$

(Cantone *et al.*, *Cell* 2009)

- We attempt identification in the class of models with *S*-structure
  - Different but similar analytical form
  - Test for flexibility of the approach
  - Known delays can be accounted for

Results

- Comparison with TSNI (di Bernardo *et al.*)
- A few protein concentration datapoints
- Rates simulated from the model:
  ``What-if'' study, extensions possible
- PPV=TD/TD+FD and Se=TD/TD+FU
  (T=True, D=Detected, U=Undetected edges)

Table 2. Average performance (standard errors in parentheses) on the IRMA datasets for different noise levels

	Switch-on data		Switch-off data	
σ_ϵ	PPV	Se	PPV	Se
0.07	0.98 (0.07) [0.98 (0.07)]	0.53 (0.08) [0.53 (0.08)]	0.91 (0.12) [0.88 (0.13)]	0.58 (0.07) [0.56 (0.08)]
0.1	0.95 (0.10) [0.94 (0.11)]	0.46 (0.08) [0.46 (0.08)]	0.85 (0.14) [0.80 (0.14)]	0.51 (0.09) [0.48 (0.09)]
0.3	0.67 (0.23) [0.64 (0.24)]	0.29 (0.10) [0.27 (0.10)]	0.58 (0.25) [0.52 (0.25)]	0.25 (0.11) [0.22 (0.11)]

Indices PPV and Se are reported for both the signed (in square brackets) and unsigned (without square brackets) directed graph.

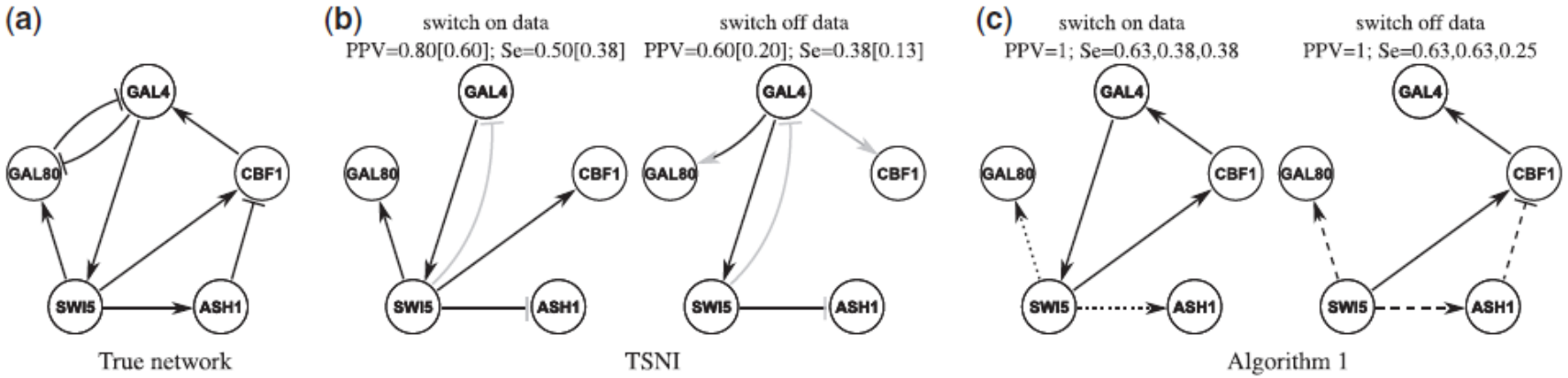


Fig. 1. (a) True network of interactions in IRMA. Results obtained by (b) the TSNI algorithm (Cantone *et al.*, 2009) and by (c) Algorithm 1. Grey arcs (respectively, grey-end markers) denote incorrect direction (respectively, sign) of the inferred interactions. Values of PPV and Se for the signed directed graph, when different from the unsigned case, appear in square brackets. The three values of Se in (c) refer to increasing noise levels, while dashed and dotted arcs denote interactions inferred only for $\sigma_\epsilon < 0.3$ and $\sigma_\epsilon < 0.1$, respectively.

(Porreca *et al*, *Bioinformatics* 2010)

Conclusions

- Algorithmic procedure for learning gene network dynamics from data
- Generalizes to any model class with monotonicity properties
  - Recent work by Belta and Julius along these lines
- Applicable to existing data, provided suitable preprocessing
  - E.g. $g_i(x) = \kappa_i^1 + \kappa_i^2 b_i(x) = \dot{x}_i + \gamma_i x_i$

    (Ronen *et al*, *PNAS* 2002, Brown *et al*, *Biotechnol J* 2008,...)
  - Recently: synthesis rates and variances from concentration profiles, via bootstrapping or deconvolution methods (Porreca *et al.*, CDC 2010)
- More properties of *S*- and unate-structure models to exploit
  - Quasi-convexity (To be presented at IFAC 2011. Submitted to *J Robust Nonlin Contr*, Special issue on System Identification of Biological Systems)
- Application to real *E.coli* carbon starvation response data

... Thank you!

eugenio.cinquemani@inria.fr