

Qualitative response of interaction networks: application to the validation of biological models

Anne Siegel, O. Radulescu, M. Le Borgne, C. Guziolowski, P.
Veber

Institute for Computer Science of Rennes (IRISA) [CNRS - INRIA - Université de Rennes 1]

July 2007, Zurich

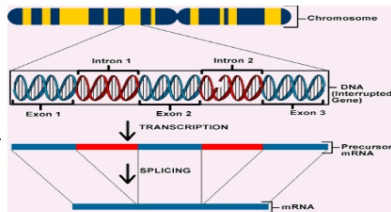
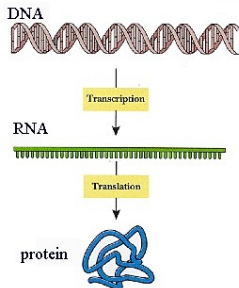


From genes to systems

- ▶ Proteins are obtained from DNA by genomic processes
- ▶ Not enough genes to explain diversity

Tracks ?

- ▶ Combinatorial genomic processes ?
- ▶ Role of non-coding DNA (98%) ? RNA ?

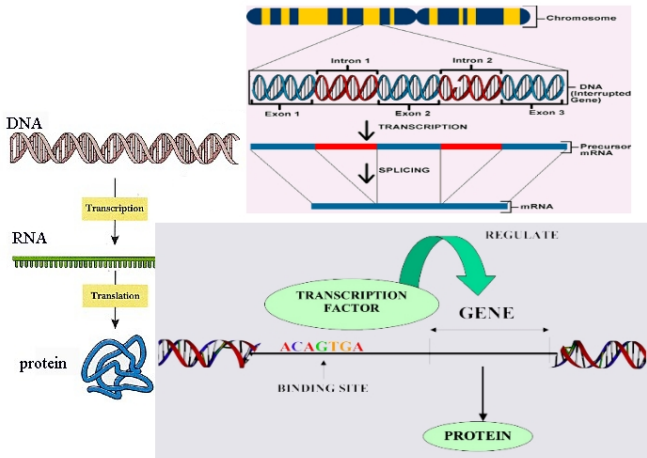


From genes to systems

- ▶ Proteins are obtained from DNA by genomic processes
- ▶ Not enough genes to explain diversity

Tracks ?

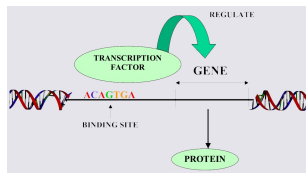
- ▶ Combinatorial genomic processes ?
- ▶ Role of non-coding DNA (98%) ? RNA ?
- ▶ Genetic and metabolic control of transcription ?



System Biology

System biology

Study of biological systems taking into account all its constituents and their relationships.



Model

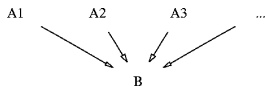
Mathematical description of the component of a system, their relationship and their evolution.

Tasks

- ▶ Construction of network models
- ▶ Simulation models
- ▶ Prediction of properties and behaviors

Flux modelling : mass action kinetics

- ▶ The production rate of a product is the sum of incoming influences

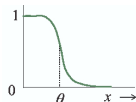


$$\frac{dB}{dt} = K_1 v(A_1) + K_2 v(A_2) + \dots$$

- ▶ Kinetic functions $v(A)$ depend on the type of process it describes
 - ▶ Chemical reaction : linear function
 - ▶ Enzyme reaction : sigmoid
 - ▶ Transcriptional regulation : threshold function



$$\begin{cases} \frac{dx}{dt} = k_x F_{S_x}^-(y) - k_{-x}x \\ \frac{dy}{dt} = k_y F_{S_y}^-(x) - k_{-y}y \end{cases}$$



$$F_S^-(x) = \frac{x^n}{x^n + S^n}$$

Differential models

Mathematical model

- ▶ Concentration vector of species $X(i) : \mathbf{X}$
- ▶ Control parameters \mathbf{P}
- ▶ Differential dynamics

$$\frac{d\mathbf{X}}{dt} = \mathbf{F}(\mathbf{X}, \mathbf{P})$$

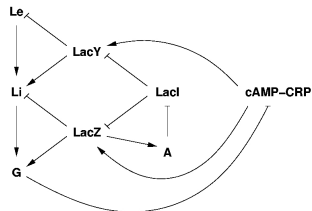
How precise is the knowledge about \mathbf{F}

- ▶ Metabolic or signalling networks : good knowledge of the coefficients
Simulations of the dynamics and study of phase space
- ▶ Regulatory genetic network : very few knowledge
 - ▶ The literature provides signs of fluxes dependencies
 - ▶ For threshold models : knowledge on comparisons between thresholds

Interaction graph

Definition

- ▶ $i \rightarrow j$ when a change in i has an influence of the production of j .



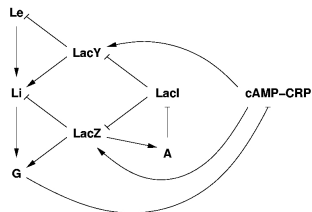
- ▶ Differential setting : it corresponds to the **signs of the Jacobian matrix**

$$i \rightarrow j \quad \text{ssi} \quad \frac{\partial F_{X(j)}}{\partial X(i)} \neq 0$$

Interaction graph

Definition

- ▶ $i \rightarrow j$ when a change in i has an influence on the production of j .



- ▶ Differential setting : it corresponds to the **signs of the Jacobian matrix**

$$i \rightarrow j \quad \text{ssi} \quad \frac{\partial F_{x(j)}}{\partial X(i)} \neq 0$$

The interaction graph is common to all type of models

- ▶ **Discrete models/ Piecewise linear models**

Small number of qualitative states
Dynamics given by transition functions

- ▶ **Differential models**

Large number of components
Smooth dynamics

- ▶ **Stochastic models**

Very few copy of each specy
Probalistic dynamics

High-throughput data

DNA Chip

In a given cell, we compare the quantity of **thousands** of m-RNA's **at the same time** to a given state.

Massive, noisy and qualitative data

Protein-protein interactions data

Measure affinities between proteins

Chip-on-Chip : Protein-DNA data

Choose a protein and look for all parts of the genome that can bind to this protein.

Inform on potential genetic interactions

How are such data used for the system biology task ?

Inference of interaction graphs

linear models : probabilistic rule

Incorporate high-throughput biological data as prior knowledge to a Bayesian network or a more general probabilistic model.

▶ Lee et al. (2002)

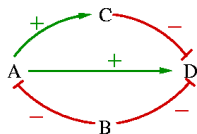
▶ Segal et al. (2002)

▶ Ideker et al., Yeang et al.
(2001-05)

▶ Vert et al. (2006)

▶ ...

Most of the methods use a **probabilist linear model** to achieve the inference task.



$$P(A) = P(A|B)P(B) + P(A|\neg B)P(\neg B)$$

$$B \text{ inhibits } A : P(A|\neg B) \simeq 1 \text{ and } P(A|B) \simeq 0$$

Inference of interaction graphs

linear models : probabilistic rule

Incorporate high-throughput biological data as prior knowledge to a Bayesian network or a more general probabilistic model.

▶ Lee et al. (2002)

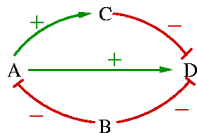
▶ Segal et al. (2002)

▶ Ideker et al., Yeang et al.
(2001-05)

▶ Vert et al. (2006)

▶ ...

Most of the methods use a **probabilist linear model** to achieve the inference task.



$$P(A) = P(A|B)P(B) + P(A|\!B)P(\!B)$$

$$B \text{ inhibits } A : P(A|\!B) \simeq 1 \text{ and } P(A|B) \simeq 0$$

Intuitive underlying rule ?

The variations of a specy are mainly explained by the variations of at least of of its predecessors.

If D increases, then either A have increased, or B or C have decreased ?

Small variations

Assume that

- ▶ Start from a steady state $\mathbf{F}(\mathbf{X}_1, \mathbf{P}_1) = \mathbf{0}$
- ▶ Stress the system by acting on the parameters
- ▶ Wait for a new steady state $\mathbf{F}(\mathbf{X}_2, \mathbf{P}_2) = \mathbf{0}$
- ▶ Compute the sign of the variations $\mathbf{X}_2 - \mathbf{X}_1$.

Linearize the steady state equation

$$\mathbf{F}(\mathbf{X}, \mathbf{P}) = \mathbf{0}$$

If i is a node such that $X(i)$ is not directly influenced by the parameters ($\frac{\partial \mathbf{F}_{X(i)}}{\partial P(k)} = 0$) and ($\frac{\partial \mathbf{F}_{X(i)}}{\partial X(i)} \neq 0$), then the variations of $X(i)$ between two steady states are related to the variations of its predecessors in the graph :

$$\delta X(i) = - \left(\frac{\partial \mathbf{F}_{X(i)}}{\partial X(i)} \right)^{-1} \sum_{k \neq i, k \rightarrow i} \frac{\partial \mathbf{F}_{X(i)}}{\partial X(k)} \delta X(k).$$

What about large variations ?

$$\delta X(i) = - \left(\frac{\partial \mathbf{F}_{X(i)}}{\partial X(i)} \right)^{-1} \sum_{k \neq i, k \rightarrow i} \frac{\partial \mathbf{F}_{X(i)}}{\partial X(k)} \delta X(k).$$

- ▶ Consider the species that influence i : $\hat{X}^{(i)} = (X(k_1), \dots, X(k_p))$, $k_j \neq i, k_j \rightarrow i$. Then $\mathbf{F}_{X(i)} = \mathbf{F}_{X(i)}(X(i), \hat{X}^{(i)})$
- ▶ If $\frac{\partial \mathbf{F}_{X(i)}}{\partial X(i)} < -C$ and $\mathbf{F}_{X(i)}(\{\mathbf{X}, X(i) = 0\}) > 0$ then $\mathbf{F}_{X(i)}(\cdot, \hat{X}^{(i)})$ has exactly one zero

$$X(i) = \Phi_i(\hat{X}^{(i)}).$$

- ▶ By the implicit function theorem :
$$d\Phi_i = - \left(\frac{\partial \mathbf{F}_{X(i)}}{\partial X(i)} \right)^{-1} \sum_{k \neq i, k \rightarrow i} \frac{\partial \mathbf{F}_{X(i)}}{\partial X(k)} dX(k).$$
- ▶ Consider a differential path $C^{1,2}$ between the steady states that satisfies $\mathbf{F}_i(\Phi_i(C^{1,2}), \hat{X}^{(i)}(C^{1,2})) = 0$ and does not go through the singular points $\mathbf{F}_{X(i)}$.

$$\Delta X_i = \int_{C^{1,2}} - \left(\frac{\partial \mathbf{F}_{X(i)}}{\partial X(i)} \right)^{-1} \sum_{k \neq i, k \rightarrow i} \frac{\partial \mathbf{F}_{X(i)}}{\partial X(k)} dX(k).$$

- ▶ Since Φ_i is bounded when $\hat{X}^{(i)} \in C^{1,2}$, $C^{1,2}$ is near from experimental observations and the signs remain constant.

The intuitive rule is true under reasonable assumptions

Theorem

Assume that

- ▶ *Specy i autoregulates itself negatively* : $\frac{\partial \mathbf{F}_{X(i)}}{\partial X(i)} < -C, \quad C > 0.$
- ▶ *There is no direct influence from \mathbf{P} on $X(i)$*
- ▶ *When i is absent, the system produces it* $\mathbf{F}_{X(i)}(\{\mathbf{X}, X(i) = 0\}) > 0$
- ▶ *For every predecessor $k \rightarrow i$, the sign of the action $\frac{\partial \mathbf{F}_{X(i)}}{\partial X(k)}$ is constant during the experimentation*

Then the variations of the species between two steady states (for different parameters) satisfy the following relationship :

$$\text{sign}(\Delta X(i)) \simeq \sum_{k \neq i, k \rightarrow i} \text{sign} \left(\frac{\partial \mathbf{F}_{X(i)}}{\partial X(k)} \right) \times \text{sign}(\Delta X(k)).$$

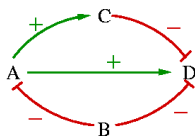
Addition, multiplication and equality hold in the **sign algebra** $\{[+], [-], [?]\}$

$$[+] + [-] = [?], \quad [+] \simeq [?] \quad [-] \simeq [?] \quad [-] \neq [+]$$

Example when interaction signs are known

$$\text{sign}(\Delta X(i)) \approx \sum_{k \neq i, k \rightarrow i} \text{sign}(i \rightarrow k) \times \text{sign}(\Delta X(k)).$$

Usual sign rules and additional rules : $++- = ?$ $+ \neq -$



- ▶ The variation of C is given by the variation of A
 $\text{sign}(\Delta C) \approx \text{sign}(\Delta A)$
- ▶ the variation of A is the opposite of the variation of B
 $\text{sign}(\Delta A) \approx -\text{sign}(\Delta B)$
- ▶ the variation of D must be equal to the variation of A , $-B$ or $-C$.
 $\text{sign}(\Delta D) \approx \text{sign}(\Delta A) - \text{sign}(\Delta B) - \text{sign}(\Delta C)$

$$\text{sign}(\Delta C) \approx \text{sign}(\Delta A)$$

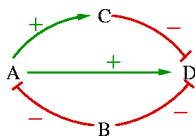
$$\text{sign}(\Delta A) \approx -\text{sign}(\Delta B)$$

$$\text{sign}(\Delta D) \approx \text{sign}(\Delta A) - \text{sign}(\Delta B) - \text{sign}(\Delta C)$$

Example when interaction signs are known

$$\text{sign}(\Delta X(i)) \approx \sum_{k \neq i, k \rightarrow i} \text{sign}(i \rightarrow k) \times \text{sign}(\Delta X(k)).$$

Usual sign rules and additional rules : $++- = ?$ $+ \neq -$



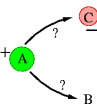
- ▶ The variation of C is given by the variation of A
 $\text{sign}(\Delta C) \approx \text{sign}(\Delta A)$
- ▶ the variation of A is the opposite of the variation of B
 $\text{sign}(\Delta A) \approx -\text{sign}(\Delta B)$
- ▶ the variation of D must be equal to the variation of A , $-B$ or $-C$.
 $\text{sign}(\Delta D) \approx \text{sign}(\Delta A) - \text{sign}(\Delta B) - \text{sign}(\Delta C)$

$$\begin{array}{lcl} \text{sign}(\Delta C) & \approx & \text{sign}(\Delta A) \\ + & \approx & + \\ \text{sign}(\Delta A) & \approx & -\text{sign}(\Delta B) \\ + & \approx & -(-) \\ \text{sign}(\Delta D) & \approx & \text{sign}(\Delta A) - \text{sign}(\Delta B) - \text{sign}(\Delta C) \\ + & \approx & + - (-) - (+) \end{array}$$

There are 4 sets of solutions (among 16 possible)

A	B	C	D
+	-	+	+
+	-	+	-
-	+	-	+
-	+	-	-

Example when interaction signs are NOT known



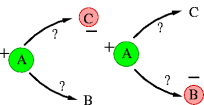
One experiment

$$\text{sign}(A \rightarrow C) = -$$

$$\text{sign}(\Delta C) \simeq \text{sign}(A \rightarrow C) \text{sign}(\Delta A)$$

$$\text{sign}(\Delta A) = +$$

$$\text{sign}(\Delta C) = -$$



Two experiments

$$\text{sign}(A \rightarrow C) = -$$

$$\text{sign}(A \rightarrow B) = -$$

$$\text{sign}(\Delta C^{(1)}) \simeq \text{sign}(A \rightarrow C) \text{sign}(\Delta A^{(1)})$$

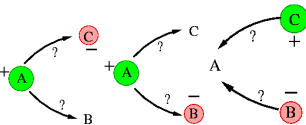
$$\text{sign}(\Delta B^{(2)}) \simeq \text{sign}(A \rightarrow B) \text{sign}(\Delta A^{(2)})$$

$$\text{sign}(\Delta A^{(1)}) = +$$

$$\text{sign}(\Delta C^{(1)}) = -$$

$$\text{sign}(\Delta A^{(2)}) = +$$

$$\text{sign}(\Delta B^{(2)}) = -$$



Three experiments

$$\text{sign}(A \rightarrow C) = -$$

$$\text{sign}(A \rightarrow B) = -$$

$$\text{sign}(\Delta A^{(3)}) = +$$

$$\text{sign}(\Delta A^{(3)}) = -$$

Incompatibility

$$\text{sign}(\Delta C^{(1)}) \simeq \text{sign}(A \rightarrow C) \text{sign}(\Delta A^{(1)})$$

$$\text{sign}(\Delta B^{(2)}) \simeq \text{sign}(A \rightarrow B) \text{sign}(\Delta A^{(2)})$$

$$\text{sign}(\Delta C^{(3)}) \simeq \text{sign}(A \rightarrow C) \text{sign}(\Delta A^{(3)})$$

$$\text{sign}(\Delta B^{(3)}) \simeq \text{sign}(A \rightarrow B) \text{sign}(\Delta A^{(3)})$$

$$\text{sign}(\Delta A^{(1)}) = +$$

$$\text{sign}(\Delta C^{(1)}) = -$$

$$\text{sign}(\Delta A^{(2)}) = +$$

$$\text{sign}(\Delta B^{(2)}) = -$$

$$\text{sign}(\Delta B^{(3)}) = +$$

$$\text{sign}(\Delta C^{(3)}) = -$$

Setting constraints from high throughput data

Variables

- ▶ signs of the **variation of products** $\Delta X(i, \eta)$ in each considered experimentation
(underlying hypothesis : data concern stationary state shifts)
- ▶ signs of **interactions** $s(i \rightarrow k)$
(underlying *restrictive* hypothesis : every actor has a constant action on its target)

Constraints

- ▶ **litterature knowledge** set up the signs of some interactions.
- ▶ **Chip-Chip data** tell that some signs are zero.
- ▶ **qualitative data** set up the sign of some variations.
- ▶ **General constraint** : the variation of an *internal* product is explained by the variation of one of its predecessors

$$\text{sign}(\Delta X(i, \eta)) \simeq \sum_{k \neq i, k \rightarrow i} \text{sign}(s(i \rightarrow k)) \times \text{sign}(\Delta X(k, \eta)).$$

Studying constraints for biological purpose

Questions asked by biologists

- ▶ **Validation** : consistency between knowledge and data
- ▶ **Corrections** of the model ?
- ▶ **Prediction** of new information : variation for nonobserved products or sign of unknown interaction.
- ▶ **Key nodes**
 - ▶ For the validation of the model
 - ▶ For the understanding of behaviors

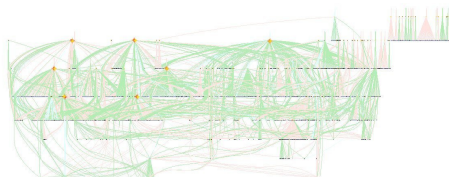
Two mains tools to realize these tasks

- ▶ Enumeration of solutions by **Decision Diagrams** (Pyquali)
 - ▶ Compact representation of the solutions in $\{+, -\}$
 - ▶ Elimination of variables
- ▶ Solver for constraints expressed in **Answer Set Programming** (Clasp)
 - ▶ Provides one solution for a given set of constraints.

Application to consistency

- ▶ **Biological question** Are the different pieces of information coherent with each other ?
- ▶ **Translation into constraints framework** Do the system of constraint admit at least a solution ?

Example : the network of transcriptional interactions for E. Coli given by Regulon DB is not internally coherent.



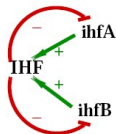
- ▶ Large scale network with hierarchical structure (87% of genes are regulated by 13%)
- ▶ 160 doubled signed interactions
- ▶ **1100 constraints**, 1258 variables

Number of nodes	1258
Number of interactions	2526
Nodes without successor	1101
Nodes with more than 80 successors	7
protein complex	4

Application to corrections

- ▶ **Biological question** When I have contradicting data and knowledge, what should I change ?
- ▶ **Origin of errors**
 - ▶ Errors in experimental data or knowledge
 - ▶ Missing interaction between nodes
 - ▶ Non-constant signed action between an actor and its target
 - ▶ (Missing node)
- ▶ **Constraints framework** What is the minimal set of equations that raise inconsistency ?

Automatic identification of inconsistent system



$$IHF \approx ihfA + ihfB \quad (1)$$

$$ihfA \approx -IHF \quad (2)$$

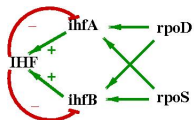
$$ihfB \approx -IHF \quad (3)$$

ihfA	ihfB	IHF	Conflict
+	+	+	(2), (3)
+	+	-	(1)
+	-	+	(1)
+	-	-	(1)
-	+	+	(1)
-	+	-	(1)
-	-	+	(1)
-	-	-	(2), (3)

Application to corrections

- ▶ **Biological question** When I have contradicting data and knowledge, what should I change ?
- ▶ **Origin of errors**
 - ▶ Errors in experimental data or knowledge
 - ▶ Missing interaction between nodes
 - ▶ Non-constant signed action between an actor and its target
 - ▶ (Missing node)
- ▶ **Constraints framework** What is the minimal set of equations that raise inconsistency ?

Manual curated answer : Adding new interactions



$$\begin{aligned}
 IHF &\approx ihfA + ihfB \\
 ihfA &\approx -IHF + rpoD + rpoS \\
 ihfB &\approx -IHF + rpoD + rpoS
 \end{aligned}$$

Consistent system (18 solutions

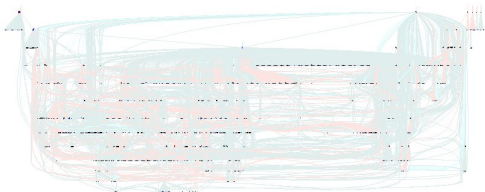
among 32)

rpoD	rpoS	ihfA	ihfB	IHF
+	+	+	+	+
+	+	+	-	+
+	+	-	+	+
-	-	-	-	-
-	-	-	+	-
-	-	+	-	-
<hr/>				
+/-	-/+	+	+	+
+/-	-/+	+	-	+
+/-	-/+	+	-	-
+/-	-/+	-	+	+

Protein	Gene	Function
σ^{70}	rpoD	Transcribes most genes in growing cells
σ^{38}	rpoS	The starvation/stationary phase sigma-factor

Large scale network corrections

New (consistent) model and data on exponential phase



Number of nodes	1529
Number of interactions	3883
Nodes without successor	1365
Nodes with more than 80 successors	10
sigma-factors	6
protein complex	4

gene	effect	gene	effect	gene	effect	gene	effect	gene	effect
acnA	+	csiE	+	gadC	+	osmB	+	recF	+
acrA	+	cspD	+	hmp	+	osmE	+	rob	+
adhE	+	dnaN	+	hns	+	osmY	+	sdaA	-
appB	+	dppA	+	hyaA	+	otsA	+	sohB	-
appC	+	fic	+	ihfA	-	otsB	+	treA	+
appY	+	gabP	+	ihfB	-	polA	+	yeiL	+
blc	+	gadA	+	lrp	+	proP	+	yfiD	+
bolA	+	gadB	+	mpl	+	proX	+	yihI	-

Model and data are inconsistent !

Correction algorithm. There was a **mistake on data** provided by RegulonDB

Good variations : $ihfA = +$ and $ihfB = +$ (confirmed by the litterature)

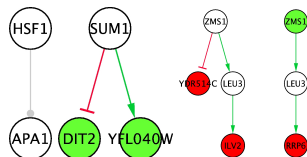
Validation of unsigned network ?

Several unsigned networks for *S. Cerevisiae* and datasets

- ▶ Unsigned interactions between transcription factors (from Chip-Chip analyses or promoteur inference) [70/83 nodes, 96/131 edges]
- ▶ Full interaction network given by Chip-Chip analyses (Lee et al, 2002) [2419 nodes 4344 edges]
- ▶ 15 quite complete stress experimental datasets (YDB)
- ▶ About 300 mutant experimentations (Hugues et al, 2000)

Result : All unsigned networks are inconsistent with the datasets

We identify inconsistent subsets for each network



Application to predictions

- ▶ **Biological question** What do the knowledge and data predict on nonobserved signed and/or products ?
- ▶ **Computer scientist question** What are the variables whose sign is the same in all solutions ?

Application 1 : E. Coli and 40 stationary phase data

Allows to predict 401 new variations (that is, 26 % of the network)

Application 2 : unsigned S. Cerevisiae

Allow to infer 15% of signs

Interaction network	Nodes	Edges	Number Exp.	Input/Output obs. simul.	Inferred	Incompatibilities
(B) Extended Lee network	70	96	15	70	29 (30.2%)	7.2%
(D) Global network	2419	4344	14	2270	21 (16%)	17.5%

Conclusion

Qualitative data and modelling

- ▶ Inferring the structure of models (statistic approach)
- ▶ **Constraints for the behavior of the system**
 - ▶ Validation, correction, predictions

To be done

- ▶ **Experiment design** : influence of nodes on the consistency of the network and its predictability ?
- ▶ Take **time-series data** into account ?
- ▶ System analysis : identify competitions processes.

Are the hypotheses so reasonable ?

- ▶ What happens when the dynamics is not smooth ?
Non-continuous/discrete models ? Is the rule still valid ?

Acknowledgments

Computer scientists

C. Guziolowski (IRISA, Rennes)

M. Le Borgne (IRISA, Rennes)

P. Veber (IRISA, Rennes)

Mathematician

O. Radulescu (departement of Mathematics, Rennes)

Biologists

S. Lagarrigue (INRA, Rennes)

P. Blavy (INRA - IRISA, Rennes)

References

Radulescu O, Lagarrigue S, Siegel A, Veber P, Borgne ML (2006) Topology and static response of interaction networks in molecular biology. *J R Soc Interface* 3 :185–96.

Siegel A, Radulescu O, Borgne ML, Veber P, Ouy J, et al. (2006) Qualitative analysis of the relation between DNA microarray data and behavioral models of regulation networks. *Biosystems* 84 :153–74.

Veber P, Borgne ML, Siegel A, Lagarrigue S, Radulescu O (2004/2005) Complex qualitative models in biology : A new approach. *Complexus* 2 :140–151.

Guziolowski C, Veber P, Borgne ML, Radulescu O, Siegel A (2007) Checking consistency between expression data and large scale regulatory networks : A case study. *Journal of Journal of Biological Physics and Chemistry*

Veber P, Guziolowski C, Borgne ML, Radulescu O, Siegel A (2007) Inferring the role of transcription factors in regulatory networks. Submitted.