Qualitative analysis of the relation between DNA microarray data and behavioral models of regulation networks

A. Siegel[1], O. Radulescu[2], M. Le Borgne[1], P. Veber[1], J. Ouy[1], S. Laguarrigue[3]

Introduction

Analysis of DNA microarrays New experimental techniques such as DNA microarrays or mass spectrometry have become a very useful tool for the discovery and the understanding of cellular genetic regulations. They allow to measure average gene expression for a very high number of targets. However, experimental results are usually rather noisy and are consequently difficult to interpret. Much efforts have been done to analyze them in a statistical framework, despite the usual small number of samples at hand (opposite to the large number of observed variables). Clustering methods were used to detect coexpressed genes [KF02]; there were several attempts to propose algorithms that infer regulation networks from microarray data [YVK04, NRF04]. In this paper, we adopt a different approach, namely to consider microarray data as qualitative information that can be compared to the predictions of a qualitative behavioral regulation model.

Cellular processes are usually represented as networks of interacting molecules. Most models which are built from experimental data are quantitative and simulation oriented, but some work has been done to propose qualitative descriptions. Metabolic path analysis is probably the most representative of this kind of approach [PSP+04]. Qualitative analysis is important as for large networks, quantitative data are most of the time missing and difficult to obtain. On the other hand, there exists a wealth of qualitative knowledge scattered in the scientific literature. In this paper we show how to make use of this knowledge for the analysis of microarray results.

Qualitative models and experiments Microarrays can be seen as qualitative data if, for one gene, one only considers the sign of the difference in its expression level between two situations – usually the effects of particular conditions are compared to a reference situation. These data can be supplemented with the same kind of informations (signs of variations) for metabolite levels, obtained by biochemical measurements. Our goal is to assess if these experimental measurements are compatible with what is known about the interactions among the targets.

This knowledge is represented as a labeled graph, where nodes represent molecules (such as mRNA, proteins or metabolites) and edges represent interactions, labeled by elements of $\{+,-\}$. Signs on the edges have a dynamic interpretation : a $-$ on an edge (resp. a $+$) between two molecules A and B means that an increase of the concentration of A decreases (resp. increases) the rate of production of B. Such information about behavioral interactions can be collected either from interaction databases (Bind [BG03], IntAct Project, Amaze [LACa04]) or directly extracted from the literature.

Qualitative analysis Analysis of such a qualitative graph is difficult when the graph has numerous edges and nodes; a naive approach is prone to errors. In this paper, we present a formal analysis of such graphs, based on a mathematical model: an experiment is modeled as a steady state shift of a differential dynamical system. This leads us to equations connecting transcript variations, that we interpret qualitatively as a linear system in the sign algebra [Kui94, TMD03]. This system represents a generalization of the discrete Laplace equation on an oriented graph, and as the Laplace equation, it has solutions that are determined by values on the entering boundary [CY00, TS90].

We introduce three complementary qualitative analyses based on this approach. The modeling and the algorithms will be illustrated on a simplified model of lipid metabolism in liver cells.

[1] Projet Symbiose. Institut de Recherche en Informatique et Systèmes Aléatoires, IRISA-CNRS 6074, Campus de Beaulieu, 35042 Rennes Cedes, France

[2] Institut de Recherche Mathématique de Rennes, UMR-CNRS 6625, Campus de Beaulieu, 35042 Rennes Cedex, France

[3] ENSAR - INRA Génétique Animale, 65, rue de St Brieuc, 35042 Rennes cedex, France

1 Working example: regulation of the synthesis of fatty acids

The example studied in this paper deals with the genetic regulation of lipid metabolism in liver. The approach that we describe allows to perform partial studies on a set of nodes that were chosen because of their mutual interactions. The choice of the observed nodes was not optimized for this approach and we shall see how our analysis suggests improvements of this choice.

In liver, citrate is transformed into Polyunsaturated fatty acids (PUFA), thanks to a metabolic pathway composed of four enzymes, namely ACL (acyl-CoA synthetase), ACC (acetyl-Coenzyme A carboxylase), FAS (fatty acid synthase) and SCD1 (Stearoyl-CoA desaturase 1).

Nonesterified fatty acids or their CoA derivatives seem to be the main signal involved in the transcriptional effect of long-chain fatty acids. The effects of fatty acids are mediated either directly through their specific binding to various nuclear receptors (PPARα – peroxisome proliferator activated receptors, LXR – nuclear receptor subfamily 1, group H, member 3, HNF-4alpha) leading to changes in the trans-activating activity of these transcription factors; or indirectly as the result of changes in the abundance of regulatory transcription factors (SREBP-1c – sterol regulatory element binding-protein–, ChREBP, etc.) [PLMG04].

We wish to keep the following established relations in the model that will be used as an example in this paper. The products we consider are enzymes ACL, ACC, FAS, SCD1 and SCAP; nuclear receptor PPAR, LXR, SREBP, as they are synthesized from the corresponding genes; trans-activating active form of the transcription factors, that is, LXR-a (denoting a complex LXR:RXR), PPAR-a (denoting a complex PPAR:RXR) and SREBP-a (denoting the cleaved form of SREBP-1c, a representative of the clivage enzymes being SCAP – src family associated phosphoprotein); arachidonic acid (20:4n-6), denoted PUFA, as a representative of polyunsaturated fatty acids.

Relations between these products are the following: SREBP-a is an activator of the transcription of ACL, ACC, FAS, and SCD1; LXR-a is an activator of the transcription of SREBP and FAS, it also indirectly activates ACL and ACC; PPAR indirectly inhibits the production of ACL, ACC, FAS, and SCD1; PUFA activates the formation of PPAR-a from PPAR, and inhibits the formation of LXR-a from LXR; SCAP represents the activators of the formation of SREBP-a from SREBP.

As a working example, we suppose that during an experimentation, animals (chicken and mice) were kept in a diet during several hours. Then, the transcriptional quantity of LXR, SREBP, PPAR, ACL, ACC and SCD1 were extracted from a DNA microarray analysis of their liver. Biochemical measures also provided the variation of PUFA. We assume that some experimentations have shown that ACL, LXR, PPAR and PUFA increases, SREBP, ACC, FAS and SCD1 decreases.

2 Mathematical framework

We justify here the approach and impose some applicability limits. The basic mathematical object is a qualitative graph representing the present knowledge on interactions and their qualitative effects for the biological process under study. We assume that experimental data give variations of the concentrations of some molecules between two external conditions. The corresponding nodes will be called *observed nodes*.

First modeling assumption: differential model We assume that the various interactions involved in the model evolve according to a differential dynamics:

$$\frac{dX}{dt} = F(X)$$

The state vector X is indexed by $I = \{1, \ldots, n\}$.

2.1 Interaction graph

Interaction graph The interaction graph is defined as $M = (I, A, s)$, where I gathers all the products in the model together with an extra generic node $\mathcal{E}$ to represent the exterior world.

The set $A \subset I \times I$ contains oriented edges joining $j \in I$ and $i \in I$ whenever $\frac{\partial F_i}{\partial X_j} \neq 0$. Every edge is labeled by $s(j,i) = sign(\frac{\partial F_i}{\partial X_j}) \in \{+,-\}$. The sign of the interactions is the only input to the model; no other information such as the form of the functions F or the absolute values of the partial derivatives $\frac{\partial F_i}{\partial X_j}$ will be used.

In our example, we interpret the biological information in a differential model setting. Whenever the literature proves an action (positive or negative) from j to i, we consider an edge from j to i in the interaction graph of the differential model. Moreover, we assume the existence, beside the other interactions, of a negative self-interaction on all nodes of the

graph, implying $s(i,i) = -$ for any i. These result for instance from degradation processes.

For the biological problem that is used as an illustration we consider two slightly different models: I_1 = {PUFA,LXR,LXR-a,PPAR,PPAR-a,SREBP,SREBP-a,ACL,ACC,FAS,SCD1,SCAP} $\cup$ $\{\mathcal{E}\}$, and I_2 = {PUFA,LXR,LXR-a,PPAR,PPAR-a,SREBP,SREBP-a,ACL,ACC,FAS,SCD1}$\cup\{\mathcal{E}\}$. In the second model we assume that the influence of SCAP can be neglected, which is correct if the variation of SCAP is small. The interaction graph for I_1 and I_2 are shown in Fig.2.1 and Fig.2.2.

By specifying the connections of the node $\mathcal{E}$ to the other nodes we specify which are the entering nodes of the system. This is an important modeling hypothesis.

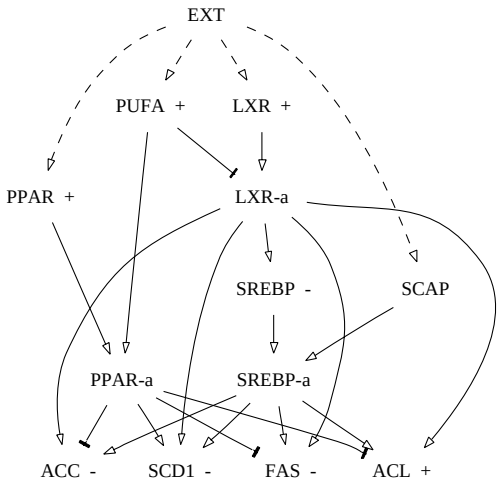

Figure 2.1: Interaction graph I_1 for a model of regulation of the synthesis of fatty acids. The node EXT stands for $\mathcal{E}$ and represents the exterior world. Self-regulation loops on nodes are omitted for sake of clarity. When available, signs correspond to observed variations.

Path on graph A path P is a sequence of nodes $\{i_1, i_2, \ldots, i_p\}$ such that $(i_k, i_{k+1}) \in A$. Let a_P denote the product of partial derivatives $a_P = \prod_{k=1}^{k=p-1} \frac{\partial F_{i_{k+1}}}{\partial X_{i_k}}$. The sign of the path is the product $s(P) = sign(a_P)$. A path from j to i will be denoted $j \rightsquigarrow i$.

Analyzed subnetwork, observed nodes Let $G \subset I$ be a subset of the nodes. The nodes in G inherit the connections from M. G will be the set

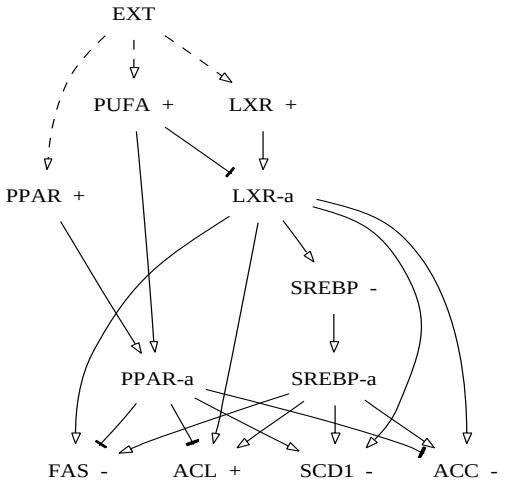

Figure 2.2: Interaction graph I_2 for a simplified model of regulation of the synthesis of fatty acids. The influence of SCAP is supposed to be negligible.

on which we perform the analysis; this may be any chosen set of nodes.

The set of observed nodes will be denoted $\mathcal{O}$. On $\mathcal{O}$ one can define the nodes variations $\delta X_i, i \in \mathcal{O}$, given by experimental data. The sign of these variations is denoted by $s(\delta X_i)$. In our examples, only some nodes are observed:

$$\mathcal{O} = \{\text{PPAR,LXR,ACL,SREBP,FAS,SCD1,PUFA}\}.$$

2.2 Modeling the influence of the exterior: boundary

Our problem, like many other combinatorial problems, gives rise to Laplace equations on graphs [CY00]. Let us adapt this framework here; the nodes of the analyzed subgraph G are partitioned into two subsets: the entering boundary and the inside nodes.

Entering boundary The entering boundary of G, denoted $\gamma^{in}G$ is formed by all the nodes of G that are supposed to have entering connections from the outside: $\gamma^{in}G = \{i \in G | \exists j \in I \setminus G, \quad (j,i) \in A\}$.

Notice that when $G = I \setminus \{\mathcal{E}\}$, the entering boundary is entirely made of those nodes that have connections from $\mathcal{E}$. Specifying this set is a modeling hypothesis, since specifying a boundary excludes other entering points in the system.

In the two models of our working example, entering boundaries are supposed to be

$\urcorner^{in}G_1$ = {PPAR, LXR,PUFA,SCAP} and $\urcorner^{in}G_2$ = {PPAR, LXR,PUFA}.

Simple entering boundary The entering boundary of G is simple if there is no path entirely contained in G that connects two different points of the boundary. i.e. $i \in \urcorner^{in}G$, $j \rightsquigarrow i \subset G$ and $j \in \urcorner^{in}G \implies j = i$.

The entering boundaries for both models G_1 and G_2 are simple.

Predecessors, incoming shells We denote by $pred(i)$ the set of predecessors of i: $pred(i) = \{j \in I, (j,i) \in A\}$.

We say that the set of predecessors is an incoming shell if $\urcorner^{in}(\{i\} \cup pred(i)) = pred(i)$; it is a simple incoming shell if $\urcorner^{in}(\{i\} \cup pred(i))$ is simple.

There are several situations when $pred(i)$ is not a simple incoming shell:

- a predecessor j of i has no connection from the exterior of $\{i\} \cup pred(i)$. Then j is not on $\urcorner^{in}(\{i\} \cup pred(i))$ and $pred(i)$ is not an incoming shell. A typical case is $j = \mathcal{E}$, implying that any successor of $\mathcal{E}$ does not have a simple incoming shell;
- there is an edge connecting two predecessors of i;
- i has at least two predecessors and one of its predecessors j forms with i a two-nodes loop. Then there are paths from all other predecessors of i to j and the incoming shell is not simple.

In our example, the predecessors of LXR-a,PPAR-a,SREBP,SREBP-a,ACL,ACC,FAS,SCD1 are simple incoming shells. For instance the list of predecessors of SREBP is $pred$(SREBP) = {LXR-a}, which is a simple incoming shell.

The predecessors of LXR, PUFA, SCAP are not incoming shells for a common reason: all these nodes have $\mathcal{E}$ as predecessor which by definition has only exit arcs.

2.3 Steady state shifts

Second modeling assumption: steady state shift In our second modeling assumption, we assume that the two experimental situations which are compared correspond to two steady states of the biological process. Steady states are characterized by $F(X) = 0$. The steady state shift is the result of external influences on the process. For the observed nodes, the data give the total variation of the observed variables for a change of the exterior conditions.

This point of view, which is assumed in the rest of the paper, is not correct in all situations. It cannot be assumed for time series where observations are made at relatively close instants. In this situation, it is not true that the system has reached steady state at each instant when it is observed.

For our example, this hypothesis is justified since data come from metabolism of animals that where kept under normal diet or fasted during a long time before the measurements.

Third modeling assumption: small changes and linear response Our third modeling assumption is that changes of the exterior conditions are small and that a linear response approximation can be used. Because we are interested in the signs of variations only, the smallness condition can be replaced by a monotonicity condition : the sign of all partial derivatives $\frac{\partial F_i}{\partial X_j}$ is considered to be constant within the experimental range of variation of the external conditions.

Discrete Laplace equation For a node i which is not in the entering boundary, we derive an equation for variations by differentiation of the steady state equation:

$$\sum_{j \in G,(j,i) \in A} \frac{\partial F_i}{\partial X_j} \delta X_j \ = \ 0.$$

A similar equation is derived for a node i in the boundary, implying the total influence of the exterior:

$$\sum_{j \in G,(j,i) \in A} \frac{\partial F_i}{\partial X_j} \delta X_j = - \sum_{k \in I \setminus G,(k,i) \in A} \frac{\partial F_i}{\partial X_k} \delta X_k.$$

2.4 Main mathematical result

Our main result allows one to describe the effects on subnetworks of changes of external conditions. The quantitative expression of this result is the following theorem, whose proof is given in the Appendix and which follows from the discrete Laplace equation.

Theorem 1
Let G be a subnetwork with simple entering boundary $\urcorner^{in}G$.

1. *The response of $i \in G \setminus \urcorner^{in}G$ to small changes on the entering boundary satisfies the following relation:*

$$\delta X_i = \sum_{j \in \urcorner^{in} G} \sum_{j \rightsquigarrow i \in \mathcal{P}} \frac{a_{j \rightsquigarrow i}}{C_{j \rightsquigarrow i}} \delta X_j, \qquad (2.1)$$

where $\mathcal{P}$ denotes the set of elementary path included in G, $a_{j \rightsquigarrow i} = \prod_{k=0}^{n-1} \frac{\partial F_{i_{k+1}}}{\partial X_{i_k}}$ and $C_{j \rightsquigarrow i}$ denotes a coefficient depending on the Jacobian matrix.

2. *If the network contains no positive loop, then all $C_{j\rightsquigarrow i}$ are strictly positive.*

In our qualitative analysis the only available information in Eq. 2.1 is the sign of some terms. To cope with this, the following corollary of Theorem 1 will be our main tool. Remind that $s(\delta X_i)$ stands for the sign of the variation of X_i.

Corollary 1
The linear response of a subnetwork G with no positive loops and simple entering boundary $\top^{in}G$ satisfies the following relation in the sign algebra:

$$\forall i \in G \setminus \top^{in}G,\ s(\delta X_i) = \sum_{j\in\top^{in}S} \sum_{j\rightsquigarrow i} s(j\rightsquigarrow i)s(\delta X_j), \tag{2.2}$$

By sign algebra, we mean that + and − are embedded in the set of subsets $\{\{+\},\{-\},?\}$, where $? = \{+,-\}$ means undetermined sign. The subsets are then provided with the composition rules: $\{+\}+\{-\} =?$, $\{+\}+\{+\} = \{+\}$, $\{-\}+\{-\} = \{-\}$, $\{+\}\times\{-\} = \{-\}$, $\{+\}\times\{+\} = \{+\}$, $\{-\}\times\{-\} = \{+\}$. An equality in an equation means that the sign corresponding to the l.h.s. and the sign of the r.h.s. have a non empty intersection [Kui94, TMD03].

Corollary 1 applies to any subnetwork. In particular, for the subnetwork $\{i\}\cup pred(i)$ we obtain:

Corollary 2
If the set of predecessors of the node i is a simple incoming shell, then the following relation holds in the sign algebra:

$$s(\delta X_i) = \sum_{k\in pred(i)} s(k,i)s(\delta X_k). \tag{2.3}$$

Supposing that the predecessors of all nodes that are not successors of $\mathcal{E}$ are simple incoming shells, we can use Corollary 2 for our analysis. Without having got a full proof yet, we conjecture that solutions of systems generated by this corollary should fulfill also all equations of the type of Corollary 1.

The equations that Corollary 2 provides for the working example I_1 (including SCAP) are the following (for ease of notations, $s(\delta A)$ is denoted A).

System 1
```
(1) PPAR-a = PPAR + PUFA
(2) LXR-a = -PUFA + LXR
(3) SREBP = LXR-a
(4) SREBP-a= SREBP + SCAP
(5) SCD1 = LXR-a  + SREBP-a + PPAR-a
(6) FAS=  LXR-a + SREBP-a - PPAR-a
(7) ACL= LXR-a  + SREBP-a - PPAR-a
(8) ACC= LXR-a + SREBP-a - PPAR-a
```

Known values are the following:

Values 1
```
PPAR=+, PUFA=+, LXR=+, SREBP=-, SCD1=-,
FAS=-, ACL=+, ACC=-
```

The system associated with the example I_2 (excluding SCAP) is similar to System 1 except for equation (4).

System 2
```
...
(4) SREBP-a= LXR-a
...
```

3 Qualitative analysis

We develop here three approaches of qualitative analysis. The first one solves System 1. The second approach proposes an algorithmic research for solutions and incompatibilities in the system obtained from Corollary 2. The last one uses the concept of pure balances.

3.1 Solution of the system of qualitative equations

Let us solve System 1 by using the rules of the sign algebra.

Property 1
System 1 is compatible. LXR-a and SCD1 are balanced in the negative direction. The variation of SCAP has to be positive.

System 2 is incompatible. Simple change of data on ACL renders it compatible. LXR-a and SCD1 are balanced in the negative directions, FAS and ACC are not balanced.

Proof From Eq.1 PPAR-a=+. From Eq.3 LXR-a=SREBP=−.

We have already a biologically interesting result: by Eq.2, LXR-a is supposed to be undetermined, that is, the result of a balance between a positive action induced by LXR and a negative action induced by PUFA. Eq.3 implies that the balance on LXR-a should incline towards the negative direction.

From Eq.5 it follows that there is a balance on SCD1. Data on SCD1 imply that this balance incline towards the negative direction.

The only possibility for Eq.7 to be compatible is SREBP-a=+. From Eq.4, this is possible provided that SCAP=+. Then SCD1,FAS,ACL,ACC and SREBP-a result from balances.

In System 2, Eq.4 implies that SREBP-a=−, which is inconsistent with the value forced by value of A-CL and Eq.7. Eqs.6,8 together with data values on FAS, ACC imply that no balance is necessary in lhs of Eqs.6,8 in order to get the correct signs of FAS and ACC. Therefore there is necessarily an error, either in the interactions or in the data. Supposing that interactions are correct, considering that ACL is wrong is the simplest way to make the system compatible (because Eqs.6 and 8 do not force signs of FAS and ACC). ■

Such a result is the most precise that one can deduce from the qualitative system of equations. However, such an analysis can be done only on small examples. In the two next subsections, we intend to develop two algorithmic methods that allow to detect first, incompatibilities, second, places of balances and, if possible, the orientation of balances.

3.2 Detection of incompatibilities: graph valuation algorithm

The set of solutions of a linear qualitative system such as System 1 has in general a complex structure. Moreover, a solution of the qualitative system might give only limited information on the real system which it comes from. The point that remains always true is: a solution of the quantitative system always gives a solution of the qualitative one. So, it is interesting to show that a qualitative system has no solution compatible with the observations. This implies a contradiction between our model and the observations.

Solving the qualitative system is equivalent to giving qualitative values to each node, such that observed nodes are given the observed value and Eqs. 2.3 are satisfied. Such a valuation will be called compatible.

Property 2
In a network with no positive loop, if the set of predecessors of any node i which is not a successor of $\mathcal{E}$ is a simple incoming shell, the graph valuation algorithm described below provides a new graph such that any compatible valuation of the nodes in the initial graph is a compatible valuation of the new graph.

Basic ideas underlying the algorithm One natural means to reduce the number of unknowns in a system of equations is to compute each unknown in every equation where it is possible and substitute the value in the other equations. Incompatibilities may appear when two equations lead to two different values for one unknown.

In qualitative algebra, the usual rules for solving equations do not apply. The algorithm applies, as far as possible, two safe rules. Unfortunately, these rules are not always sufficient for the problem of contradiction detection.

Eqs. 2.3 shows that the sign of δX_i can be computed if for all predecessor k, the signs of δX_k are known and all the terms $s(k,i)s(\delta X_k)$ have the same value. This elementary rule is used in the forward propagation step of the algorithm.

Let us suppose that we have $\delta X_i = \sum_{k\in K_1} s(k,i)s(\delta X_k) + \sum_{k\in K_2} s(k,i)s(\delta X_k)$, where $s(\delta X_i)$ is known as is the sign of $\sum_{k\in K_1} s(k,i)s(\delta X_k)$ and the two signs are opposite ($K_1 \cup K_2 = pred(i)$, $K_1 \cap K_2 = \emptyset$). Then the sign of $\sum_{k\in K_2} s(k,i)s(\delta X_k)$ is equal to $s(\delta X_i)$. This can be represented in a graph by adding a new node with value $s(\delta X_i)$ and edges such that the equation $\sum_{k\in K_2} s(k,i)s(\delta X_k) = s(\delta X_i)$ is implied by relation 2.3. This is the basic step of backward propagation when K_2 has more than one element and K_1 is not empty.

If K_2 contains only one node k_0 with no value, we must have $s(\delta X_{k_0}) = s(k_0,i)s(\delta X_i)$ with the preceding configuration and Eq. 2.3.

Application to the detection of incompatibilities Since the valuations introduced by the algorithm use only local application of Eq. 2.3 and computed values are not changed, the resulting graph may contain incompatibilities. These incompatibilities are detected simply by checking Eq. 2.3 on each node.

Biological application 1
If the graph output by the algorithm described below contains a node where Eq. 2.3 is not satisfied, then experimental data and the model are incompatible.

Graph valuation algorithm Although not necessary, a breadth-first traversal of the graph is more efficient in forward propagation. For the sake of simplicity, we don't keep track of the nodes used in the rules for localization of incompatibilities. This is easy to do after implementation of the main steps of the algorithm.

1. *Initialization.* All the nodes are given a value in $\{+,-,?\}$. Observed nodes are initialized with the value corresponding to the observed variations of the concentration of the molecules. The other nodes are initialized with the value ?.

2. *Forward propagation.* Let F be a FIFO buffer for nodes. Put all the successors of the nodes with values different from ? in F. Let i the first node in F:

   - if $val(i) \neq ?$ remove i from F;
   - if $val(i) = ?$ and for all predecessors j of i, $val(j) \neq ?$ then
     $val(i) = \sum_{k \in pred(i)} s(k,i)val(k)$;
     remove i from F and put the successors of i in F;
   - if a predecessor of i has value ?, do nothing.

   This step ends when for each node i in F, there is at least a predecessor of i with value ? or F is empty.

3. *Backward propagation.* Examine all the nodes i with value different from ?. If i has at least a predecessor j with value ? and δX_i and $val(i) \times \sum_{k \in pred(i), s(\delta X_k) \neq ?} s(k,i)val(k)) = -$ then:

   - if j is the unique predecessor of i with value ? then $val(j) = s(j,i)val(i)$;
   - if there is more than one node as j, add a new node n with value $val(i)$, remove the edges (j,i) and add an edge (n,i) with $s(n,i) = +$, and edges (j,n) with $s(j,i) = s(j,n)$.

4. *The loop.* Begin by a forward propagation. Alternate forward and backward propagations until a loop doesn't add any new value.

   The algorithm always halts because when a node is given a value different from ? this value is never changed and each propagation step tends to decrease the number of nodes with value ?.

Working example including SCAP In the example I_1 the initialization is PUFA=+,LXR=+,LXR-a=?,PPAR=+,PPAR-a=?,SREBP=-,SREBP-a=?,ACL=+,ACC=+,FAS=-,SCD1=-,SCAP=?.

The forward propagation implies PPAR-a=+ (from PUFA=+ and PPAR=+). The backward propagation gives LXR-a=- (from SREBP=-), SREBP-a=+ (from ACL=+ and LXR-a=-), SCAP=+ (from SREBP-a=+ and SREBP=-). Hence the system is compatible and determined.

Working example excluding SCAP Initialization for I_2 is similar as the one of I_1. The forward propagation implies PPAR-a=+ (from PUFA=+ and PPAR=+) and SREBP-a=- (from SREBP=-).

The backward propagation gives LXR-a=- (from SREBP=-), and LXR-a=+ (from ACL=+ and SREBP-a=-). A conflict occurs, so that the system is incompatible.

In example I_2 (without SCAP), Eq.2.3 is not satisfied in node SREBP. The algorithm could be improved in order to keep track of the nodes involved in application of the rules used to determinate the values of LXR-a and SREBP. In this case, the nodes are ACL, SREBP-a, SREBP and LXR-a. The nodes ACL and SREBP are observed and that incompatibility involves LXR-a and SREBP-a. The observations or the model may be wrong. In the later case, a modeling error should be tracked first in the subgraph containing the nodes involved in the incompatible valuation.

In example I_1 the algorithm does not detect any incompatibility. The node SCAP is given a value during the second backward propagation and Eq. 2.3 is checked everywhere. The decision to consider the action of SCAP as negligible in example I_2 was incompatible with data and accurately located by the algorithm.

Improvements The preceding algorithm is generally not complete for the resolution of linear qualitative systems. As a consequence, the conclusions of the algorithm must be considered only when it detects an incompatibility or all values are determined.

A complete set of rules, based on some kind of Gaussian elimination in qualitative algebra exists for square systems [TMD03]. The design of an algorithm based on these rules is under consideration and could improve the use of qualitative reasoning for biological networks.

3.3 Detection of balances

In this last section, we intend to go further in the notion of balances and their detection; this allows for a partial propagation of the information contained in the graph.

Indeed, as explained above, some nodes can have a value different from ? but the incoming edges $\sum_{k \in pred(i)} s(k,i)s(\delta X_k)$ is undetermined. This means that the node i is influenced by actions of different signs, that we call balances. Some of them derive from other balances, others are purely collected by the node. We intend to automatically detect pure balances and give their biological meaning.

Influences As an application of Theorem 1, we want to define the influence of observed nodes on the other nodes. For this purpose we restrict ourselves to

paths from the set $\mathcal{AP} = \{j \rightsquigarrow i = (j, j_1, \ldots, j_l = i),\ j \in \mathcal{O},\ j_n$ has a simple incoming shell for all $n > 0\}$.

Let $j \rightsquigarrow i = (j, j_1, \ldots, j_l = i) \in \mathcal{AP}$. Applying Theorem 1 for each subnetwork $\{j_n\} \cup pred(j_n)$ and replacing successively the variables implies that the influence of the path $j \rightsquigarrow i$ on i is given by the sign of the path

$$s(\delta X_i) = s(\delta X_j)s(j \rightsquigarrow i) + \sum_{k \in Pred(j \rightsquigarrow i)} s(\delta X_k)s(k \rightsquigarrow i), \tag{3.1}$$

where $Pred(j \rightsquigarrow i) = \{j \in \cup_{n>0} pred(j_n)\} \setminus \{j, j_1, \ldots, i\}$.

We say that $s(\delta X_j)s(j \rightsquigarrow i)$ is the *influence of* $j \rightsquigarrow i$ *on the node* i.

Let $\mathcal{I}^+(i)$ and $\mathcal{I}^-(i)$ be the set of positive and negative influences on i:

$$\mathcal{I}^+(i) = \{j \rightsquigarrow i \in \mathcal{AP},\ s(\delta X_j)s(j \rightsquigarrow i) = +\}$$
$$\mathcal{I}^-(i) = \{j \rightsquigarrow i \in \mathcal{AP},\ s(\delta X_j)s(j \rightsquigarrow i) = -\}.$$

Intuitively, $j \rightsquigarrow i \in \mathcal{I}^+(i)$ indicates that δX_j contributes positively to the variations of X_i through the path $j \rightsquigarrow i$. Negative contributions are in $\mathcal{I}^-(i)$.

Balances and dominated influences A node i is said to *carry a balance* as soon as both the sets of positive and negative influences are nonempty: $\mathcal{I}^+(i) \neq \emptyset$ and $\mathcal{I}^-(i) \neq \emptyset$, meaning that two paths provide contributions with opposite signs on i.

If a node i carries a balance, the final variation of the node tells which contributions has been dominated at this step by other contributions on i so that they can not influence the successors of i. This leads to define dominated paths on a node i: if i is observed, $j \rightsquigarrow i \in \mathcal{AP}$ is *potentially dominated* when its influence is not coherent with the observation on i, that is, $s(\delta X_j)(j \rightsquigarrow i) \neq s(\delta X_i)$. In the case of a nonobserved node, $j \rightsquigarrow i \in \mathcal{AP}$ is potentially dominated as soon as the node is balanced. This provides the following definition of positive and negative potentially dominated set $\mathcal{C}^+(i)$ and $\mathcal{C}^-(i)$ on the node i:

$$\begin{array}{rl} X_i \text{ observed}, & \left\{ \begin{array}{l} \mathcal{C}^\epsilon(i) = \emptyset \\ \mathcal{C}^{-\epsilon}(i) = I^{-\epsilon}(i) \end{array} \right. \\ \text{sign}(\delta X_i) = \epsilon & \\ X_i \text{ nonobserved}, & \left\{ \begin{array}{l} \mathcal{C}^+(i) = \mathcal{I}^+(i) \\ \mathcal{C}^-(i) = \mathcal{I}^-(i) \end{array} \right. \\ \mathcal{I}^+(i) \neq \emptyset \text{ and } \mathcal{I}^+(i) \neq \emptyset & \\ X_i \text{ nonobserved}, & \left\{ \begin{array}{l} \mathcal{C}^+(i) = \emptyset \\ \mathcal{C}^-(i) = \emptyset. \end{array} \right. \\ \mathcal{I}^+(i) = \emptyset \text{ or } \mathcal{I}^+(i) = \emptyset & \end{array}$$

To identify influences that are dominated by an intermediate node in a path, a path $j \rightsquigarrow i \in \mathcal{AP}$ is said to be a *carried influence* to i if it is a consequence of a previous potentially dominated influence; they are defined depending on their signs as follows:

$$\mathcal{T}^+(i) = \{j \rightsquigarrow i \in \mathcal{I}^+(i),\ \exists k \neq i \in j \rightsquigarrow i,\ j \rightsquigarrow k \in \mathcal{C}^{s(k \rightsquigarrow i)}(k)\},$$
$$\mathcal{T}^-(i) = \{j \rightsquigarrow i \in \mathcal{I}^-(i),\ \exists k \neq i \in j \rightsquigarrow i,\ j \rightsquigarrow k \in \mathcal{C}^{-s(k \rightsquigarrow i)}(k)\}.$$

Pure balanced influences A pure balance holds on i if not all influences coming to i are carried: a node is said to collects a pure balance if $\mathcal{B}^-(i)$ or $\mathcal{B}^+(i)$ are nonempty, with:

$$\mathcal{B}^+(i) = \mathcal{C}^+(i) \setminus \mathcal{T}^+(i),$$
$$\mathcal{B}^-(i) = \mathcal{C}^-(i) \setminus \mathcal{T}^-(i).$$

These are paths that are sure to be involved in a competitive process for the node i.

Consequently, a nonobserved node i collects a pure balance iff two competitive paths contribute to i, these path being not previously dominated. The result of the balance is undetermined.

Pure balances on an observed node i correspond to paths whose influence is contradictory to the observation on the variable, and such that no information allows to say that the influence was previously dominated by another path of the model. This implies that a new balance occurs at the node i: the variation of i tells which actions have dominated the others.

A special case appears when an observed node i collects a pure balance with $s(\delta X_i) = +$ and $\mathcal{T}^+(i) = \mathcal{I}^+(i)$ (or, symmetrically, $s(\delta X_i) = -$ and $\mathcal{T}^-(i) = \mathcal{I}^-(i)$). This suggests that the experimental data about X_i is incorrect, since reversing the sign of i suppresses the pure balance collected by i. The other possibility is that all "good" influences on i are carried by a balance on a predecessor of i; then i is a balanced node and pure influences in $\mathcal{B}^-(i)$ are all dominated by carried influences that belong to $\mathcal{T}^+(i)$. From the above discussion it follows:

Biological application 2
When the interaction graph contains no positive loop, if a node i collects a pure balance, that is, $\mathcal{B}^+(i) \neq \emptyset$ or $\mathcal{B}^-(i) \neq \emptyset$, then

- *either two pathways with a different sign act competitively on the node; this holds when $\mathcal{B}^+(i) \neq \emptyset$ and $\mathcal{B}^-(i) \neq \emptyset$;*
- *or there might be an error in the model or data; this is suggested when $\mathcal{B}^+(i) = \emptyset$ or $\mathcal{B}^-(i) = \emptyset$: if the variations of δX_i are reversed, i does not collect a pure balance anymore.*

In the case when the graph contains no loop, pure balanced nodes can be identified by a simple computation on the graph.

Property 3
In a network with no loop, if the set of predecessors of any node which is not a successor of $\mathcal{E}$ is a simple incoming shell, a node i collects a pure balance if and only if $B^+(i) \neq 0$ or $B^-(i) \neq 0$ in the pure balance computation described below. The balanced case in Theorem 2 occurs when $(I^+ - t^+)(I^- - t^-)(i) \neq 0$, the possible error case occurs when $(I^+ - t^+)(I^- - t^-)(i) = 0$.

Pure balance computation Let I^+, I^- denote the number of influences, t^+, t^- denote the number of transported dominated influences and B^+, B^- denote the number of pure balanced influences. Let also p^+ and $p-$ denote the experimental information.

No loop hypothesis The definition of pure balanced on a node i depends on balances collected by the predecessors of i. An algorithmic computation of their number can be done as soon as the graph is acyclic. If it is not acyclic, a special study of the combinatorics of the graph is needed to compute pure balances.

Initialization For every observed simple node on the boundary, let $I^+(i) = I^-(i) = t^+(i) = t^-(i) = B^+(i) = B^-(i) = 0$. For every observed node in G, let $p^+(i) = 1$ iff i is observed and $s(\delta X_i) = +$, $p^+(i) = 0$ otherwise. Let $p^-(i) = 1$ iff i is observed and $s(\delta X_i) = -$, $p^-(i) = 0$ otherwise.

Propagation We suppose that the interaction graph contains no loop, hence we can suppose that all the functions I^+, I^-, t^+, t^-, B^+, B^- are computed for every predecessor of a node i. Then the functions are computed as follows for the node i.

$I^+(i) = \sum_{k \in pred(i)} I^{s(k,i)}(k) + p^{s(k,i)}(k)$
$I^-(i) = \sum_{k \in pred(i)} I^{-s(k,i)}(k) + p^{-s(k,i)}(k)$
$t^+(i) = \sum_{k \in pred(i)} t^{s(k,i)}(k) + B^{s(k,i)}(k)$
$t^-(i) = \sum_{k \in pred(i)} t^{-s(k,i)}(k) + B^{-s(k,i)}(k)$
X_i observed, $s(\delta X_i) = \epsilon$
$$\begin{cases} B^\epsilon(i) = 0 \\ B^{-\epsilon}(i) = (I^{-\epsilon} - t^{-\epsilon})(i) \end{cases}$$
X_i nonobserved, $(I^+ - t^+)(I^- - t^-)(i) = 0$
$$\begin{cases} B^+(i) = 0 \\ B^-(i) = 0 \end{cases}$$
X_i nonobserved, $(I^+ - t^+)(I^- - t^-)(i) \neq 0$
$$\begin{cases} B^+(i) = (I^+ - t^+)(i) \\ B^-(i) = (I^- - t^-)(i). \end{cases}$$

Working example A computation on the graph related to the working examples I_1 and I_2 of regulation of lipid synthesis show that 7 nodes among 11 are balanced in both models: LXR-a, SREBP, SREBP-a, ACL, ACC, FAS, SCD1. There are 14 potentially dominated path arriving in an observed node and 5 paths arriving in nonobserved nodes. They are all listed in the Appendix.

The balance computation shows that among the 19 potentially balanced paths, only 6 appear to be pure (see the list in the Appendix). Pure balanced nodes are LXR-a, ACL and SCD1.

- LXR-a collects two competitive paths, so that this node is in the nonobserved balanced situation.
- ACL collects two pure negative influences and no pure positive influence. This suggests that the data on ACL is false: when one suppose that the variation is positive, no more pure balance occurs on ACL. If the data on ACL is proved to be true, then the system has to be studied more precisely, as in the previous subsection, to check whether there is a real incompatibility between the model and the data.
- SCD1 is an observed node that collects two pure negative influences and one pure positive influence. This suggests that a real balance occurs on SCD1. The observed positive variation on SCD1 suggests that the negative pathways through S-REBP dominates the positive pathway through PPAR.

Consequently, two points can be extracted from such an analysis: at the SCD1 level, the negative pathways through SREBP dominates the positive pathway through PPAR; at the ACL level, the data on ACL should be checked carefully or the model about edges leading to ACL should be studied in detail.

4 Remarks and conclusion

In this paper we showed how a qualitative theoretical model of a mixed genetic/metabolic network can be confronted to experimental data such as DNA-microarray data.

Using a linear response approximation we have been able to obtain quantitative equations that should be satisfied by the theoretical variations of genes and metabolic products under changes of the external conditions. These equations are analogous

to the discrete Laplace equations on graphs that have been studied in the context of Markov chains [CY00]. In order to calculate the variations we have to compute the inverse of the Jacobian. This inverse is in fact the discrete Green function of the problem.

Contrary to [CY00] we do not want explicit formulas for the Green function (difficult to obtain for irregular graphs), but we want to emphasize how one node collects influences from the other nodes. This formulation of the problem is particularly adapted for qualitative analysis and explicitly uses the topology of the interaction graph. In particular, we find that if the network has no positive loops then the signs of the influences are just products of the signs of the interactions along the paths of the graph, which coincides to wide-spread intuition.

Our purpose was to show the possibilities of this approach, rather than to present it in full generality here. We have left the discussion of the case of positive loops (that can produce counter-intuitive phenomena such as negative impedances) to a future publication.

We have developed three complementary methods to analyze the compatibility between the model and the experimental data.

- The first one consists in solving a qualitative system of equations. Incompatibilities, balances and potential errors can be obtained with this approach, but we have not yet investigated general algorithms and automatization.
- The graph valuation algorithm allows one to reduce the qualitative system of equations so that it can be studied more easily. The valuation is an automated method to partially or totally solve the qualitative equations. The existence of a solution allows one to predict unknown variations. In the absence of a solution it is possible to detect incompatibilities between the model and the data and the place where these incompatibilities occur.
- The pure balance computation allows one to detect the variables in the system that really are influenced by competitive pathways and to find which pathway wins the competition; eventually, this computation detects the presence of errors in the data.

The analysis of the examples taken from the regulation of lipid metabolism suggests that this type of methods could be used to optimize experimental protocols. The optimization criteria are not formalized here. Nevertheless, as a general rule, we would like to have entering boundaries that are completely observed. In one of the examples, SCAP (which is on the entering boundary of any set containing SREBP-a and its predecessors) is not observed. Instead, we have predicted the sign of its variation. The experimental knowledge of SCAP variation would have made possible the comparison to the theoretical value and the detection of existing incompatibilities.

Appendix: Proofs

Proof of Theorem 1 Restricting the steady state equations to the subset G one gets:

$$F_j(\{X_j\},\{X_i\}) = 0, j \in G, i \in I \setminus G. \tag{4.1}$$

By differentiation of Eq.4.1 we find that:

$$\sum_{j\in G,(j,i)\in A} \frac{\partial F_i}{\partial X_j}\delta X_j = -C_i \delta X_i^f, \tag{4.2}$$

where:

$$\delta X_i^f = \begin{cases} \frac{1}{C_i}\sum_{k\in I\setminus G,(k,i)\in A} \frac{\partial F_i}{\partial X_k}\delta X_k & \text{if } i \in \urcorner^{in} G \\ 0 & \text{if not.} \end{cases} \tag{4.3}$$

Eq. 4.3 emphasizes the physical significance of the real constants C_i that by analogy with electrical networks can be called boundary impedances (ratios of external forces and forced variations on the boundaries).

From Eq. 4.2 it follows:

$$\delta X_i = -\sum_{j\in G} B_{ij} C_j \delta X_j^f, \tag{4.4}$$

where $B = \left([\frac{\partial F_i}{\partial X_j}]_{i,j\in G}\right)^{-1}$ is the inverse of the Jacobian matrix.

Let us consider the following well-known property of the inverse matrix [Blo79]:

$$A_{ij}^{-1} = (-1)^{i+j}\frac{det(M_{ji})}{detA} \tag{4.5}$$

where M_{ij} is the minor obtained by deleting line i and column j in matrix A.

Furthermore, it is a simple exercise [Blo79] to show that:

$$M_{ji} = \begin{cases} (-1)^{i+j}\sum_{j\sim i}(-1)^{l_{j\sim i}-1} a_{j\sim i} det(A_{(j\sim i)^c}) & \\ \qquad\qquad \text{if } i \neq j & \\ detA_{\{j\}^c} \qquad \text{if } i = j. & \end{cases} \tag{4.6}$$

where $j \rightsquigarrow i$ is any path leading from j to i with no loop, $l_{j\rightsquigarrow i}$ is the number of nodes in the path, $a_{j\rightsquigarrow i}$ is the product of elements of A along this path, $A_{(j\rightsquigarrow i)^c}$ is the principal minor defined by the set of indices complementary to those in the path.

If i is not on the entering boundary the sum in Eq.4.4 contains only terms with $j \neq i$, because $\delta X_i^f = 0$. Using Eqs.4.5,4.6 one gets: $\delta X_i = -\sum_{j\in{}^{in}G} \sum_{j\rightsquigarrow i} (-1)^{l_{j\rightsquigarrow i}-1} a_{j\rightsquigarrow i} C_j \frac{detA_{(j\rightsquigarrow i)^c}}{detA} \delta X_j^f$.

If i is on the entering boundary, and if this boundary is simple, then the sum in Eq.4.4 contains only one term: $\delta X_i = -C_i \frac{detA_{ic}}{detA} \delta X_i^f$. In order to have $\delta X_i = \delta X_i^f$ one should have:

$$C_i = -\frac{detA}{detA_{\{i\}^c}} \tag{4.7}$$

This consistent choice of the constants C_i will be adopted throughout the rest of the paper yielding:

$$\delta X_i = \begin{cases} \sum_{j\in{}^{in}G} \sum_{j\rightsquigarrow i} \frac{a_{j\rightsquigarrow i}}{C_{j\rightsquigarrow i}} \delta X_j^f & \\ \qquad\qquad \text{if } i \notin {}^{in}G & \\ \delta X_i^f + \sum_{j\in{}^{in}G, j\neq i} \sum_{j\rightsquigarrow i} \frac{a_{j\rightsquigarrow i}}{C_{j\rightsquigarrow i}} \delta X_j^f & \\ \qquad\qquad \text{if } i \in {}^{in}G & \end{cases} \tag{4.8}$$

where $C_{j\rightsquigarrow i} = (-1)^{l_{j\rightsquigarrow i}-1} \frac{detA_{\{j\}^c}}{detA_{\{j\rightsquigarrow i\}^c}}$.

Remark 1 If the entering boundary is simple then Eq.4.8 becomes $\delta X_i = \delta X_i^f$ when $i \in {}^{in}G$.

Remark 2 Eq.4.8 is a generalization of the fact that solutions of Laplace equation on a graph are entirely determined by their values on the boundary [CY00, TS90]. Intuitively, one would expect that the signs of terms in the sum should be the signs of $a_{j\rightsquigarrow i}\delta X_j^f$ but this is not true in general. We have the following property:

Property 4
If the interaction graph contains no positive loops, then

$$C_{j\rightsquigarrow i} > 0$$

which means that the terms in the $j \rightsquigarrow i$ sum in Eq.4.8 have the signs of $a_{j\rightsquigarrow i}\delta X_j^f$.

Proof If there are no positive loops, then it can be shown (see Soulé [Sou03]) that the signs of principal minors of the Jacobian are alternating. Thus,

$$\begin{aligned} sign(detA_{\{j\}^c}) &= (-1)^{\#G-1} \\ sign(detA_{(j\rightsquigarrow i)^c}) &= (-1)^{\#G-l_{j\rightsquigarrow i}}. \end{aligned}$$

■

Theorem 1 follows from Eq.4.8 and Property 4 ■

Proof of Property 3 The formulas given in the pure balance computation simply compute the number of influences and contradictions by using the information provided by the predecessor of a node. If i is a node, then positive influences $I^+(i)$ are the union of positive influences $I^+(k)$ on any predecessor of i with a positive arc, negative influences of any predecessor of i with a negative arc, and influences directly starting in k if it is an observed node.

Positive carried influences on i are given by carried influences on each predecessor of i with the appropriate sign and balances identified on these predecessors. Then one simply has to compare carried influences on i with the total number of influences to get pure balanced influences. ■

List of potentially balanced paths in the working example In the working example, there are 14 potentially dominated paths arriving in an observed node.

LXR[+] →+ LXR-a →+ SREBP[-]
LXR[+] →+ LXR-a →+ ACC[-]
LXR[+] →+ LXR-a →+ SREBP →+ SREBP-a →+ ACC[-]
LXR[+] →+ LXR-a →+ FAS[-]
LXR[+] →+ LXR-a →+ SREBP →+ SREBP-a →+ FAS[-]
LXR[+] →+ LXR-a →+ SCD1[-]
LXR[+] →+ LXR-a →+ SREBP →+ SREBP-a →+ SCD1[-]
PUFA[+] →+ PPAR-a →− SCD1[-]
PPAR[+] →+ PPAR-a →− SCD1[-]
PPAR[+] →+ PPAR-a →− ACL[+]
PUFA[+] →+ PPAR-a →− ACL[+]
PUFA[+] →− LXR-a →+ ACL[+]
PUFA[+] →− LXR-a →+ SREBP →+ SREBP-a →+ ACL[+]
SREBP[-] →+ SREBP-a →+ ACL[+]

There are also 5 potentially balanced path related to nonobserved nodes.

LXR[+] →+ LXR-a
PUFA[+] →− LXR-a
LXR[+] →+ LXR-a →+ SREBP →+ SREBP-a
PUFA[+] →− LXR-a →+ SREBP →+ SREBP-a
SREBP[+] →+ SREBP-a

Pure balanced paths are the following.

LXR[+] →+ LXR-a
PUFA[+] →− LXR-a
PUFA[+] →+ PPAR-a →− SCD1[-]
PPAR[+] →+ PPAR-a →− SCD1[-]
SREBP[-] →+ SREBP-a →+ ACL[+]
PPAR[+] →+ PPAR-a →− ACL[+]
PUFA[+] →+ PPAR-a →− ACL[+]

References

[BG03] Hogue CW Bader GD, Betel D. Bind: the biomolecular interaction network database. *Nucleic Acids Res.*, 31(1):248–50, 2003.

[Blo79] DM. Bloom. *Linear algebra and geometry.* Cambridge University Press, Cambridge, 1979.

[CY00] F. Chung and ST. Yau. Discrete green's functions. *J. Combin. Theory Ser.A*, 91:191–214, 2000.

[KF02] N. Kaminski and N. Friedman. Practical approaches to analyzing results of microarray experiments. *American Journal of Respiratory and Cell Molecular Biology*, 27:125–13, 2002.

[Kui94] B.J. Kuipers. *Qualitative reasoning. Mdeling and simulation with incomplete knowledge*. MIT Press, 1994.

[LACa04] C. Lemer, E. Antezana, F. Couche, and al. The amaze lightbench: a web interface to a relational database of cellular processes. *Nucleic Acids Res.*, 32:D443–D448, 2004.

[NRF04] I. Nachman, A. Regev, and N. Friedman. Inferring quantitative models of regulatory networks from expression data. *Bioinformatics*, 20:i248 – i256, 2004.

[PLMG04] JP Pegorier, C Le May, and J. Girard. Control of gene expression by fatty acids. *J Nutr.*, 134(9):2444S–2449S, 2004.

[PSP+04] JA Papin, J Stelling, ND Price, S Klamt, S Schuster, and Palsson BO. Comparison of network-based pathway analysis methods. *Trends in Biotechnology*, 22:400–405, 2004.

[Sou03] C. Soulé. Graphic requirements for multistationarity. *Complexus*, 1(123-133), 2003.

[TMD03] L. Travé-Massuyès and P. Dague, editors. *Modèles et raisonnements qualitatifs*. Hermes sciences, 2003.

[TS90] AN. Tikhonov and AA. Samarskii. *Equations of mathematical physics*. Dover, New York, 1990.

[YVK04] Y. Yamanishi, J.-P. Vert, and M. Kanehisa. Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics*, 20:i363 – i370, 2004.