

Complex Qualitative Models in Biology: A New Approach

P. Veber^a M. Le Borgne^a A. Siegel^a S. Lagarrigue^b
O. Radulescu^c

^aProjet Symbiose, Institut de Recherche en Informatique et Systèmes Aléatoires, IRISA-CNRS 6074, Université de Rennes 1, ^bUMR Génétique animale, Agrocampus Rennes-INRA et ^cInstitut de Recherche Mathématique de Rennes, UMR-CNRS 6625, Université de Rennes 1, Rennes, France

Key Words

Equilibrium shift · Qualitative algebra · Decision diagrams · Experiment design

Abstract

We advocate the use of qualitative models in the analysis of large biological systems. We show how qualitative models are linked to theoretical differential models and practical graphical models of biological networks. A new technique for analyzing qualitative models is introduced, which is based on an efficient representation of qualitative systems. As shown through several applications, this representation is a relevant tool for the understanding and testing of large and complex biological networks.

Copyright © 2005 S. Karger AG, Basel

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

KARGER

© 2005 S. Karger AG, Basel
1424-8492/05/0024-0140
\$22.00/0
Accessible online at:
www.karger.com/cpu

Dr. Philippe Veber, Projet Symbiose, Institut de Recherche en Informatique et Systèmes Aléatoires
IRISA-CNRS 6074, Université de Rennes 1, Campus de Beaulieu
FR-35042 Rennes Cedex (France)
Tel. +33 299 847 100, Fax +33 299 847 171,
E-Mail Philippe.Veber@irisa.fr

Simplexus

Biological Extrapolation

In the past, biologists characterized whole organisms, identifying them, describing them, and classifying them. The discovery of DNA changed all that, and biology gained a prefix – molecular. While there are still countless biologists studying whole organisms, today a biologist is just as likely to investigate the interactions of molecules as members of a species.

In the present paper, Veber and colleagues propose a method for studying and computing steady-state shifts of differential systems describing biological networks. Critically, the researchers suggest that qualitative models can be linked to theoretical models of biological networks and might ultimately lead to new insights into understanding large biological systems, by which we mean many interacting molecular species.

Given recent major developments in molecular biology such as microarrays, mass spectrometry, and protein chips, the development of a quantitatively robust method for coping with putatively thousands of variables in any given biological system simultaneously would be rather timely. Network-based representations have been used widely to describe gene regulation or metabolic pathways at the cellular level. Network reconstruction relies on gathering vast amounts of quantitative data from microarrays. However, these analytical techniques are still imperfect and sensitive to interference and noise, therefore not always quantitatively reliable. Available experimental data can only be interpreted on qualitative grounds. Microarrays, for instance, which are used to compare gene activity, only tell us that gene G is more active in situation A than in situation B.

Veber and his colleagues wanted to cope with this situation. Previously, they devised a novel framework for comparing two sets of experimental conditions – situation A and situation B, for example. With

1 Introduction

Understanding the behaviour of a biological system from the interplay of its molecular components is a particularly difficult task. A model-based approach proposes a framework to express some hypotheses about a system and make some predictions out of it, in order to compare with experimental observations. Traditional approaches [see 1 for an interesting review] include ordinary differential equations or stochastic processes. While they are powerful tools to acquire a fine grained knowledge of the system at hand, these frameworks need accurate experimental data on chemical reaction kinetics, which are scarcely available. Furthermore, they also are computationally demanding and their practical use is restricted to a limited number of variables.

As an answer to these issues, many approaches were proposed that abstract from quantitative details of the system. Among others, let us stress the work done on gene regulation dynamics [2], hybrid systems [3] or discrete event systems [4, 5]. The goal of such qualitative frameworks is to enable system level analysis of a biological phenomenon. This appears as a relevant answer to a recent technical breakthrough in experimental biology:

- microarrays, mass spectrometry, protein chips currently allow to measure thousands of variables simultaneously
- obtained measurements are rather noisy, and may not be quantitatively reliable

Microarrays, for instance, are used for comparing the activity of genes between two experimental settings. A microarray experiment gives a differential measure between two experimental settings. It delivers information on the relative activity of each gene represented on the array. Despite many attempts made to quantify the output of microarrays, the essential output of the technique says, for example, that a gene *G* is more active in situation *A* than in situation *B*.

In this paper, we use a framework developed by Siegel et al. [6] for the comparison of two experimental conditions, in order to derive qualitative constraints on the possible variations of the variables. Our main contribution is the use of an efficient representation for the set of solutions of a qualitative system. This representation allows to solve systems with hundreds of variables. Moreover, this representation opens the way to a finer analysis of qualitative systems. This new approach is illustrated by solving three important problems:

- checking the accordance of a qualitative system with qualitative experimental data
- minimally correcting corrupted data in discordance with a model
- helping in the design of experiments

Our main focus here is to show how to use large qualitative models and qualitative interpretations of experimental data. In this respect our work could be used as an extension to what was proposed by Gutierrez-Rios et al. [7], where basically the authors propose to analyze pangenomic gene expression arrays in *Escherichia coli*, using simple qualitative rules.

In the first section we establish links between differential, graphical and qualitative models.

2 Mathematical Modelling

In this section we show how qualitative models can be linked to more traditional differential models. Differential models are central to the theory of metabolic control [8, 9]. They also have been applied to various aspects of gene network dynamics. The purpose of this section is to lay down a set of qualitative equations describing steady-state shifts of differential models. For the sake of completeness, we rederive in a simpler case results that have been established in greater generality [6, 10].

2.1 Modelling Assumptions

Let us consider a network of interacting cellular constituents, numbered from 1 to n .

this approach they side-step the data and gene perturbation information from microarrays and their ilk and build a network not from scratch, but by gathering information from the literature. This allowed them to analyze even incomplete models and to compare models and data. As such, the emergent mathematical results connect network topology and information from steady-state shift experiments, which are used by chemists to reveal reaction mechanisms.

Gene networks are essentially graphical models of molecular interactions involved in the expression machinery of genes. The particular genes being used by a cell under specific conditions depend on these interactions and the set of active genes reveal a cell's specialization. Liver cells, for instance, express different genes to neurons or muscle cells. In order to understand cell function a good model is essential as too is knowledge of the topology of the networks that allow the cell to function. Each one of these problems is difficult and there are no well-established methods to solve them, as there are no machines that photograph or read gene networks. Gene network reconstruction is a painful process of knowledge accumulation and consists in interpreting the response of genes to perturbations such as deleting or silencing some genes, overexpressing other genes or changing the external conditions. In this process errors are inevitable.

Qualitative methods, however, have been used for error detection and correction since the 1980s in various applications such as electronic circuit design and diagnosis. In the current paper, Veber and his colleagues have built on this approach and developed a method for applying qualitative constraints to possible deviations in a gene network model. They have created an efficient way of representing the set of solutions of a qualitative system, which they explain, allows one to solve systems with hundreds of variables.

These constituents may be proteins, RNA transcripts or metabolites for instance. The state vector X denotes the concentration of each constituent.

Differential Dynamics

X is assumed to evolve according to the following differential equation:

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

where F is an (unknown) non-linear, differentiable function. A steady-state X_{eq} of the system is a solution of the algebraic equation:

$$F(X_{eq}) = 0.$$

Steady states are asymptotically stable if they attract all nearby trajectories. A steady state is non-degenerated if the Jacobian calculated in that steady state is non-vanishing. According to the Grobman-Hartman theorem, a sufficient condition to have non-degenerated asymptotically stable steady states is $Re(\lambda_i) < -C, C > 0, i = 1, \dots, n$, where λ_i are the eigenvalues of the Jacobian matrix calculated at the steady state.

Experiment Modelling

Typical two-state experiments such as differential microarrays are modelled as steady-state shifts. We suppose that under a change of the control parameters in the experiment, the system goes from one non-degenerated stable steady state to another one. The output of the two-state experiment can be expressed in terms of concentration variations for a subset of products, between the two states. We suppose that the signs of these variations were proven to be statistically significant.

Interaction Graph

The only knowledge we require about the function F concerns the signs of the derivatives

$$\frac{\partial F_i}{\partial X_j}.$$

These are interpreted as the action of the product j on the product i . It is an activation if the sign is +, an inhibition if the sign is -. A null value means no action.

An interaction graph $G(V,E)$ is derived from the Jacobian matrix of F :

(1) with nodes $V = \{1, \dots, n\}$ corresponding to products and

(2) (oriented) edges

$$E = \{(j, i) \mid \frac{\partial F_i}{\partial X_j} \neq 0\}.$$

Edges are labelled by

$$s(j, i) = \text{sgn}(\frac{\partial F_i}{\partial X_j}).$$

The set of predecessors of a node i in G is denoted $\text{pred}(i)$. The interaction graph is actually built from information gathered in the literature. In consequence in some places it may be incomplete (some interactions may be missing), in others it may be redundant (some interactions may appear several times as direct and indirect interactions). It is an important issue that neither incompleteness nor redundancy do not introduce inconsistencies and this will be addressed in section 5.

Negative Diagonal in the Jacobian Matrix

For any product i , we exclude the possibility of vanishing diagonal elements of the Jacobian

$$\frac{\partial F_i}{\partial X_i}.$$

This can be justified by taking into account degradation and dilution (cell growth) processes that can be represented as negative self-loops in the interaction graph, that is for all $i, (i, i) \in E$ and $s(i, i) = -$.

Discussion

In our mathematical modelling we suppose that the system starts and ends in non-degenerated stable steady states. Of course this is not always the case for several reasons: there is too much waiting time to reach steady state; one can end up in a limit cycle and oscillate instead of reaching a steady state. All these possibili-

As proof of principle, the researchers have used an example derived from the physiology of higher organisms – the genetic regulation of fatty acid synthesis in liver. The liver has two ways of making fatty acids: saturated and monounsaturated fatty acids are produced from citrates via a four enzyme route, while polyunsaturated fatty acids (PUFA) are produced from essential fatty acids from the diet using just two enzymes. PUFAs play a key role in several biological functions, such as regulation of gene expression that in turn impacts on lipid, carbohydrate, and protein metabolism, various receptors mediate their actions.

The receptors involved in PUFA mediation are the variables in the model and are abbreviated as PPAR, LXR, and SREBP. The receptors are synthesized from the corresponding genes and are then modified to make the active forms. Another protein SCAP cleaves SREBP and interacts with yet another protein family denoted INSIG, which hints at just how complicated the molecular mechanism is. Also included in the model are the final products ACL, ACC, FAS, SCD1, D5D, D6D, and PUFAs themselves. All these variables then have different interactions: SREBP activates transcription of ACL, ACC, FAS, SCD1, D5D and D6D. LXR activates transcription of SREBP and FAS, and indirectly activates ACL, ACC and SCD1, and so on, all aimed at regulating the production of PUFA in the liver.

An experiment to investigate just how these systems work would involve re-feeding fasting laboratory animals and quantifying the various liver products using microarray analysis. Currently, there are no methods available that would allow researchers to feed the vast amounts of data retrieved from a study into an analytical model. Furthermore, quantitative data is incomplete and highly inaccurate. Qualitative modelling is a solution that might cope with incomplete and inaccurate data, explains Veber. Indeed, the present researchers are not the only ones to recognize this

ties should be considered with caution. Actually this hypothesis might be difficult to check from the two states only. Complementary strategies such as time series analysis could be employed in order to assess the possibility of limit cycle oscillations.

Positive self-regulation is also possible but introduces a supplementary complication. In this case for certain values of the concentrations degradation exactly compensates the positive self-regulation and the diagonal elements of the Jacobian vanish (this is a consequence of the intermediate value theorem). We can avoid dealing with this situation by considering that the positive self-regulation does not act directly and that it involves intermediate species. This is a realistic assumption because a molecule never really acts directly on itself (transcripts can be autoregulated but only via protein products). Thus, all nodes can keep their negative self-loops and all diagonal elements of the Jacobian can be considered to be non-vanishing. Although the positive regulation may imply vanishing higher order minors of the Jacobian, this will not affect our local qualitative equations.

2.2 Quantitative Variation of One Variable

We focus here on the variation of the concentration of a single chemical species represented by a component X_i of the vector X . Since we have adopted a *static* point of view, we are only interested in the variation of X_i between two non-degenerated stable steady states X_{eq}^1 and X_{eq}^2 independently of the trajectory of the dynamical system between the two states.

Let us denote by \hat{X}_i the vector of dimension n_i obtained by keeping from X all coordinates j that are predecessors of i in the interaction graph. Then, under some additional assumptions described and discussed by Radulescu et al. [10], we have the following result:

Theorem 2.1

The variation of the concentration of species i between two non-degenerated steady states X_{eq}^1 and X_{eq}^2 is given by

$$X_{eq_i}^1 - X_{eq_i}^2 = \int_S - \left(\frac{\partial F_i}{\partial X_i} \right)^{-1} \sum_{k \in \text{pred}(i)} \frac{\partial F_i}{\partial X_k} dX_k \quad (1)$$

where S is the segment linking

$$\hat{X}_{eq_i}^1 \text{ to } \hat{X}_{eq_i}^2.$$

Full proof is given by Radulescu et al. [10]. The above formula is a quantitative relation between the variation of concentrations and the derivatives

$$\frac{\partial F_i}{\partial X_j}.$$

Now our next move will be to introduce a qualitative abstraction of this relation.

2.3 Qualitative Equations

We propose here to study equation 1 in sign algebra. By sign algebra, we mean the set $\{+, -, ?\}$, where $?$ represents an undetermined sign. This set is provided with the natural commutative operations:

$$\begin{aligned} + + - = ? & \quad + + + = + & \quad - - - = - & \quad + \times - = - \\ + \times + = + & \quad - \times - = + & \quad ? + - = ? & \quad ? + + = ? \\ ? + ? = ? & \quad ? \times - = ? & \quad ? \times + = ? & \quad ? \times ? = ? \end{aligned}$$

Equality in sign algebra \approx is defined as follows:

\approx	+	-	?
+	T	F	T
-	F	T	T
?	T	T	T

Importantly, qualitative equality is not an equivalence relation, since it is not transitive. This implies that computations in qualitative algebra must be carried out with care. At least two major properties should be emphasized:

- If a term of a sum is indeterminate (?) then the whole sum is indeterminate.

problem. Boolean or multivalued logical models, for instance, also provide qualitative descriptions. Compared to these models Veber and his colleagues use an alternative algebra, better adapted to the problem of equilibrium shifts, and, finally, provide reasonable algorithms to code and solve qualitative equations.

In order to compare model and data, one needs a mathematical description of the relation between the two, a comparison calculus, in other words. This can be done by predicting how the equilibrium of the genetic regulation system changes under different conditions, for instance, how lipid metabolism changes during fasting and normal feeding. A gene network behaves similarly to an elastic medium, according to Veber and colleagues' findings. Perturbations somewhere in the network are transported along the graph in the direction of the regulation arrows and obey an analogue of quantitative elastic moduli. Nonetheless, under certain mild conditions, the signs of variations of gene expression is the qualitative sum of the signs of influences coming from its neighbours in the network. For instance, gene A is positively regulated by B and also by C. The qualitative equation is $dA = dB + dC$.

This is purely dependent on topology and does not imply the need for a choice of Boolean function of B and C (AND, OR, etc.) to give A as is required in a Boolean gene network (a parallel field of qualitative gene modelling). In three valued sign algebra $(+, -, ?)$ a + variation of A ($dA = +$) and a + variation of B ($dB = +$) is compatible either with $dC = +$ or with $dC = -$. Furthermore, the situation $dA = -, dB = +$ is only compatible with $dC = -$. The latter example is illustrative of the predictive power of qualitative equations: if one knows that A decreases and B increases, and one knows that the only influence on A comes from B and C, then there is no need to measure C, as it should decrease.

Veber and colleagues explain that the qualitative equations can be reduced to a

- If one hand of a qualitative equality is indeterminate, then the equality is satisfied whatever the value of the other hand is.

A *qualitative system* is a set of algebraic equations with variables in $\{+, -, ?\}$. A *solution* of this system is a valuation of the unknown which satisfies each equation, and in such a way that no variable is instantiated to $?$. This last requirement is important since otherwise any system would have trivial solutions (like all variables to $?$).

Theorem 2.2

Under the assumptions and notations of theorem 2.1, if the sign of

$$\frac{\partial F_i}{\partial X_j}$$

is constant, then the following relation holds in sign algebra:

$$s(\Delta X_i) \approx \sum_{k \in \text{pred}(i)} s(k, i) s(\Delta X_k) \quad (2)$$

where $s(\Delta X_k)$ denotes the sign of

$$X_{eq_k}^1 - X_{eq_k}^2.$$

By writing equation 2 for all nodes in the graph, we obtain a system of equations on signs of variations, later referred to as *qualitative system* associated with the interaction graph G . This will be used extensively in the next sections.

2.4 Link between Qualitative and Quantitative

The qualitative system obtained from equation 2 is a consequence of the quantitative relations that result from Theorem 2.1. So the sign function maps a quantitative variation between two equilibrium points onto a qualitative solution of equation 2. The converse is not true in general. For a given solution S of the qualitative system, there might be no equilibrium change ΔX in the differential quantitative model, such that each real-valued component of ΔX has the sign given by S .

However, some components of the solution vectors are uniquely determined by the qualitative system. They take the same sign value in every solution vector. For such so-called hard components, the sign of any quantitative solution (if it exists) is completely determined by the qualitative system.

We will use the previous properties to check the coherence between models and experimental data. By experimental data we mean the sign of the observed variation in concentration for some nodes. In particular, if the qualitative system associated with an interaction graph G has no solution given some experimental observations, then no function F satisfying the sign conditions on the derivatives can describe the observed equilibrium shift, meaning that either the model is wrong or some data are corrupted. In the next section, we introduce a simplified model related to lipid metabolism, and illustrate the above-described formalism.

3 Toy Example: Regulation of the Synthesis of Fatty Acids

In order to illustrate our approach, we use a toy example describing a simplified model of genetic regulation of fatty acid synthesis in liver. The corresponding interaction graph is shown in figure 1.

Two ways of production of fatty acids coexist in the liver. Saturated and monounsaturated fatty acids are produced from citrates thanks to a metabolic pathway composed of four enzymes, namely ACL (ATP citrate lyase), ACC (acetyl-coenzyme A carboxylase), FAS (fatty acid synthase) and SCD1 (stearoyl-CoA desaturase 1). Polyunsaturated fatty acids (PUFA) such as arachidonic acid and docosahexaenoic acid are synthesized from essential fatty acids provided by nutrition; D5D (delta-5 desaturase) and D6D (delta-6 desaturase) catalyze the key steps of the synthesis of PUFA.

PUFA plays pivotal roles in many biological functions; among them, they regu-

finite set of clauses in polynomial time, so resolution of this qualitative system is an NP-complete problem. However by encoding the qualitative equations associated with the problem as algebraic equations, they offer a solution that can be computed with greatly reduced time and complexity and so without the need for vast computer power.

One additional benefit of the technique developed by Veber and colleagues is in experiment design. Models are built on accumulated knowledge. However, one would like to optimize the process by choosing the most economic way to gain the necessary knowledge. For instance, one might decide which genes to observe or perturb in an experiment. The choice is based on an estimate of the amount of information that such an experiment would bring. In the approach of Veber and colleagues, this choice is connected to the number of solutions of the qualitative equations constrained by a certain result of the experiment. The higher this number, the less informative will be the experiment and so an alternative route could be taken.

The researchers now plan to validate their approach further on metabolic pathways in yeast and *Escherichia coli*. These organisms have large pathways represented by vast quantities of microarray data that are publicly available. Despite the scale of such databases, Veber and colleagues have found that 200 variable sets can be handled within minutes.

David Bradley of Sciencebase.com

late the expression of genes that impact on lipid, carbohydrate, and protein metabolism. The effects of PUFA are mediated either directly through their specific binding to various nuclear receptors (PPAR α = peroxisome proliferator-activated receptors, LXR α = liver X receptor α , HNF-4 α) leading to changes in the transactivating 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.) [11].

3.1 Variables in the Model

We consider in our model nuclear receptors PPAR α , LXR α , SREBP-1c (denoted by PPAR, LXR, SREBP, respectively, in the model), as they are synthesized from the corresponding genes and the transactivating active forms of these transcription factors, that is LXR-a (denoting a complex LXR α :RXR α), PPAR-a (denoting a complex PPAR α :RXR α) and SREBP-a (denoting the cleaved form of SREBP-1c). We also consider SCAP (SREBP cleavage activating protein), a key enzyme involved in the cleavage of SREBP-1c, that interacts with another family of proteins called INSIG (showing the complexity of molecular mechanism).

We also include in the model 'final' products, that is enzymes ACL, ACC, FAS, SCD1 (implied in the fatty acid synthesis from citrate), D5D, D6D (implied in PUFA synthesis) as well as PUFA themselves.

3.2 Interactions in the Model

Relations between the variables are the following. SREBP-a is an activator of the transcription of ACL, ACC, FAS, SCD1, D5D and D6D [11, 12]. LXR-a is a direct activator of the transcription of SREBP and FAS, it also indirectly activates ACL, ACC and SCD1 [13]. Notice that these indirect actions are kept in the model because we do not know whether they are only SREBP-mediated.

PUFA activates the formation of PPAR-a from PPAR, and inhibits the formation of LXR-a from LXR as well as the formation of SREBP-a (by inducing the degradation of mRNA and inhibiting the cleavage) [11]. SCAP represents the activators of the formation of SREBP-a from SREBP, and is inhibited by PUFA.

PPAR directly activates the production of SCD1, D5D, D6D [14–16]. The dual regulation of SCD1, D5D and D6D by SREBP and PPAR is paradoxical because SREBP transactivates genes for fatty acid synthesis in liver, while PPAR induces enzymes for fatty acid oxidation. Hence, the induction of the D5D and D6D gene by PPAR appears to be a compensatory response to the increased PUFA demand caused by induction of fatty acid oxidation.

3.3 Fasting-Refeeding Protocols

The fasting-refeeding protocols represent a favourable condition for studying lipogenesis regulation; we suppose that during an experiment, animals (such as rodents or chicken) were kept in a fasted state for several hours. Then, hepatic mRNA of LXR, SREBP, PPAR, ACL, FAS, ACC and SCD1 are quantified by DNA microarray analysis. Biochemical measures also provide the variation of PUFA.

A compilation of recent literature on lipogenesis regulation provides hypothetical results of such protocols: SREBP, ACL, ACC, FAS and SCD1 decline in the liver during the fasted state [17]. This is expected because fasting results in an inhibition of fatty acid synthesis and an activation of the fatty acid oxidation. For the same reason, PPAR is increased in order to trigger oxidation. However, Tobin et al. [18] showed

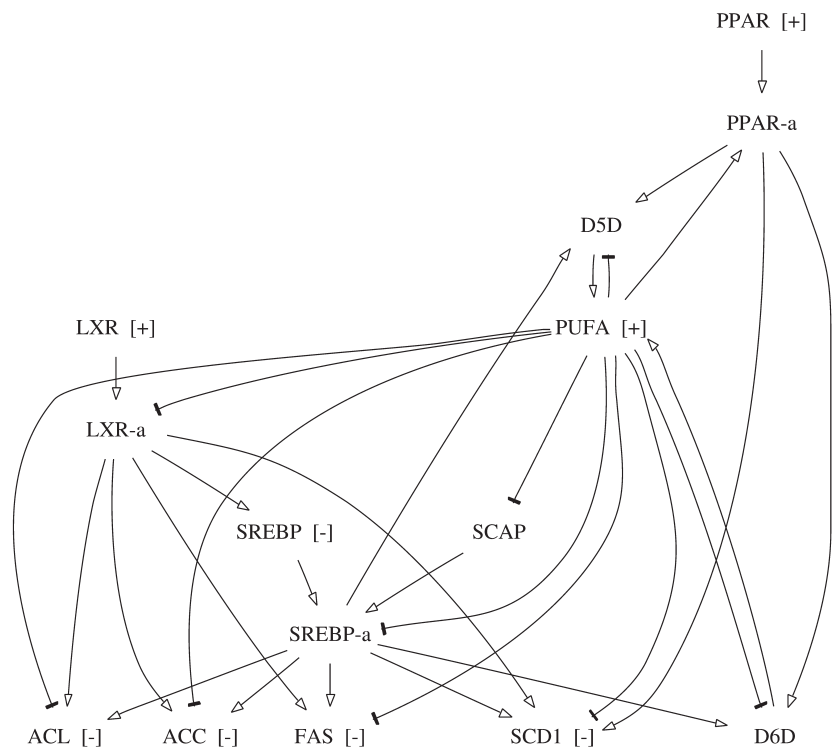


Fig. 1. Interaction graph for the toy model. Self-regulation loops on nodes are omitted for the sake of clarity. Observed variations are depicted next to each vertex, when available.

that rats fasting for 24 h increased the hepatic LXR mRNA, although LXR positively regulates fatty acid synthesis in its activated form. Finally, PUFA levels can be considered to be increased in liver following starvation because of the important lipolysis from adipose tissue as shown by Lee et al. [19] in mice after 72 h fasting.

3.4 Qualitative System Derived from the Graph

As explained in the previous section, we derive a qualitative system from the interaction graph shown in figure 1. For ease of presentation, we denote by A the sign of variation for species A . (See table below).

In the next section, we propose an efficient representation for such qualitative systems.

4 Analysis of Qualitative Equations: a New Approach

4.1 Resolution of Qualitative Systems

The resolution of (even linear) qualitative systems is an NP-complete problem [see for instance 20, 21]. One can show this by reducing the satisfiability problem for a finite set of clauses to the resolution of a qualitative system in polynomial time.

Let us consider a collection $C = \{c_1, \dots, c_n\}$ of clauses on a finite set V of variables. Let $\{+, -, ?\}$ a sign qualitative algebra. In order to reduce the satisfiability problem to the resolution of a qualitative

system, let us code *true* into $+$ and *false* into $-$. If c is a clause, let us denote by \bar{c} the encoding of c in a qualitative algebra formula. The following encoding scheme provides a polynomial procedure to code a clause into a qualitative formula:

Clause	Sign algebra
$a \in V$	$\rightarrow \bar{a}$
$c_1 \vee c_2$	$\rightarrow \bar{c}_1 + \bar{c}_2$
$\neg c$	$\rightarrow -\bar{c}$

The satisfiability problem for the set of clauses C is then reduced to finding a solution of the qualitative system:

$$\{\bar{c}_i \approx + / i = 1, \dots, n\}$$

So an NP-complete problem can be reduced to the resolution of a qualitative system in polynomial time (with respect to the size of the problem). This shows that solving qualitative systems is an NP-complete problem. For example, the only pair of values which are not solution of $-\bar{a} + \bar{b} \approx +$ are $(+, -)$. This corresponds to the only pair (*true*, *false*) that does not satisfy $\neg a \vee b$.

Several heuristics were proposed for the resolution of qualitative systems. For linear systems, sets of rules have been designed [20]. This set is complete: it allows to find every solution. It is also sound: every solution found by applying these rules is correct. The rules are based on an adaptation of Gaussian elimination. However, only heuristics exist for choosing the equa-

tion and the rule to apply to it. In case of a dead end, when no rule applies any longer, it is necessary to backtrack to the last decision made. As a result programmes implementing qualitative resolution are not very efficient in general and only minor problems can be resolved in reasonable time. For this reason we propose an alternate way to solve qualitative systems (linear or not).

4.2 Qualitative Equation Coding

Our method is based on a coding of qualitative equations as algebraic equations over Galois fields $\mathbb{Z}/p\mathbb{Z}$ where p is a prime number greater than 2. The elements of these fields are the classes modulo p of the integers. If \bar{x} denotes the class of the integer x modulo p , a sum and a product are defined on $\mathbb{Z}/p\mathbb{Z}$ as follows:

$$\bar{x} + \bar{y} = \overline{x+y} \quad \bar{x} \times \bar{y} = \overline{x \times y}$$

Galois fields have two basic properties which we use extensively:

- Every function $f : (\mathbb{Z}/p\mathbb{Z})^n \rightarrow \mathbb{Z}/p\mathbb{Z}$ with n arguments $\mathbb{Z}/p\mathbb{Z}$ is a polynomial function
- if \oplus denotes the operation $f \oplus g = f^{(p-1)} + g^{(p-1)}$, then every equation system $p_1(X) = 0, \dots, p_k(X) = 0$ has the same solutions than the unique equation $p_1 \oplus p_2 \oplus \dots \oplus p_k(X) = 0$.

The following table specifies how the sign algebra $\{+, -, ?\}$ is mapped onto the Galois field with three elements $\mathbb{Z}/3\mathbb{Z}$ used for that coding.

Sign algebra	$\mathbb{Z}/3\mathbb{Z}$
$+$	$\rightarrow 1$
$-$	$\rightarrow -1$
$?$	$\rightarrow 0$

Sign algebra	$\mathbb{Z}/3\mathbb{Z}$
$e_1 + e_2$	$\rightarrow -\bar{e}_1 \cdot \bar{e}_2 \cdot (\bar{e}_1 + \bar{e}_2)$
$e_1 \times e_2$	$\rightarrow \bar{e}_1 \cdot \bar{e}_2$
$e_1 \approx e_2$	$\rightarrow \bar{e}_1 \cdot \bar{e}_2 \cdot (\bar{e}_1 - \bar{e}_2)$

Finally a qualitative system $\{e_1, \dots, e_n\}$ is coded as the polynomial $\bar{e}_1 \oplus \dots \oplus \bar{e}_n$.

System 1	Observations 1
1 PPAR-a = PPAR + PUFA	PPAR = +
2 LXR-a = -PUFA + LXR	PUFA = +
3 SREBP = LXR-a	LXR = +
4 SREBP-a = SREBP + SCAP - PUFA	SREBP = -
5 ACL = LXR-a + SREBP-a - PUFA	ACL = -
6 ACC = LXR-a + SREBP-a - PUFA	ACC = -
7 FAS = LXR-a + SREBP-a - PUFA	FAS = -
8 SCD1 = LXR-a + SREBP-a - PUFA + PPAR-a	SCD1 = -
9 SCAP = -PUFA	
10 D5D = PPAR-a + SREBP-a - PUFA	
11 D6D = PPAR-a + SREBP-a - PUFA	

A similar coding for the qualitative algebra $\{+, -, \mathbf{0}, \mathbf{?}\}$ uses the Galois field $\mathbb{Z}/5\mathbb{Z}$ and will not be presented here.

With this coding, every qualitative system has a solution if and only if the corresponding polynomial has a solution without null component. Null solutions are excluded since $\mathbf{?}$ solutions are excluded for qualitative systems. In general we will have to add polynomial equations $X^2 = 1$ to insure this.

4.3 An Efficient Representation of Polynomial Functions

Recall that our purpose is to efficiently solve an NP-complete problem. There is no hope to find a representation of polynomial functions allowing to solve polynomial systems of equations in polynomial time. The coding of a qualitative system as a polynomial equation is obviously polynomial in the size of the system (number of variables plus number of equations). So finding the solution of a polynomial system of equations is itself an NP-complete problem. It is more or less the satisfiability problem.

Nevertheless, there exists a representation of polynomial functions on Galois fields which gives, in practice, good performances for polynomials with hundreds of variables. This kind of representation was first used for logical functions which may be considered as polynomial functions over the field $\mathbb{Z}/2\mathbb{Z}$. This representation is known as BDD (binary decision diagrams) and is widely used in checking logical circuits [22] and in model checkers as nuSMV [23].

We present here this representation for the field $\mathbb{Z}/3\mathbb{Z}$. Generalizations to other Galois fields could be treated as well. The starting point is a generalization of Shannon decomposition for logical functions:

$$\begin{aligned} p(X_1, X) &= (1 - X_1^2)p_{[X_1=0]}(X) \\ &+ X_1(-X_1 - X_1^2)p_{[X_1=1]}(X) \\ &+ X_1(X_1 - X_1^2)p_{[X_1=2]}(X) \end{aligned}$$

where p is a polynomial function with n variables. This decomposition leads to a tree representation of the polynomial function: the variable X_1 is the root and has three children. Each of these is obtained by instantiating X_1 to $-1, 0$ or 1 in $p(X_1, X)$. This representation is exponential (3^n) as each non-constant node has 3 children. It also depends on a chosen order on the variables.

Then a key observation [see 22] is that several subtrees are identical. They have the same variable as the root variable and isomorphic children. If we decide to represent each type of tree only once, then the tree representation is transformed into a direct acyclic graph. With this representation there is no more redundancy among subtrees. The result may be a dramatic decrease in the size of the representation of a polynomial function (fig. 2).

A property of the Shannon-like decomposition is that many operations on polynomial functions are recursive with respect to this decomposition. More precisely let

$$\begin{aligned} p^i(X_1, X) &= (1 - X_1^2)p_0^i(X) \\ &+ X_1(-X_1 - X_1^2)p_1^i(X) \\ &+ X_1(X_1 - X_1^2)p_2^i(X) \end{aligned}$$

$i = 1, 2$ be two polynomial functions with $p_\alpha(X) = p_{[X_1=\alpha]}(X)$, $\alpha = 0, 1, 2$. Then for binary operations Δ on polynomial functions,

$$\begin{aligned} p^1 \Delta p^2 &= (1 - X_1^2)(p_0^1 \Delta p_0^2) \\ &+ X_1(-X_1 - X_1^2)(p_1^1 \Delta p_1^2) \\ &+ X_1(X_1 - X_1^2)(p_2^1 \Delta p_2^2) \end{aligned}$$

This kind of recursive formula leads to an exponential complexity of any computation. Again, it is possible to take advantage of the redundancy by using a cache to remember each operation. This technique is known as memoization in formal calculus. A 40% cache hit rate is commonly observed.

More complex operations on polynomial functions are also implemented with a recursive scheme and memoization. Let us just mention quantifier elimination as among the most useful for our purpose.

This representation of polynomial functions on Galois fields has also several drawbacks:

- The memory size heavily depends on the order of variables. The libraries implementing formal computations always have to reordering algorithms.
- For each order, there exist polynomial functions which are exponential in memory size.

Nevertheless, in practice, this representation has proved to be very efficient for polynomial functions with several hundred variables. The computations performed on our toy model and on another real size one used a programme named SI-GALI which is devoted to polynomial functions on $\mathbb{Z}/3\mathbb{Z}$ representation. Several algorithms were added to this programme in order to answer questions of biological interest.

5 Qualitative Models and Experimental Data

In this section, we show how to compute some properties of a qualitative system, and eventually get some insights into the biological model it represents. The algorithms we derive heavily rely on the representation introduced above. Hence, not only can they deal in practice with computationally hard problems efficiently, but also they are expressed in a rather simple and generic fashion.

Let M be a qualitative model represented by its associated interaction graph $G(V, E)$. Recall that V is the set of variables. Let V_O be the set of observed variables, and $o_i \in \{+, -\}$ for $i \in V_O$ the experimental observations. As explained in the previous section, the qualitative system derived from M can be coded as a polynomial function $P_M(X_1, \dots, X_n)$. Roots of P_M corre-

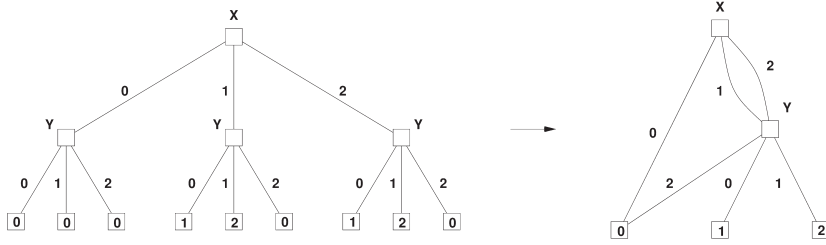


Fig. 2. From tree representation to direct acyclic graph for $X^2(Y + 1)$. The tree has 13 nodes while the DAG representing the same function has 5 nodes.

spond to solutions of the qualitative system.

5.1 Satisfiability of the Qualitative System

A property of the coding described above is that the system has no solution if P_M is equal to the constant polynomial 1. Alternatively if $P_M = 0$, the qualitative equations do not constrain the variables at all.

Now if some observations o_i for $i \in V_O$ are available, checking their consistency with the model M boils down to instantiating $X_i = o_i$ in $P_M(X_1, \dots, X_n)$, for all $i \in V_O$, and testing whether the resulting polynomial is different from 1.

We computed the polynomial P_L associated with our toy example (see section 3) and it has roots. Recall that it does not guarantee the existence of some (quantitative) differential model conforming to the interaction graph depicted in figure 1. Satisfiability of the qualitative system is only a necessary condition for the model to be correct.

The polynomial obtained by instantiating these observations into P_L is different from 1, meaning that our model does not contradict generally observed variations during fasting.

Large size models might advantageously be reduced using standard graph techniques. First we look for connected components in the interaction graph. A graph with several connected components represents a coherent qualitative model if each

component is coherent. Second, a node without successor except itself appears only in its associated equation. If this node is not observed, its associated qualitative equation adds no constraint on the other nodes. So, at least for satisfiability checking, this node can be suppressed and its qualitative equation removed from the system. This procedure is applied iteratively, until no node can be deleted. The resulting graph leads to a new qualitative system which is satisfiable if the initial system is satisfiable.

5.2 Correcting Data or Model

If the qualitative system, given some experimental observations, is found to have no solution, it is of interest to propose some

correction of the data and/or the model. By correction, we mean inverting the sign of an observed variable or the sign of an edge of the interaction graph. In the general case, there are several possibilities to make the system satisfiable, and we need some criterion to choose among them. We applied a parsimony principle: a correction of the data should imply a minimal number of sign inversions.

In the following, we show how to compute all minimal corrections for the data. Given $(o_i)_{i \in V_O}$ a vector of experimental observations which is not compatible with the model, we compute all $(o'_i)_{i \in V_O}$ vectors which are compatible with the data and such that the Hamming distance between o and o' is minimal. By Hamming distance, we mean the number of differences between o and o' . The set of such o' vectors might be very large, but again, by encoding it as the set of roots of a polynomial function, we obtain a compact representation.

This procedure can be extended in a straightforward manner to corrections of edge signs in the interaction graph. This is done by considering these signs as variables of the model. For ease of presentation, we only detail data correction.

Input:

P , a polynomial function on variables V
 $i \in V$

Output:

C , a polynomial function encoding all minimal corrections
 d , minimal number of corrections

if P is constant then

if $P = 0$ then

Result: $C = 0, d = 0$

else

Result: $C = 1, d = \infty$

end

else

let P_0, P_1, P_2 be the Shannon decomposition of P with respect to variable X_i ,
and (C_j, d_j) the result obtained by recursively applying the algorithm on P_j and $i + 1$ for $j \in \{0, 1, 2\}$

let $d'_j = \begin{cases} d_j + 1 & \text{if } i \in V_O \text{ and } o_i \neq j \\ d_j & \text{otherwise} \end{cases}$ and $C'_j = \begin{cases} (X_i - j) \oplus C_j & \text{if } i \in V_O \\ C_j & \text{otherwise} \end{cases}$

Result: $d = \min d'_j, C = \prod_{j, d'_j = d} C'_j$

end

Algorithm 1: Algorithm for experimental data correction.

Let us illustrate this algorithm on our toy example: during fasting experiments, synthesis of fatty acids tends to be inhibited, while oxidation, which produces ATP, is activated. In particular ACC, ACL, FAS and SCD1 are implied in the same pathway to produce saturated and monounsaturated fatty acids. Expectedly, they are known to decline together at fasting. Suppose we introduce some wrong observation, say for instance an increase of ACL, while keeping all other observations given above. The polynomial obtained from P_L including these new observations is equal to 1, and hence has no solution. Applying algorithm 1, we recover this error. Now if we wrongly change two values, say ACL and ACC to 1, the algorithm proposes a different correction, namely to change the observed value of SREBP to 1, which is more parsimonious.

5.3 Experiment Design

It is often the case that not all variables in the system under study can be observed. Biochemical measurements of metabolites can be costly and/or time consuming. By experiment design, we mean here the choice of the variables to observe so that an experiment might be informative.

Let $P_M(X_O, X_U)$ be the polynomial function coding for the qualitative system M . X_O (or X_U) denotes the state vector of observed (or unobserved) variables. The polynomial function representing the admissible values of the observed variables is obtained by elimination of the quantifier in $\exists X_U P_M(X_O, X_U)$. Let $P_M^O(X_O)$ denote the resulting polynomial function.

For some choice of observed variables, it might well be that P_M^O is null, which basically means that the experiment is totally useless. Remark that no improvement can be found by taking a subset of X_O . The solution is either to add new observed variables or to choose a completely different set of observed variables.

In order to assess the relevance of a given experiment (namely of a given observed

subset), we suggest to compute the following ratio: number of consistent valuations for observed variables versus the total number of valuations of observed variables. A very stringent experiment has a low ratio. An experiment having a ratio value of one is useless.

Again this computation is carried out in a recursive fashion. Let P be a polynomial function representing the set of admissible observed values. Let $Rat(p)$ the percentage of solutions of $P(X) = 0$ in the space $(\mathbb{Z}/p\mathbb{Z})^n$, where n is the number of variables X . If P is constant then $Rat(P) = 1$ (or $Rat(P) = 0$) if $P = 0$ (or $P \neq 0$). Else, let P_1, P_2, P_3 be a Shannon-like decomposition of $P(X)$ with respect to some variable of P . Then it is easy to prove:

$$Rat(P) = (Rat(P_0) + Rat(P_1) + Rat(P_2))/3$$

The relevance of this approach was assessed on our toy example: for each subset O of variables in the model, containing at most four variables, we computed $Rat(P_L^O)$. Expectedly, the lowest ratios (i.e. the most stringent experiments) were achieved observing four variables: either {SCAP, PUFA, PPAR-a, PPAR}, or {SREBP, SCAP, PUFA, LXR-a}, or {SREBP, PPAR-a, PPAR, LXR-a}.

Interestingly, the procedure captures what might be thought of as control variables, like PUFA/SCAP, SREBP/LXR-a and PPAR/PPAR-a. The first two pairs control the activation of fatty acid synthesis; the third one controls fatty acid oxidation.

Indeed one can go even further: if we isolate some kind of control variables, we are naturally interested in knowing how they constrain other variables. Achieving this amounts to computing the set of variables the value of which is constant for all solutions of the system (the so-called hard components). Recall that these hard components of qualitative solutions are also important with respect to the hypothetical differential model which is abstracted in the qualitative one. Indeed, all solutions of the quantitative equation for equilibrium change have the same sign pattern on the hard components. Algorithm 2 describes a recursive procedure which finds the set of hard components together with their value.

Let us set some of our previously found control variables of the toy example to a given value, say PUFA to 1, and LXR to -1. Then applying the algorithm 2, the corresponding polynomial has the following hard components:

Input: P , a polynomial function on variables V

Output:

the set $W \subset V \times \{0, 1, 2\}$ of hard components, together with their values
a boolean b which is true if P has at least one root

if P is constant **then**

if $P = 0$ **then**

return $(\emptyset, true)$

else

return $(\emptyset, false)$

end

else

let P_0, P_1, P_2 be the Shannon decomposition of P with respect to variable X_i ,
and (W_j, b_j) the result obtained by recursively applying the algorithm on P_j for $j \in \{0, 1, 2\}$

let $W = \{(v, v') | v \in V, v' \in \{0, 1, 2\}, \forall j b_j \Rightarrow (v, v') \in W_j\}$

if there exists a unique j_0 such that b_{j_0} is true **then**

add (i, j_0) to W

end

return $(W, b_0 \vee b_1 \vee b_2)$

end

Algorithm 2: Determination of hard components

ACL	= -1	FAS	= -1
ACC	= -1	LXR-a	= -1
SCAP	= -1	SREBP	= -1
SREBP-a	= -1	PPAR	= -1
PPAR-a	= -1		

which expectedly corresponds to the inhibition of fatty acid synthesis.

5.4 Real Size System

We have used our new technique to check the consistency of a database of molecular interactions involved in the genetic regulation of fatty acid synthesis. In the database, interactions were classified as behavioural or biochemical.

- A behavioural interaction describes the effects of a variation of a product concentration. It is either direct or indirect (unknown mechanism).
- A biochemical interaction may be a gene transcription, a reaction catalyzed by an enzyme. Such molecular interactions can be found in existing databases. They need a behavioural interpretation.

All the behavioural interactions were manually extracted from a selection of scientific papers. Biochemical interactions were extracted from public databases available on the Web (Bind [24], IntAct [25], Amaze [26], KEGG [27] or TransPath [28]). A biochemical interaction may be linked to a behavioural interpretation in the database.

The database is used to generate the interaction graph. While behavioural interactions directly correspond to edges in the graph, biochemical interactions are given a simplified interpretation. Roughly, any increase of a reaction input induces an increase of the outputs.

The interaction graph which is built from the database contains more than 600 vertices and more than 1,400 edges. It is clear that, nevertheless, the obtained graph is not a comprehensive model of genetic regulation of fatty acid synthesis in the liver. Anyway our aim is to see how far this model can account for experimental ob-

servations, and propose some corrections when it cannot.

We used our technique to check the coherence of the whole model. After reducing the graph with standard graph techniques as described in section 5.1, we found that the model was incoherent. The reduced graph has about 150 nodes. We developed a heuristic to isolate minimal incoherent subsystems. It turned out that all the contradictions we detected resulted from arguable interpretations of the literature.

6 Conclusion

In this paper we proposed a qualitative approach for the analysis of large biological systems. We rely on a framework more thoroughly described by Siegel et al. [6], which is meant to model the comparison between two experimental conditions as a steady-state shift. This approach fits well with state of the art biological measurement techniques, which provide rather noisy data for a large amount of targets. It is also well suited to the use of biological knowledge, which is most of the time descriptive and qualitative.

This qualitative approach is all the more attractive since we can rely on new analysis methods for qualitative systems. This new technique is also introduced in this paper and is original in qualitative modelling. It relies on a representation of qualitative constraints by decision diagrams. This has not only a major impact on the scalability of qualitative reasoning, but it also permits to derive many algorithms in a quite generic fashion.

We plan to validate our approach on pathways which have been published for yeast and *E. coli*. These pathways are not only of a significant size but microarray data for this species are publicly available. Concerning the scalability of the methods, qualitative systems with up to 200 variables are handled within a few minutes.

On the theoretical side, we study applications of our algebraic techniques to network reconstruction, as proposed by

Wagner [29]. The problem is to infer direct actions between products, based on large-scale perturbation data, in order to obtain the most parsimonious interaction graph. Our approach could lead to a reformulation of this problem in terms of polynomial operations. Indeed, finding a minimal regulation network from a minimal polynomial representation has already been described by Laubenbacher and Stigler [30], though it was applied to a rather different type of network. A similar approach tailored to the framework described in this paper could eventually lead to original and practical algorithms for network reconstruction.

Acknowledgement

This research was supported by ACI IMPBio, a French Ministry for Research programme on interdisciplinarity.

References

- 1 de Jong H: Modeling and simulation of genetic regulatory systems: a literature review. *J Comput Biol* 2002; 9: 69–105.
- 2 de Jong H, Gouzé J-L, Hernandez C, Page M, Sari T, Geiselmann J: Qualitative simulation of genetic regulatory networks using piecewise-linear models. *Bull Math Biol* 2004; 66: 301–340.
- 3 Ghosh R, Tomlin C: Symbolic reachable set computation of piecewise affine hybrid automata and its application to biological modelling: delta-notch protein signalling. *Syst Biol* 2004; 1: 170–183.
- 4 Chaouiya C, Remy E, Ruet P, Thieffry D: Qualitative modelling of genetic networks: from logical regulatory graphs to standard petri nets. *Lecture Notes in Computer Science*. Berlin, Springer-Verlag, 2004, vol 3099, pp 137–156.
- 5 Chabrier-Rivier N, Chiaverini M, Danos V, Pages F, Schächter V: Modeling and querying biomolecular interaction networks. *Theor Comput Sci* 2004; 325: 25–44.
- 6 Siegel A, Radulescu O, Le Borgne M, Veber P, Ouy J, Lagarrigue S: Qualitative analysis of the relation between DNA microarray data and behavioral models of regulation networks. *Biosystems*, submitted.
- 7 Gutierrez-Rios RM, Rosenblueth DA, Loza JA, Huerta AM, Glasner JD, Blattner FR, Collado-Vides J: Regulatory network of *Escherichia coli*: consistency between literature knowledge and microarray profiles. *Genome Res* 2003; 13: 2435–2443.
- 8 Fell D: *Understanding the Control of Metabolism*. London, Portland Press, 1997.
- 9 Heinrich R, Schuster S: *The Regulation of Cellular Systems*. New York, Chapman & Hall, 1996.
- 10 Radulescu O, Lagarrigue S, Siegel A, Le Borgne M, Veber P: Topology and linear response of interaction networks in molecular biology. *R Soc Interface*, submitted.
- 11 Jump DB: Fatty acid regulation of gene transcription. *Crit Rev Clin Lab Sci* 2004; 41: 41–78.
- 12 Nara TY, He WS, Tang C, Clarke SD, Nakamura MT: The e-box like sterol regulatory element mediates the suppression of human delta-6 desaturase gene by highly unsaturated fatty acids. *Biochem Biophys Res Commun* 2002; 296: 111–117.
- 13 Steffensen KR, Gustafsson JA: Putative metabolic effects of the liver x receptor (lrx). *Diabetes* 2004; 53(supp 1):36–52.
- 14 Matsuzaka T, Shimano H, Yahagi N, et al: Dual regulation of mouse delta(5)- and delta(6)-desaturase gene expression by SREBP-1 and PPARalpha. *J Lipid Res* 2002; 43: 107–114.
- 15 Miller CW, Ntambi JM: Peroxisome proliferators induce mouse liver stearoyl-CoA desaturase 1 gene expression. *Proc Natl Acad Sci USA* 1996; 93: 9443–9448.
- 16 Tang C, Cho HP, Nakamura MT, Clarke SD: Regulation of human delta-6 desaturase gene transcription: identification of a functional direct repeat-1 element. *J Lipid Res* 2003; 44: 686–695.
- 17 Liang G, Yang J, Horton JD, Hammer RE, et al: Diminished hepatic response to fasting/refeeding and liver x receptor agonists in mice with selective deficiency of sterol regulatory element-binding protein-1c. *J Biol Chem* 2002; 277: 9520–9528.
- 18 Tobin KA, Steineger HH, Alberti S, Spydevold O, et al: Cross-talk between fatty acid and cholesterol metabolism mediated by liver x receptor-alpha. *Mol Endocrinol* 2000; 14: 741–752.
- 19 Lee SS, Chan WY, Lo CK, et al: Requirement of PPARalpha in maintaining phospholipid and triacylglycerol homeostasis during energy deprivation. *J Lipid Res* 2004; 45: 2025–2037.
- 20 Dormoy JL: Controlling qualitative resolution. *Proceedings of the 7th National Conference on Artificial Intelligence, AAAI88*, Saint Paul, 1988.
- 21 Travé-Massuyès L, Dague P: *Modèles et raisonnements qualitatifs*. Paris, Hermès Sciences, 2003.
- 22 Bryan RE: Graph-based algorithm for boolean function manipulation. *IEEE Trans Comput* 1986; 8: 677–691.
- 23 Clarke E, Grumberg O, Long D: Verification tools for finite-state concurrent systems; in *A Decade of Concurrency – Reflections and Perspectives*. *Lecture Notes in Computer Science*. Berlin, Springer-Verlag, 1994, vol 803, pp 124–128.
- 24 Bader GD, Betel D, Hogue CW: Bind: the biomolecular interaction network database. *Nucleic Acids Res* 2003; 31: 248–250.
- 25 Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, et al: IntAct – an open source molecular interaction database. *Nucleic Acids Res* 2004; 32: D452–D455.
- 26 Lemer C, Antezana E, Couche F, et al: The aMAZE LightBench: a web interface to a relational database of cellular processes. *Nucleic Acids Res* 2004; 32:D443–D448.
- 27 Ogata H, Goto S, Sato K, Fujibuchi W, et al: Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 1999; 27: 29–34.
- 28 Schacherer F, Choi C, Gotze U, Krull M, Pistor S, Wingender E: The TRANSPATH signal transduction database: a knowledge base on signal transduction networks. *Bioinformatics* 2001; 17: 1053–1057.
- 29 Wagner A: Reconstructing pathways in large genetic networks from genetic perturbations. *J Comput Biol* 2004; 11: 53–60.
- 30 Laubenbacher R, Stigler B: A computational algebra approach to the reverse engineering of gene regulatory networks. *J Theor Biol* 2004; 229: 523–537.