

RESEARCH ARTICLE

Open Access

Applications of a formal approach to decipher discrete genetic networks

Fabien Corblin^{1,2*}, Eric Fanchon^{1*}, Laurent Trilling¹

Abstract

Background: A growing demand for tools to assist the building and analysis of biological networks exists in systems biology. We argue that the use of a formal approach is relevant and applicable to address questions raised by biologists about such networks. The behaviour of these systems being complex, it is essential to exploit efficiently every bit of experimental information. In our approach, both the evolution rules and the partial knowledge about the structure and the behaviour of the network are formalized using a common constraint-based language.

Results: In this article our formal and declarative approach is applied to three biological applications. The software environment that we developed allows to specifically address each application through a new class of biologically relevant queries. We show that we can describe easily and in a formal manner the partial knowledge about a genetic network. Moreover we show that this environment, based on a constraint algorithmic approach, offers a wide variety of functionalities, going beyond simple simulations, such as proof of consistency, model revision, prediction of properties, search for minimal models relatively to specified criteria.

Conclusions: The formal approach proposed here deeply changes the way to proceed in the exploration of genetic and biochemical networks, first by avoiding the usual trial-and-error procedure, and second by placing the emphasis on sets of solutions, rather than a single solution arbitrarily chosen among many others. Last, the constraint approach promotes an integration of model and experimental data in a single framework.

Background

A central task in molecular systems biology is to build and analyze genetic and biochemical networks in order to decipher the properties of cellular phenomena. The emphasis is not on investigating in detail one or a few molecules at a time, as is done traditionally in molecular biology, but rather on focusing on the network level.

We are specifically interested here in gene regulatory networks (GRNs) formalized as discrete genetic networks as defined by R. Thomas [1,2]. The main goal of this formalism is to obtain a qualitative understanding of the network dynamics by reasoning on discrete entities. In GRNs the molecular players are the genes and the proteins they produce. A genetic interaction corresponds to the fact that a gene g_i produces a protein p_i which influences the expression rate of another gene g_j , or g_i itself. The set of all these genetic interactions

constitute the interaction graph, to be defined formally later. In a given state of the network, each gene has a certain expression rate (the rate of production of the encoded protein) which depends on the presence or absence of a subset of proteins, the activators or repressors of the considered gene. The expression rate of a gene is maximum when all its activators are present and all its repressors are absent. In this context a basic objective is to analyze the temporal evolution of the protein concentrations in given external conditions. This can be done when the values of the model parameters have been measured. When this is not the case, the problem is then to exploit the knowledge about network behaviour (e.g. response to perturbations, phenotype change when one or several genes are knocked-out) to deduce the possible values of the parameters. A frequently used method consists in performing a large number of simulations by varying some parameters (generally one or two at a time) and selecting a posteriori the set of values that is consistent with the observed

* Correspondence: fabien.corblin@irisa.fr; eric.fanchon@imag.fr

¹Laboratoire TIMC-IMAG, UMR CNRS/UJF 5525, Domaine de la Merci, 38710 La Tronche, France

behaviour. As explained below, we propose a different approach for this inference problem.

Beyond these basic functionalities (simulation, inference of parameter values), the construction of GRN models consistent with experimental data requires more sophisticated tools. It often occurs that a proposed model displays inconsistencies with part of the data. In such cases it is necessary to critically analyze the hypotheses used in building the model and to revise them. This analysis can be done “by hand” for small networks, e.g. up to three genes, but requires the use of computational tools to cope with the complexity of larger networks. In still other situations the observational constraints are weak with respect to the number of variables, and the number of solutions is very large. In such cases, it is interesting to derive properties that are shared by all the solutions, or subsets of them, in order to get a better understanding of the model properties and to design new experiments having the potential to substantially reduce the set of solutions.

Fundamentally, we want to provide the biologist studying GRNs with a software environment allowing to perform such tasks. The available knowledge is partial and bears on both the *structure* of the network of interest (the set of interactions) and the *behaviour* of the network in various conditions. The first kind of knowledge is said to be structural, or local (each interaction is a piece of information and can be studied in itself), whereas the second kind is said to be behavioural, or global (the network as a whole is giving rise to a given behaviour). The network architecture and its behaviour are closely inter-related. This relation is implemented formally as a set of constraints in a straightforward manner in our software environment, named GNBox (Genetic Networks toolBox - Additional file 1). More precisely, the philosophy of this approach is to represent a given problem, or set of problems, as a set of formulae linking variables. In our case this entails the specification of (i) the rules defining the updating scheme (how the successors of a state are computed); (ii) the network architecture (set of interactions); (iii) the observations about the behaviours of the network (partial information about paths), or any working hypothesis about the system; (iv) the query itself (e.g. number of stationary states, possible values of initially unknown parameters). The set of constraints thus defined is then submitted to a solver whether there exists solutions or not. A distinctive feature of the constraint approach is constraint propagation. It implements deduction rules and allows in favorable conditions to reduce drastically the search space, thus limiting enumeration. Of course some amount of enumeration is usually still necessary, but the aim of the game is to reduce it as much as possible. This formal relation is “executable” and allows not only

to perform basic functionalities such as simulation or reverse-engineering, but also to assert and obtain properties on both the behaviours and the interactions. More specifically, we implemented in this constraint-based setting four main functionalities: (i) proof of consistency or inconsistency of a constraint pool, (ii) constraint relaxation in case of inconsistency (model revision), (iii) prediction of properties in case of consistency, (iv) search for minimal models, with respect to the number of thresholds, for example.

In this article we present our approach and we show how it can be applied successfully to the analysis of three different biological problems. In the section Methods we present the formalism we developed. We present the formal definition of interaction graphs and of the evolution rules of Thomas networks. These notions are required to express the queries implementing the functionalities mentioned above. Other notions related to the specification of interaction compositions facilitate the expression of properties involving kinetic parameters. The implementation is discussed in [3]. In the section Results and Discussion we present three applications which differ by both the type of knowledge available and the type of biological questions asked. These applications permit (i) to illustrate the different functionalities of GNBox, (ii) to show the feasibility of this constraint-based approach on realistic biological problems, and (iii) to support the idea that a formal and declarative approach is very interesting to decipher the properties of GRNs, in order to assist in their construction.

Methods

We present briefly in this section the constraint technology, the constraint-based formalization of Thomas networks, the constraint-based formalization of biological properties of these networks, and the features of our software environment GNBox to elucidate GRNs.

Below we use the following mathematical notations: an integer x taking values between min and max is denoted $x \in min..max$, a Boolean b is an integer such as $b \in 0..1$, $b1 \Leftrightarrow b2$ means that the Boolean $b1$ is equal (or equivalent) to the Boolean $b2$, $b1$ implies $b2$ is denoted $b1 \Rightarrow b2$, the Boolean equal to $b1$ and $b2$ is denoted $b1 \wedge b2$, the Boolean equal to $b1$ or $b2$ is denoted $b1 \vee b2$, the Boolean equal to the conjunction of a list of Boolean b_i is denoted $\bigwedge_i b_i$, the Boolean equal to the disjunction of a list of Boolean b_i is denoted $\bigvee_i b_i$.

Constraint technology

We propose to implement the approach (particularly the link between network structure and behaviour) using Constraint Logic Programming (CLP) technology, with a

finite domain solver. CLP is a programming paradigm based on first order logic. CLP considers specific classes of logical terms and proposes efficient resolution methods of equations over these terms (constraints). A CLP program is a logic formula, and its execution is the construction of a proof of consistency (or inconsistency) of this formula. The formula is consistent when it is possible to find an instantiation of the variables which satisfies the formula. Logicians call such an instantiation a model. A CLP program is reversible in the sense that it permits to impose and obtain partial knowledge over all the variables of the formula (including in our case the variables describing the interactions and behaviours). For example, let say that $p(x, y)$ is a predicate defining a relationship between two entities x and y . If a measurement allows to reduce the domain of values of x , this information can be added as an additional constraint, and a *query* can be submitted about the possible values of y . The solver will try to propagate the additional information on x to reduce the domain of y , taking into account $p(x, y)$. Conversely, if the measurement has been performed on y , this information can be propagated to x through $p(x, y)$. This is reversibility. It must be said that different kinds of solvers exist, characterized by the type of variables (e.g. finite domain integers, reals) and the type of propagation rules used, among other things.

As all the variables describing interactions and behaviours have finite integer domains in the discrete framework used, the use of a constraint solver over finite domains is very well suited. In addition, the expressive power of first order logic and constraints over integers allows the definition of very general properties and functionalities. Finally, in order to be able to take advantage of the very efficient Boolean Satisfiability (SAT) solvers available, the GNBox environment is able to translate the CLP formalization into a Boolean formula in Conjunctive Normal Form (conjunction of disjunctions of Boolean variables or their negation, the input format used by most SAT solvers). Details on the translation into CNF can be found in [3]. In this way we combine the expressive power of CLP with the efficiency of SAT solvers.

Formalization of Thomas networks

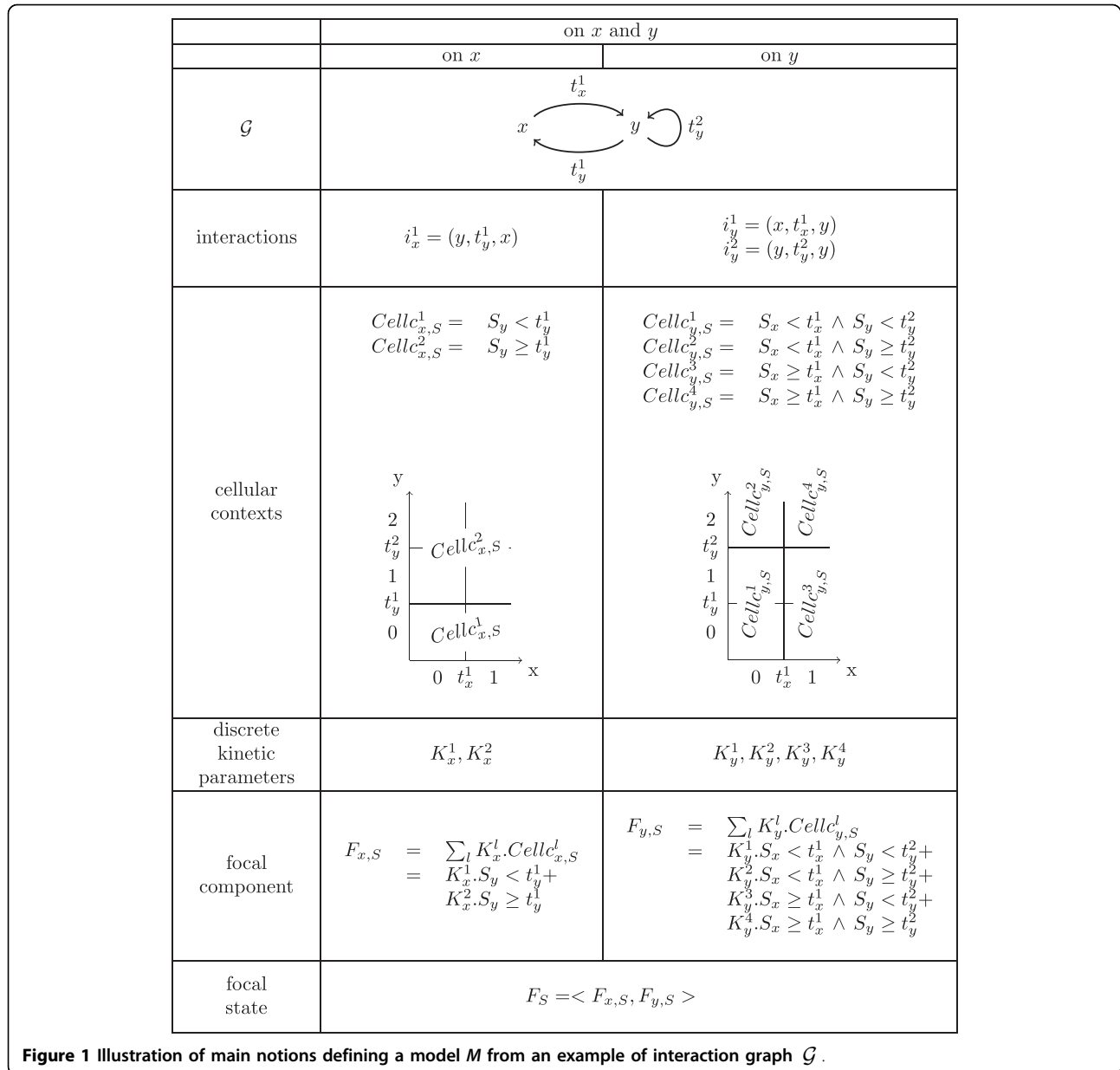
In this subsection we present a constraint-based formalism to impose, check and infer properties about discrete genetic networks as defined by R. Thomas. We first introduce the notions needed to define and formalize the interaction graph and the evolution rules of Thomas networks. We define in the next subsection the notions of composition of interactions, additivity and observability properties which are useful to express hypotheses about kinetic parameters. All the presented notions of

this subsection and the next one are illustrated with the example of Figure 1 and Figure 2, and will be put into use in the biological applications of the section Results.

The structure of a GRN is represented in an abstract way by an interaction graph \mathcal{G} . The nodes of \mathcal{G} are genes. Each node is associated to a concentration variable representing the concentration of the protein produced by the corresponding gene. The oriented edges of \mathcal{G} represent *interactions* between these genes, denoted by i_c^r for the interaction on the gene (component) c of index r , $r \in 1..r_c$, where r_c is the number of interactions on c . An integer variable representing a discrete *threshold* labels each edge. In papers using the formalism of R. Thomas edges in interaction graphs are also labeled by a sign [1]. We choose more primitive interaction graphs without these labels in order to generalize the formalization and to facilitate the expression of hypotheses about the way interactions compose on target nodes. The presence of an interaction from gene a to gene b (with a threshold t) indicates that protein a can potentially modify the expression rate of gene b . Furthermore, this change in expression rate, when it actually exists, takes place when concentration of a crosses threshold t . In other words, this interaction indicates that the rate of production of protein b can be influenced by the position of protein a with respect to threshold t . It has to be noted that such an interaction does not actually impose a difference in production rate. Rather, the absence of such an interaction forbids the existence of such a difference in production rate. Such an interaction is represented by the triplet (a, t, b) (labelled edge). In the example in Figure 1, the rows “ \mathcal{G} ” and “interactions” give respectively the definition of the interaction graph \mathcal{G} and the sets of interactions for the target genes x and y .

The network structure being defined, the next step is to define the network state and dynamics. A *state* S of the network is a list of gene product concentrations (protein or RNA). The concentrations are discretized according to the thresholds appearing in \mathcal{G} . The concentration of the product of gene c in state S is the integer $S_c \in 0..max_c$, where max_c is the maximal value of the discrete concentration of protein c . The threshold of component c of index p is $t_c^p \in 1..max_c$, where the index p takes its values in $1..max_c$ (obviously if a concentration is cut by max_c thresholds the associated discretized variable will take $max_c + 1$ values). For a given system with n genes, c_i , $i \in 1..n$, is the variable associated to the i^{th} gene, and state S is the ordered list $\langle S_{c_1}, \dots, S_{c_n} \rangle$. So, the discrete concentration space contains $\prod_{c_i} (max_{c_i} + 1)$ states.

We are now in a position to explain how the successor states of a given state are computed. Each state S is associated to a so-called *focal state*, denoted by



$F_S, F_S = \langle F_{c_1, S}, \dots, F_{c_n, S} \rangle$, and belonging to the same state space. The focal state gives the direction of evolution (tendency) for each concentration. Consider for instance $S = \langle 0, 0 \rangle$ and $F_S = \langle 1, 1 \rangle$ in a 2-dimensional system. The successor of S is not $\langle 1, 1 \rangle$ as is the case in synchronous updating schemes. Rather $F_S = \langle 1, 1 \rangle$ indicates the direction of evolution of each component taken separately. Here both are increasing and $S = \langle 0, 0 \rangle$ has 2 successors, $\langle 1, 0 \rangle$ and $\langle 0, 1 \rangle$. In other words two transitions are possible from S , and this type of updating scheme is often called asynchronous, but nondeterministic is a better term. What is the basis of this non-determinism? If the numerical values (real numbers) of the initial concentrations, together with those of the model

parameters were known, it would be possible to determine the exact successor. In the discrete abstraction considered here this information is not available and consequently both possibilities must be taken into account. Non-determinism is a fundamental property of this abstraction due to the information loss induced by the partition of concentration space into rectangular domains. We chose this formalism in this study because it is well founded and it is a good match to the qualitative knowledge generally available in Systems Biology at present. Nevertheless, it should be kept in mind that our constraint approach is not tied to Thomas networks. Other types of discrete dynamical rules could be implemented, e.g. Kauffman-like Boolean networks with

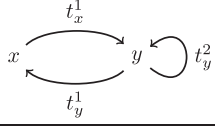
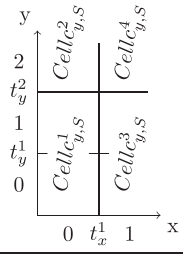
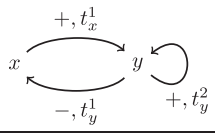
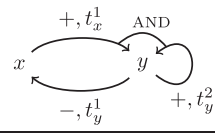
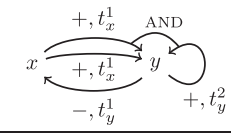
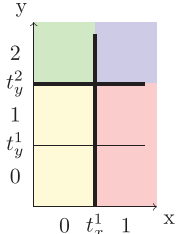
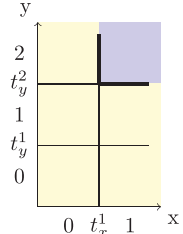
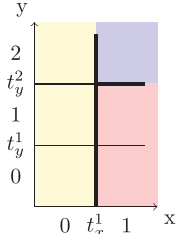
\mathcal{G}			
interactions on x	$i_x^1 = (y, t_y^1, x)$		
interactions on y	$i_y^1 = (x, t_x^1, y)$ $i_y^2 = (y, t_y^2, y)$		
cellular contexts on y			
interaction compositions on x	$ic_x^1 = \neg i_x^1$		
interaction compositions on y	$ic_y^1 = i_y^1$ $ic_y^2 = i_y^2$	$ic_y^1 = i_y^1 \wedge i_y^2$	$ic_y^1 = i_y^1$ $ic_y^2 = i_y^1 \wedge i_y^2$
informal graphical representation of interaction compositions			
compositions of cellular contexts of y			

Figure 2 Examples of interaction compositions and resulting compositions of cellular contexts over example in Figure 1.

parallel (synchronous) [4] or block-sequential updating scheme [5,6].

To implement the Thomas evolution rules we need first to specify the equations which link a state S to its focal state F_S . These equations are named *focal equations*. A set of rules then links state S , the focal state F_S associated to S , and the successor states of S . We stress here that these rules must be viewed as relationships linking different kinds of unknowns. As explained above (reversibility), the use of these relationships depends on the available information in a given state of knowledge. If the concentration values making state S are all known, together with the position of its focal state, then the successors of S can be computed. But the relationships can be exploited in other ways, too.

The system of focal equations contains different kinds of parameters: constant concentrations associated to *input genes* (that is genes that are influenced by no genes in the network and whose state is fixed by external conditions), parameters related to reaction kinetics (similar to those that would appear in a differential description), and thresholds t_c^p . The set of all these parameters is denoted by P . The parameters are among the unknowns of the system of constraints because their values are in general not known, or only partially known. The evolution rules, once formalized (see below), lead to a first set of logical constraints. To this first set are added structural constraints over the parameters derived from experimental data, and working hypotheses. The set of solutions of such a system of

constraints defines a set of instantiated models (i.e. models in which all parameters are instantiated). The couple composed of a focal equation system and a set of structural constraints is called a *parameterized constrained model* M , or just *model*. A typical query includes one or several structurally-related models (when data are available on several mutants), and some additional behavioural constraints. If the resulting system (set of all constraints of the query) is under-constrained this set contains a large number of solutions. If it is over-constrained it is empty. In our context, this last case is interpreted as a contradiction between, on one hand, the experimental evidence and, on the other hand, the network structure or the hypotheses. More sophisticated queries are presented in the application part below, to illustrate the high-level functionalities mentioned in the introduction.

The parameterized focal equation system of a model M is completely defined by an interaction graph \mathcal{G} . In fact these two entities contain exactly the same information (as long as kinetic parameters are not instantiated nor constrained). The set of interactions of \mathcal{G} having the gene c as target induces a partition of the concentration space according to the thresholds t_c^p of these interactions. This partition defines a set of regions called the *cellular contexts* of c . As long as the concentration of the proteins c' regulating c do not cross one of the t_c^p thresholds, the system stays in the same cellular context, because from the viewpoint of gene c the regulatory conditions have not changed. This means that all the states S belonging to the same cellular context of c have the same focal component $F_{c,S}$ of the focal state F_S . The value of $F_{c,S}$ being generally unknown, a formal parameter K_c^l called *discrete kinetic parameter* is introduced for each cellular context of c with index l . These parameters are the discrete version of the ratio of protein production rate over degradation rate. When the value of K_c^l is high in some cellular context, this is interpreted by saying that in the states belonging to that context the production rate of the protein associated to gene c is high, and/or its degradation rate low. But in the qualitative setting of Thomas formalism it should be kept in mind that we have only access to a discretized version of the production rate to degradation rate ratio. The number of cellular contexts for a given gene c is $l_c = 2^{r_c}$, and so is the number of K_c^l parameters.

We have introduced the main notions unformally (interaction, threshold, state, focal state, focal equation, cellular context, discrete kinetic parameter), and will now present formal definitions which are directly usable in constraint form.

Definition 1 Let c be a component, and let S be a state. The **focal component** of c in S , denoted by $F_{c,S}$ is defined by the following **focal equation** of c :

$$F_{c,S} = \sum_{l=1}^{l_c} K_c^l \cdot \text{Cell}c_{c,S}^l$$

where $K_c^l \in 0..max_c$ is the **discrete kinetic parameter** of c with index l , $l \in 1..l_c$ and $\text{Cell}c_{c,S}^l$ is a condition true if S belongs to the **cellular context** of c with index l . The indexing convention is the following: l is equal to $V + 1$ where V is the decimal value of the binary number composed of the Booleans $S_{c'} \geq t_c^p$ with $i_c^r = (c', t_c^p, c)$, these Booleans being arranged in increasing order of r (this is just meant at providing a unique numbering of the cellular context and is not fundamental).

The above formula means that if state S' belongs to the cellular context of index l' for gene c (that is $\text{Cell}c_{c,S'}^{l'}$ is true) then the focal component $F_{c,S'}$ is equal to $K_c^{l'}$.

Example 1 The row “cellular contexts” in Figure 1 describes formally and graphically, for a given order of thresholds, the cellular contexts for each component x and y of the considered example. Component x is the target of only one interaction and is thus associated to two cellular contexts, y is the target of two interactions and has 4 cellular contexts. The row “discrete kinetic parameters” gives the list of these parameters. The subscripts and superscripts make the correspondence with the associated cellular contexts (K_x^l with $\text{Cell}c_{x,S}^l$, etc.). Finally the row “focal component” gives the equations describing the focal components $F_{x,S}$ and $F_{y,S}$ of a state S . The row “focal state” in Figure 1 describes the focal state $F_S = \langle F_{x,S}, F_{y,S} \rangle$ of S .

The focal state defines the direction of the dynamic transitions starting in S . In the Thomas networks, the authorized transitions are such that:

1. S' and S are the same state or are neighbors,
2. S' and S differ on at most one component.
3. S' is in the “direction” of the focal state F_S .

The first property (formally $\forall c \ S_c - 1 \leq S'_c \leq S_c + 1$) is due to the fact that the concentrations evolve continuously, thus jumps over states are not allowed. The second is commonly called asynchronicity. The third one is specific to the discretization of evolution equations due to Thomas. We explained above that when two concentrations are increasing in a given state, it is not known in this kind of abstraction which will reach first its next threshold, and consequently which transition will occur first. In this situation both transitions are taken into account leading to two successors for the state considered (of course this generalizes to more than two). This is intimately connected to the non-determinism inherent to abstractions based on phase space partition.

The rules have the following consequences: $S = S' \Leftrightarrow F_S = S$ (stationarity of S), $S'_c = S_c + 1 \Rightarrow F_{c,S} > S_c$ (rising transition according to c) and $S'_c = S_c - 1 \Rightarrow F_{c,S} < S_c$ (downward transition according to c).

It is possible to specify a knock-out or ectopic expression *mutation*. For each non-input mutated gene c set to a constant value v , the constraint $\Lambda_l K_c^l = v$ must be introduced. For a mutated input gene to the value v the input parameter of the model is set to this value v . In some cases it is necessary to use several models in the same query, one model corresponding to the wild type and the others to mutants. In such cases we introduce constraints specifying that for all couples of models (M^α, M^β) the thresholds of M^α are equal to those of M^β , and the parameters K_c^l of M^α associated to genes c which are not mutated in M^α and M^β are equal to those of M^β . The constraints between the input parameters of M^α and M^β depends of the considered biological application (see Constraint 4, in the section Results and Discussion).

A user of GNBox must describe the structure of the studied GRN (possible interactions between genes), and can use the language *LG1* to specify the existence of a behaviour. The language *LG1* is composed of the predicate *path*($M, Path, L$) which is true if *Path* is a succession of L states authorized by the model M (achieving a formal link between a model and its behaviours), and a language to impose arithmetic constraints between variables of *Path*. Language *LG1* is used to formalize observations on the behaviour of the system. Our approach allows to specify (declare) partial information. For example only a few concentrations may have been measured. Absence of information is absence of constraints.

Interaction compositions

The interaction graph \mathcal{G} lists the interactions individually but does not contain information on the manner in which different interactions are composed when they have the same target gene. The information about the way to compose interactions is embodied in relationships linking the parameters contained in $P(K_c^l, t_c^p, \dots)$. However, the manual interpretation of instantiations or properties over parameters of P is not convenient, especially for users not acquainted with the formalism of Thomas networks. For this reason we designed a higher level language *LG2* to impose, check and infer properties about the way to compose the interactions in \mathcal{G} in the manner of the traditional notion of the logic of regulation (NEG, AND, OR gates). It should be understood that this is not fundamental to the approach but merely a facility to handle relationships between parameters induced by the composition of interactions. The user always has the choice to work directly on these relationships.

We explained above that the specification of a set of interactions for a gene c partitions (in cellular contexts) the concentration space by hyperplanes (corresponding to thresholds of interactions acting on c). *LG2* permits, for every c , to partition the concentration space in union of cellular contexts of c , named *compositions of cellular contexts*, via the definition of *interaction compositions*. Any union of cellular contexts can be specified, and in particular an union of disconnected regions. Similarly to the semantic of a set of interactions, the semantic of a set of interaction compositions is the following: all the states belonging to a given composition of cellular contexts of c have the same evolution tendency of the concentration of protein c . The borders between these regions are constituted of parts of threshold hyperplanes of interactions taking part in the composition. We name these borders interaction compositions. An interaction composition for c , denoted by ic_c^{rc} , $rc \in 1..rc_c$, rc_c being the number of interaction composition on c , permits to indicate where it is possible to have a change in the evolution trend of component c . Informally, one can see an interaction composition as a new artificial species which interacts on c and which induces a new partition of state space into two regions. First, let us remark again that an interaction $i_c^r = (c', t_c^p, c)$ induces a partition of state space into two regions by the hyperplane associated to the threshold t_c^p . By convention, the part where the states S are such that $S_{c'} \geq t_c^p$ is true is said to *satisfy* i_c^r . An interaction composition also partition the state space into two regions, but the border is not necessarily a hyper-plane defined by a single threshold. An interaction composition can have the following forms:

- an interaction i_c^r .
- $\neg i_c^r$. The region where the state S are such that $S_{c'} < t_c^p$ is said to satisfy $\neg i_c^r$.
- $ic \wedge ic'$, where ic and ic' are interaction compositions. The region where the states S satisfies both ic and ic' is said to satisfy $ic \wedge ic'$.
- $ic \vee ic'$, where ic and ic' are interaction compositions. The region where the states S satisfies ic , or ic' , or both, is said to satisfy $ic \vee ic'$.

Example 2 The sixth row in Figure 2 gives three possible sets of interactions compositions for y , related to the example in Figure 1. The four first rows recall the context of example in Figure 1. The fifth row gives the set of interaction compositions over x . The seventh row shows for each of these couple of sets a graphical representation of the detailed structure of the network, with signs + and - over interactions and bridges, denoted by AND, to express a conjunction between two interaction compositions. Finally, the last row shows the resulting compositions of cellular contexts for y .

The first case (second column of last row) leads to the same partition of the discrete concentration space of y (the areas described by the cellular contexts are the same that those described by the compositions of cellular contexts).

The second case expresses with the sole interaction composition on y , $i_y^1 \wedge i_y^2$, that either the concentration of x and y are above t_x^1 and t_y^2 , respectively (the tendency of y is unique in this region), or the concentration of x or y are below t_x^1 and t_y^2 , respectively (the tendency of y is unique in this region).

The third case expresses quite the same of the second case but permits that x interacts on y whatever the concentration of y . So, we obtain three compositions of cellular contexts because the fact that x can interact on y all along the border of the threshold t_x^1 .

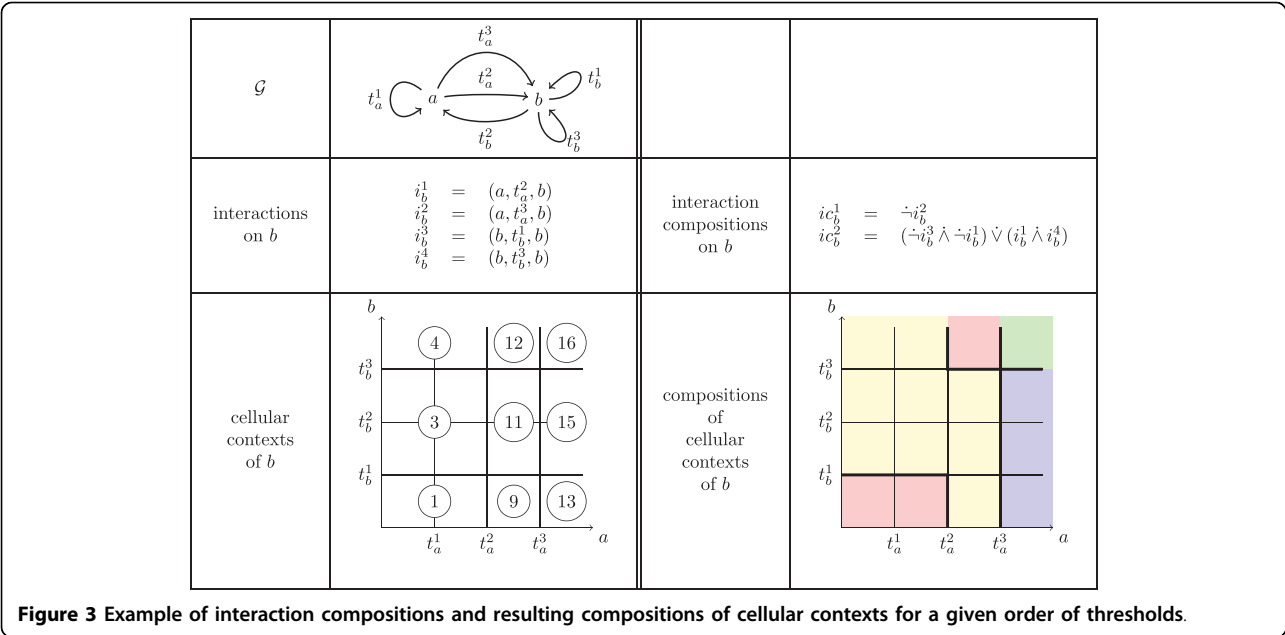
Example 3 The Figure 3 gives an example of a set of interaction compositions and resulting composition of cellular contexts. In the first column, we can see an interaction graph \mathcal{G} with two components a and b , a set of four interactions over b , and a partition of the concentration space into nine non empty cellular contexts. The indexes l of the conditions $\text{Cell } c_{b,S}^l$ appear in circles. The other cellular contexts are empty according to the order of the values of the thresholds ($t_a^1 < t_a^2 < t_a^3, t_b^1 < t_b^2 < t_b^3$). Note that usually this order is not known and the values of thresholds for a same species can be equal. In the second column (to make a parallel with the interactions and cellular contexts) we assumed to have two interaction compositions. We obtain a partition of the concentration space into four non empty compositions of cellular contexts (the pink region being the union of the two disconnected cellular contexts 1 and 12).

Additivity and observability properties

The language $LG2$ allows to define specific effects of an interaction composition on a component c . Here by effect we mean a shift in the position of the focal component F_c when the border associated to the interaction composition is crossed. Biologically, an increase of the tendency of c can be due to an increase of the expression rate of gene c , or a decrease of the degradation rate of the corresponding protein. In other formalisms these effects are specified by labelling the arcs of the interaction graph with signs (we have used this in the 7th row in Figure 2). A $+$ sign (respectively a $-$ sign) for an interaction of a gene a on b in the signed interaction graph means informally that the interaction of a on b is an *activation* (respectively an *inhibition*). However, the exact meaning of the terms activation and inhibition is not clear, especially when several interactions combine on a gene: Does an activation of b by a forbid an inhibition of b by a or not? Is an activation of b by a necessarily observed all along the border associated to the interaction or not? Two properties are used to clarify formally these questions.

The first one, called *additivity*, is the systematic non-strict increase of tendency of c when a border is crossed in some predefined direction. In other words the effect on c of the interaction composition adds to the effect of all other interaction compositions on c . The direction in which the border is crossed for this property is the one given by the passage from a state where the interaction composition is not satisfied to a state where it is satisfied.

The second property, called *observability*, is the existence of a strict increase of the tendency on c . This means that the effect on the tendency of c exists at least



at one crossing point (where the border associated to the interaction composition is crossed in the same direction as the additivity property). In contrast to the additivity property, observability property requires only the existence of an effect somewhere along the border.

To define these effects more formally we introduce for each interaction composition ic_c^{rc} on c a set, denoted by Adj_c^{rc} , containing all couples of states $(S0, S1)$ such that (i) $S0$ is adjacent to $S1$, (ii) $S0$ is a state in the region where ic_c^{rc} is not satisfied, and (iii) $S1$ is a state in the region where ic_c^{rc} is satisfied.

Example 4 For the interaction composition ic_b^2 of the example given in Figure 3 we get $Adj_b^2 = \{(\langle 0, 1 \rangle, \langle 0, 0 \rangle), (\langle 1, 1 \rangle, \langle 1, 0 \rangle), (\langle 2, 0 \rangle, \langle 1, 0 \rangle), (\langle 1, 3 \rangle, \langle 2, 3 \rangle), (\langle 2, 2 \rangle, \langle 2, 3 \rangle), (\langle 3, 2 \rangle, \langle 3, 3 \rangle)\}$. Each of the couples $(S0, S1)$ of this set is represented in Figure 4 by a kind of arrow symbol, where the 'o' end is associated to state $S0$, and the 'l' end to state $S1$.

LG2 allows to specify that an interaction composition ic_c^{rc} has an additive effect, denoted by $a(ic_c^{rc})$, i.e. that the difference of trend of c is positive or zero all along the border defined by ic_c^{rc} . The exact semantics of $a(ic_c^{rc})$ is: for every couple $(S0, S1)$ of Adj_c^{rc} the trend of c in $S0$ is less than or equal to the trend of c in $S1$. Since the trend of a state is equal to the trend of all the states in the same cellular context, the additivity constraints are expressed as relations between discrete kinetic parameters K_c^l .

Example 5 For the example in Figure 2 (with the given order of thresholds) we have $Adj_x^1 = \{(\langle 0, 0 \rangle, \langle 1, 0 \rangle), (\langle 0, 1 \rangle, \langle 1, 1 \rangle), (\langle 0, 2 \rangle, \langle 1, 2 \rangle)\}$, and $a(ic_x^1) \Leftrightarrow (K_x^2 \leq K_x^1)$ due to the negative sign associated to the interaction of y on x .

For the first case of interaction compositions on $Adj_y^1 = \{(\langle 0, 0 \rangle, \langle 1, 0 \rangle), (\langle 1, 1 \rangle, \langle 1, 1 \rangle), (\langle 0, 2 \rangle, \langle 1, 2 \rangle)\}$, $a(ic_y^1) \Leftrightarrow (K_y^1 \leq K_y^3 \wedge K_y^2 \leq K_y^4)$, and $Adj_y^2 = \{(\langle 0, 1 \rangle, \langle 0, 2 \rangle), (\langle 1, 1 \rangle, \langle 1, 2 \rangle)\}$, $a(ic_y^2) \Leftrightarrow (K_y^1 \leq K_y^2 \wedge K_y^3 \leq K_y^4)$.

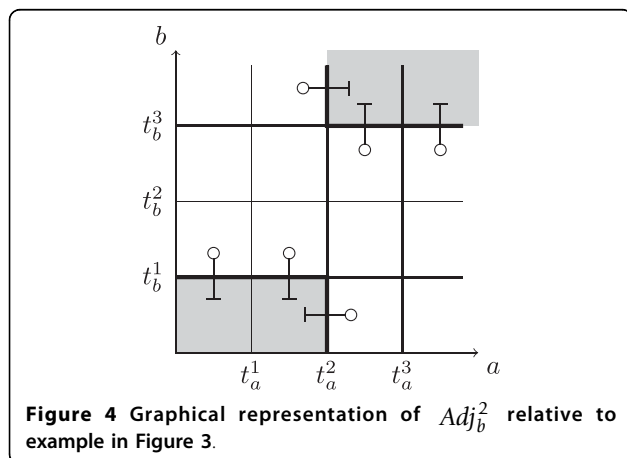


Figure 4 Graphical representation of Adj_b^2 relative to example in Figure 3.

For the second case of interaction compositions on y , $Adj_y^1 = \{(\langle 0, 2 \rangle, \langle 1, 2 \rangle), (\langle 1, 1 \rangle, \langle 1, 2 \rangle)\}$ and $a(ic_y^1) \Leftrightarrow (K_y^2 \leq K_y^4 \wedge K_y^3 \leq K_y^4)$. If this additivity property is true, the only case of activation of y is when x and y are above t_x^1 and t_y^2 respectively.

For the third case $a(ic_y^1) \Leftrightarrow (K_y^1 \leq K_y^3 \wedge K_y^2 \leq K_y^4)$ (the same as the first one in the first case because ic_y^1 is the same) and $a(ic_y^2) \Leftrightarrow (K_y^2 \leq K_y^4 \wedge K_y^3 \leq K_y^4)$ (the same as the first one in the second case). If these additivity are true, there are two cases of activation of y , one above t_x^1 and one above t_x^1 and t_y^2 . Moreover, the second case of activation is greater than the first one, due to the additivity property $a(ic_y^2)$.

If multi-arcs are present in the interaction graph (several arcs with the same origin and the same target node) the cellular contexts on each side of the border defined by the interaction composition ic_c^{rc} are not the same depending on the values of the thresholds associated to the multi-arc. In that case the additivity constraints are relations involving also thresholds. Briefly, the additivity constraint of the interaction composition ic_c^{rc} is: $\Lambda(adj(l0, l1, rc) \Rightarrow K_c^{l0} \geq K_c^{l1})$ with $adj(l0, l1, rc)$ true if it exists a couple $(S0, S1)$ of $a(ic_y^2) \Leftrightarrow (K_y^1 \leq K_y^2 \wedge K_y^3 \leq K_y^4)$ such that $\neg Cellc_{c,S0}^{l0} \wedge Cellc_{c,S1}^{l1}$ (the cellular contexts $l0$ and $l1$ are non empty, adjacent, and on each side of the border defined by ic_c^{rc}).

Example 6 For the example in Figure 3, the additivity constraint of the composition $ic_b^1 = \neg t_b^2$ is:

$$\begin{aligned} (K_b^{13} \leq K_b^9 &\Leftarrow (t_a^2 < t_a^3 \quad)) \quad \wedge \\ (K_b^{16} \leq K_b^{12} &\Leftarrow (t_a^2 < t_a^3 \quad)) \quad \wedge \\ (K_b^{15} \leq K_b^{11} &\Leftarrow (t_a^2 < t_a^3 \quad \wedge \quad t_b^1 < t_b^3)) \quad \wedge \\ (K_b^{14} \leq K_b^{10} &\Leftarrow (t_a^2 < t_a^3 \quad \wedge \quad t_b^1 > t_b^3)) \quad \wedge \\ (K_b^{13} \leq K_b^1 &\Leftarrow (t_a^2 = t_a^3 \quad)) \quad \wedge \\ (K_b^{16} \leq K_b^4 &\Leftarrow (t_a^2 = t_a^3 \quad)) \quad \wedge \\ (K_b^{15} \leq K_b^3 &\Leftarrow (t_a^2 = t_a^3 \quad \wedge \quad t_b^1 < t_b^3)) \quad \wedge \\ (K_b^{14} \leq K_b^2 &\Leftarrow (t_a^2 = t_a^3 \quad \wedge \quad t_b^1 > t_b^3)) \quad \wedge \\ (K_b^5 \leq K_b^1 &\Leftarrow (t_a^2 > t_a^3 \quad)) \quad \wedge \\ (K_b^8 \leq K_b^4 &\Leftarrow (t_a^2 > t_a^3 \quad)) \quad \wedge \\ (K_b^7 \leq K_b^3 &\Leftarrow (t_a^2 > t_a^3 \quad \wedge \quad t_b^1 < t_b^3)) \quad \wedge \\ (K_b^6 \leq K_b^2 &\Leftarrow (t_a^2 > t_a^3 \quad \wedge \quad t_b^1 > t_b^3)) \quad \wedge \end{aligned}$$

according to the identifiers l of cellular contexts for b (and so the identifiers of discrete kinetic parameters K_b^l). It can be checked with the graphic representation of cellular contexts of b in Figure 3 that for $t_a^2 < t_a^3 \wedge t_b^1 < t_b^3$ we obtain $a(ic_b^1) \Leftrightarrow (K_b^{13} \leq K_b^9 \wedge K_b^{15} \leq K_b^{11} \wedge K_b^{16} \leq K_b^{12})$. This example shows that specifying additivity properties can be much more compact than working at the level of parameters. Without language LG2 we would have to write the above formula.

In addition to the additivity property, LG2 allows to specify that an interaction composition ic_c^{rc} has an

observable effect, denoted by $o(ic_c^{rc})$, i.e. that the difference of trend of c is strictly positive at least at one position along the border defined by ic_c^{rc} . The exact semantics of $o(ic_c^{rc})$ is: for at least one couple $(S0, S1)$ of Adj_c^{rc} the trend of c in $S0$ is strictly less than the trend of c in $S1$. To be more explicit, an interaction i_c^r can be removed from the interaction graph if neither the interaction composition i_c^r (reduced to a single interaction), nor its negation $\neg i_c^r$ is observable. Briefly the observability constraint of the interaction composition ic_c^{rc} is: $\vee (adj(l0, l1, rc) \wedge K_c^{l0} < K_c^{l1})$ with $adj(l0, l1, rc)$ true if it exists a couple $(S0, S1)$ of Adj_c^{rc} such that $\neg Cellc_{c,S0}^{l0} \wedge Cellc_{c,S1}^{l1}$.

Example 7 For the example in Figure 3 with $t_a^2 < t_a^3 \wedge t_b^1 < t_b^3$ the constraint $o(ic_b^2)$ is $(K_b^3 < K_b^1) \vee (K_b^9 < K_b^1) \vee (K_b^4 < K_b^{12}) \vee (K_b^{11} < K_b^{12}) \vee (K_b^{15} < K_b^{16})$.

GNBox Features

The core functionality of the GNBox environment is to test, for a given structure of a GRN, the consistency of a set of hypotheses about the behaviours of this GRN (language *LG1*) for several mutant types, about the interaction compositions (language *LG2*), and even directly about the parameters in P . GN-Box is able to identify consistent solutions in terms of state variables that define the behaviour (*LG1*) and in terms of parameters of P . In cases where the set of hypotheses is inconsistent, it is desirable to determine the possible relaxations of hypotheses to remove the inconsistency. GNBox can identify automatically, among a defined set of questionable hypotheses, all subsets of hypotheses whose relaxation removes the inconsistency (subsets of necessarily false hypotheses). These subsets are represented as disjunctions of conjunctions of negations of hypotheses. For example, the hypotheses $H1$ and $H2$ must be relaxed or the hypothesis $H3$ must be relaxed: $(\neg H1 \wedge \neg H2) \vee \neg H3$. Also GNBox automatically identifies, among a defined set of hypotheses, all subsets of hypotheses necessarily true. These subsets are represented by disjunctions of conjunctions of hypotheses. For example, the hypotheses $H1$ and $H2$ are true or the hypothesis $H3$ is true: $(H1 \wedge H2) \vee H3$.

Results and Discussion

Application to the immunity control by the λ bacteriophage

The analysis of this network adapted from [7] illustrates mainly the capability of GNBox (i) to express constraints about reachability of states, and (ii) to find the minimal interaction graph consistent with observations.

The λ bacteriophage (or simply λ phage) is a virus that infects the bacterium *Escherichia coli*. The infection starts by the injection of the genetic material of the virus into the cytoplasm of the bacterium. We focus here on two simple observations about the evolution of the

bacterium after infection: either the viral DNA is integrated in the genetic material of the bacterium, and the cells continue to divide normally (thus reproducing the phage DNA in the same process), or the genetic material replicates in the cytoplasm of the bacterial cell to create new viral particles and then new viruses until lysis (destruction) of the cell, which leads to the release of new virus particles in the extracellular medium. The first case corresponds to the *lysogenic phase* while the second corresponds to the *lytic phase*. The decision between these two phases is made by a network of viral genes.

The model proposed in [7] contains four viral genes denoted by cI , cro , cII and n . The gene cI is expressed only in the lysogenic phase, cro is expressed only in the lytic phase and genes cII and n are not expressed in both phases. The graph \mathcal{G} of interactions between these genes is given in Figure 5. Interactions and interaction compositions (deduced from experimental data) are given in Table 1. We consider the set of all additivity and observability constraints for all these interaction compositions ($a(ic_{cl}^1)$, $o(ic_{cl}^1)$, etc.). In the following we assume that the thresholds t_c^p are ordered so that $t_c^p = p$. According to the previous section the set of interactions, the hypotheses about interaction compositions (set of interaction compositions, additivity properties, observability properties) and the hypotheses on threshold values define a parameterized constrained model (couple composed of a focal equation system derived from \mathcal{G} and a set of structural constraints). We call it M_λ and it is defined formally by the predicate $model_\lambda(M_\lambda)$. A state S for this model is represented by an ordered list $\langle S_{cI}, S_{cro}, S_{cII}, S_n \rangle$ of discrete protein concentrations. According to \mathcal{G} and hypotheses on threshold values, we have $max_{cI} = 2$, $max_{cro} = 3$, $max_{cII} = 1$, $max_n = 1$. So the concentration space contains $(2+1)*(3+1)*(1+1)*(1+1) = 48$ states.

The uninfected cell does not have any viral protein and can therefore be represented by the state $S0 = \langle S_{cI}, S_{cro}, S_{cII}, S_n \rangle = \langle 0, 0, 0, 0 \rangle$. In the lysogenic phase of

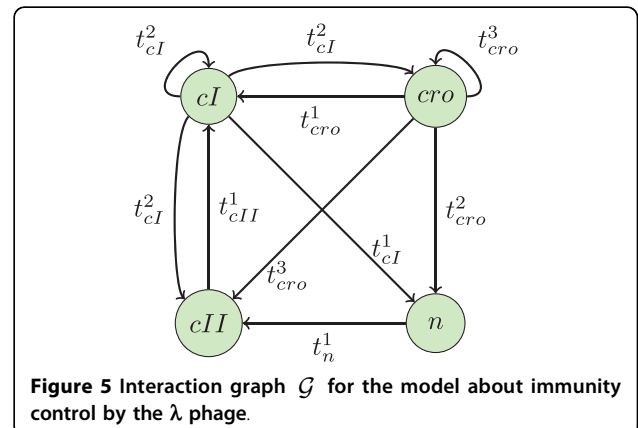


Table 1 Interactions and interaction compositions hypotheses for the model about immunity control by the λ phage

species	interactions	interaction compositions
cl	$i_{cl}^1 = (cI, t_{cl}^2, cI)$ $i_{cl}^2 = (cro, t_{cro}^1, cI)$ $i_{cl}^3 = (cII, t_{cII}^1, cI)$	$ic_{cl}^1 = i_{cl}^1$ $ic_{cl}^2 = \neg i_{cl}^2$ $ic_{cl}^3 = i_{cl}^3$
cro	$i_{cro}^1 = (cI, t_{cl}^2, cro)$ $i_{cro}^2 = (cro, t_{cro}^3, cro)$	$ic_{cro}^1 = \neg i_{cro}^1$ $ic_{cro}^2 = \neg i_{cro}^2$
cII	$i_{cII}^1 = (cI, t_{cl}^2, cII)$ $i_{cII}^2 = (cro, t_{cro}^3, cII)$ $i_{cII}^3 = (n, t_n^1, cII)$	$ic_{cII}^1 = \neg i_{cII}^1$ $ic_{cII}^2 = \neg i_{cII}^2$ $ic_{cII}^3 = i_{cII}^3$
n	$i_n^1 = (cI, t_{cl}^1, n)$ $i_n^2 = (cro, t_{cro}^2, n)$	$ic_n^1 = \neg i_n^1$ $ic_n^2 = \neg i_n^2$

the virus-host system the only viral gene expressed is cI . This phase is represented by the state $S1 = \langle S1_{cl}, S1_{cro}, S1_{cII}, S1_n \rangle = \langle 2, 0, 0, 0 \rangle$ such that the concentration of protein cI remains equal to its highest value. In the lytic phase the only viral gene expressed is cro . In a continuous description this phase is represented by a state which is not contained within a domain, but which is at the border between two adjacent domains. We could introduce in our formalism additional states corresponding to borders between domains. Such states are called *singular states* in [2]. We choose here to stick to the simpler formalism, and we represent the lytic phase as a cycle between the two following states: $S2 = \langle 0, 2, 0, 0 \rangle$ and $S3 = \langle 0, 3, 0, 0 \rangle$, such that the concentration of the protein cro remains around the highest values 2 and 3 [7]. Biological observations tell us that these two phases must be attractors of the network dynamics, and that they are reachable from the initial conditions. These observations are formalized by Constraint 1 where the lengths of the third and fourth paths for the reachability of the two phases are equal to 48 states, 48 being the total number of states of the state space.

Constraint 1

```

model_λ(M_λ)      ^
S0 = ⟨ 0, 0, 0, 0 ⟩ ^
S1 = ⟨ 2, 0, 0, 0 ⟩ ^
S2 = ⟨ 0, 2, 0, 0 ⟩ ^
S3 = ⟨ 0, 3, 0, 0 ⟩ ^
S23 = ⟨ 0, S23_cro, 0, 0 ⟩ ^
S23_cro ∈ 2..3    ^
path(M_λ, [S1, S1], 2) ^
path(M_λ, [S2, S3, S2], 3) ^
path(M_λ, [S0, ..., S1], 48) ^
path(M_λ, [S0, ..., S23], 48)

```

The GNBox environment proves the consistency of this pool of constraints in 2 seconds. All run times

mentioned in this article are obtained on a laptop with 2 GB of RAM and running at 2.4 GHz. Moreover GNBox can provide the instantiations of the parameters of P that satisfy the pool of constraints.

Example 8 An example of instantiation is:

$$\begin{aligned}
 K_{cl}^1 &= 0, K_{cl}^2 = 2, K_{cl}^3 = 0, K_{cl}^4 = 0, \\
 K_{cl}^5 &= 2, K_{cl}^6 = 2, K_{cl}^7 = 1, K_{cl}^8 = 1, \\
 K_{cro}^1 &= 3, K_{cro}^2 = 0, K_{cro}^3 = 0, K_{cro}^4 = 0, \\
 K_{cII}^1 &= 0, K_{cII}^2 = 1, K_{cII}^3 = 0, K_{cII}^4 = 0, \\
 K_{cII}^5 &= 0, K_{cII}^6 = 0, K_{cII}^7 = 0, K_{cII}^8 = 0, \\
 K_n^1 &= 1, K_n^2 = 0, K_n^3 = 0, K_n^4 = 0
 \end{aligned}$$

(remember that the indexes l of discrete kinetic parameters K_c^l are set at their creation from the numbering of interactions i_c^r . See Definition 1). The set of transitions from S to S' , denoted by $S \boxtimes S'$, for this instantiation is represented in Figure 6.

An interesting question, akin to reverse-engineering of the network, is: what are the minimal numbers of interaction compositions necessary to get a model consistent with Constraint 1 without specifying any additivity or observability constraints? In other words, we search for the minimal interaction graphs, in terms of interactions, which satisfy the observed behaviors. From a constraint point of view, this problem is specified and implemented in the following way. For each interaction composition on a gene c , a Boolean variable is created which means “all pairs of states separated only by this interaction composition have the same evolution tendency for c ”. Then GNBox searches for consistent models such that the number of these Boolean variables which are true is maximized. GNBox finds that only two interactions on cro are necessary (in two seconds). So, the minimal interaction graph contains only two interactions on cro and no interactions on the other genes. The result is surprising at first sight, but it should be borne in mind that the query contains only poor information about behaviours and no information on the interaction graph (but the limitation to possible interactions), the goal being to infer the minimum graph implied by this information. This does not preclude the existence of other interactions, but means that those are not necessary to account for the behaviours included in the query.

Finally this application lead us to the interesting question of the length L of the longest path without cycle in the state space for a given set of hypotheses Set . We call this length the *diameter* of the network for Set . This knowledge permits to restrict the length of paths in subsequent queries considering a set of hypotheses including Set . In our case Set is Constraint 1. The diameter for Set is 43. This highly combinatorial problem is

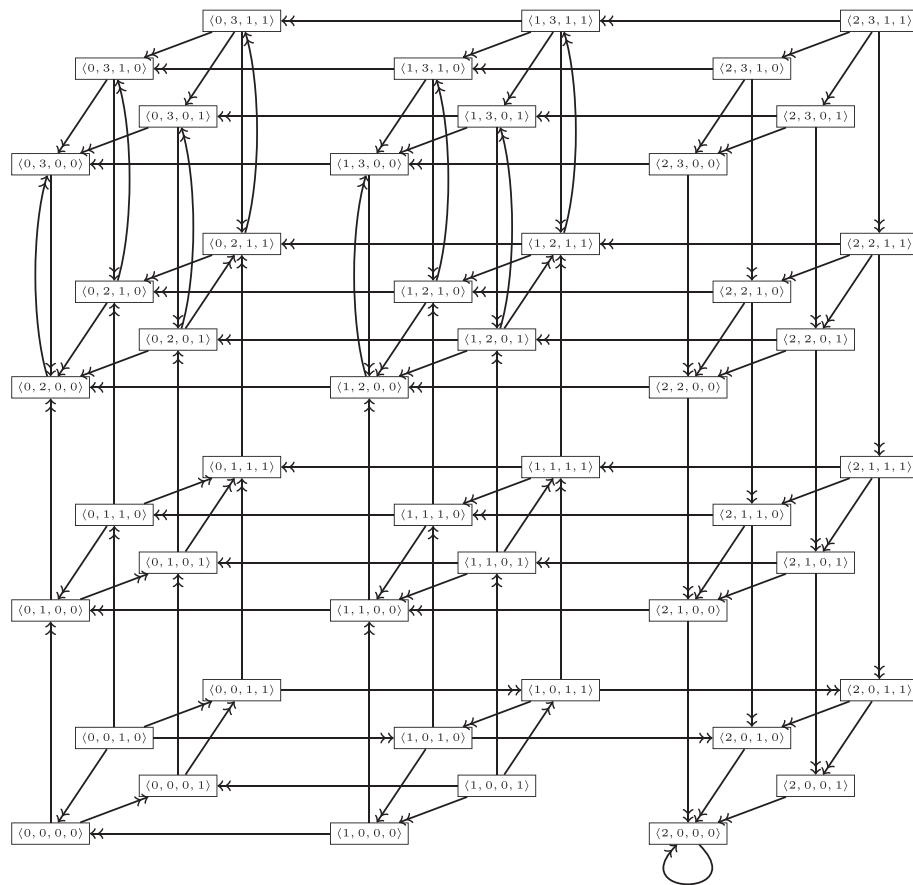


Figure 6 Set of possible transitions for the instantiation of parameters in Example 8 of the model about immunity control by the λ phage.

answered in two queries: one to prove the existence of a solution for a length L of 43 in 459 seconds, the other showing inconsistency for a length L of 44 in 489 seconds.

Application to the carbon nutritional stress response in the bacterium *Escherichia coli*

The modeling and analysis of this network is adapted from [8]. It illustrates a case of model revision coming from an inconsistency of the initial set of hypotheses. We performed a similar and more exhaustive study reported in [9]. We show that by an automatic relaxation method over biological constraints we can suggest lines of research to the biologist or, said differently, generate new hypotheses.

Populations of the bacterium *Escherichia coli* grow exponentially in favorable conditions. This state is called the *exponential phase*. In stressing conditions, when food (carbon) starts to be lacking, the populations stop growing and they enter in a state called *stationary phase*, with altered physiology and morphology. The phenomenon is reversible: the population can return to

the exponential phase if the conditions become favorable again.

The model, proposed in [8] and adapted to our formalism, contains one input node *sig* (signal, 0 in the absence of stress and 1 in the presence of stress) and five species: *crp*, *cya*, *fis*, *gyr* and *top*. The interaction graph \mathcal{G} is given in Figure 7 where the input *sig* is represented by a dotted circle filled in blue. Interactions and interaction compositions are given in Table 2. Moreover we consider the set of all additivity and observability constraints for all interaction compositions. As before, the thresholds t_c^p are ordered and equal to p . The model obtained from all these hypotheses is denoted by M_{coli} . Thus we have $max_{crp} = 2$, $max_{cya} = 2$, $max_{fis} = 3$, $max_{gyr} = 2$ and $max_{top} = 2$ (for the input *sig* we have $max_{sig} = 1$). We obtain a concentration space of $(2 + 1) * (2 + 1) * (3 + 1) * (2 + 1) * (2 + 1) = 324$ states.

The exponential phase and the stationary phase are modeled by two states, respectively S^{ns} (*ns* for “not stressed”) and S^s (*s* for “stressed”). As stated in [8], there exists partial knowledge about these states:

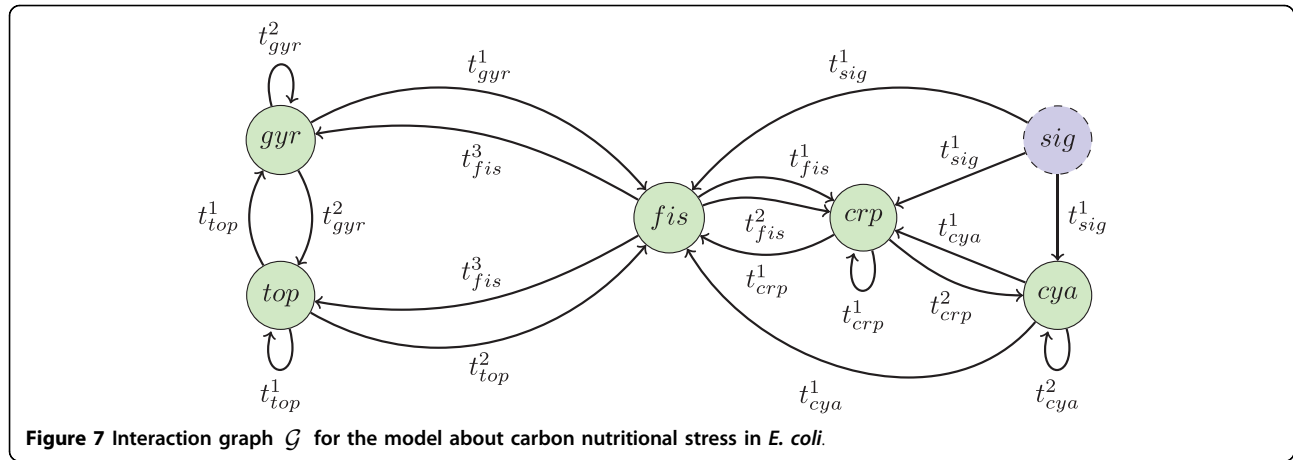


Figure 7 Interaction graph G for the model about carbon nutritional stress in *E. coli*.

Constraint 2

$$\begin{aligned} S_{crp}^{ns} = 1 & \wedge S_{crp}^s = 1 \wedge \\ S_{crp}^{ns} = 1 & \wedge S_{cya}^s = 1 \wedge \\ S_{fis}^{ns} = 1 & \wedge S_{fis}^s = 1 \wedge \\ S_{gyr}^{ns} - S_{gyr}^s & > S_{top}^{ns} - S_{top}^s \end{aligned}$$

Note that only three components are instantiated in each state and that there is a relationship between the two others which expresses the fact that the super-coiling of DNA is higher in the exponential phase. To model the presence or the absence of stress we use two

models: a model M_{coli}^{ns} without stress ($sig = 0$) and a model M_{coli}^s with stress ($sig = 1$). These two models are the same biological model M_{coli} in different conditions. So they share the same discrete kinetic parameters K_c^l . In absence of stress the system beginning in the stressed state S^s can reach the non-stressed state S^{ns} , which is steady. In presence of stress the system beginning in S^{ns} can reach S^s , which is steady. We formalize that by:

Constraint 3

$$\begin{aligned} path(M_{coli}^{ns}, [S^{ns}, S^{ns}], 2) & \wedge \\ path(M_{coli}^s, [S^s, S^s], 2) & \wedge \\ path(M_{coli}^{ns}, [S^s, \dots, S^{ns}], L) & \wedge \\ path(M_{coli}^s, [S^{ns}, \dots, S^s], L) & \end{aligned}$$

where L is the length of the third and fourth paths for the reachability of the two steady states. In the following queries we choose $L = 10$ and $L = 100$ to compare performance, but if we want a general query without any limitation on L we should choose $L = 324$ (the total number of states of the model) but the amount of memory needed to generate the pool of constraints becomes very large. We point out here that for such queries involving paths, this approach is limited to networks of medium size.

With GNBox we prove that the pool of constraints composed of constraints for models M_{coli}^s and M_{coli}^{ns} Constraint 2 and Constraint 3 is inconsistent in 2 seconds for path length $L = 10$ states, and 13 seconds for path length $L = 100$. In fact just imposing the existence of two steady states gives an inconsistency in less than 1 second, thus proving that the constraint pool is inconsistent whatever the value of L . In [8] it is noted that the proposed instantiated model is indeed inconsistent. Here we prove in addition that there exists no other instantiation of the discrete kinetic parameters (accepting the interaction compositions hypotheses) able to

Table 2 Interactions and interaction compositions hypotheses for the model about carbon nutritional stress in *E. coli*

species	interactions	interaction compositions
crp	$i_{crp}^1 = (sig, t_{sig}^1, crp)$	$ic_{crp}^1 = i_{crp}^1 \wedge i_{crp}^2 \wedge i_{crp}^3$
	$i_{crp}^2 = (crp, t_{crp}^1, crp)$	
	$i_{crp}^3 = (cya, t_{cya}^1, crp)$	
	$i_{crp}^4 = (fis, t_{fis}^1, crp)$	$ic_{crp}^2 = \neg i_{crp}^4$
	$i_{crp}^5 = (fis, t_{fis}^2, crp)$	$ic_{crp}^3 = \neg i_{crp}^5$
cya	$i_{cya}^1 = (sig, t_{sig}^1, cya)$	$ic_{cya}^1 = \neg i_{cya}^1 \vee \neg i_{cya}^2 \vee \neg i_{cya}^3$
	$i_{cya}^2 = (crp, t_{crp}^2, cya)$	
	$i_{cya}^3 = (cya, t_{cya}^2, cya)$	
fis	$i_{fis}^1 = (sig, t_{sig}^1, fis)$	$ic_{fis}^1 = \neg i_{fis}^1 \vee \neg i_{fis}^2 \vee \neg i_{fis}^3$
	$i_{fis}^2 = (crp, t_{crp}^1, fis)$	
	$i_{fis}^3 = (cya, t_{cya}^1, fis)$	
	$i_{fis}^4 = (gyr, t_{gyr}^1, fis)$	$ic_{fis}^2 = i_{fis}^4 \wedge \neg i_{fis}^5$
	$i_{fis}^5 = (top, t_{top}^1, fis)$	
gyr	$i_{gyr}^1 = (fis, t_{fis}^3, gyr)$	$ic_{gyr}^1 = \neg i_{gyr}^1$
	$i_{gyr}^2 = (gyr, t_{gyr}^2, gyr)$	$ic_{gyr}^2 = \neg i_{gyr}^2 \vee \neg i_{gyr}^3$
	$i_{gyr}^3 = (top, t_{top}^2, gyr)$	
top	$i_{top}^1 = (fis, t_{fis}^2, top)$	$ic_{top}^1 = i_{top}^1$
	$i_{top}^2 = (gyr, t_{gyr}^1, top)$	$ic_{top}^2 = i_{top}^2 \wedge \neg i_{top}^3$
	$i_{top}^3 = (top, t_{top}^1, top)$	

restore consistency. In other words it is proved that this network architecture with these hypotheses on interaction compositions is incompatible with the observations. It is thus necessary to revise the model. In [8] the authors suggest that a regulator or an interaction may be missing in the model. Here, instead, we keep the interaction graph \mathcal{G} as it is, and try to change the way interactions are composed. The set of unreliable hypotheses is the set of all additivity and observability properties about interaction compositions. We allow the relaxation of these hypotheses and we obtain the property $\neg a(ic_{gyr}^1) \vee \neg a(ic_{top}^1)$ in 7 seconds with a path length $L = 10$, and in 830 seconds with a path length $L = 100$. Discussions with the biologist lead to the conclusion that it is not acceptable to relax the additivity property of the first composition on *gyr*, $\neg(fis, t_{fis}^3, gyr)$. This suggests that the composition on *top*, (fis, t_{fis}^3, top) , is the one which is not additive and consequently that it is possible to observe an inhibition of *top* by *fis*. This inhibition effect is actually observed for another kind of stress. In [10] it is said: "when *Fis* levels are low, hydrogen peroxide treatment results in *topA* activation". This means that *fis* acts in some cellular contexts as an inhibitor of *top*. This paper shows that the protein *fis* can indeed play an inhibitory role on *top* in some contexts, and it thus gives support to the new hypothesis that *fis* plays an inhibitory role in the response to nutritional stress. It is remarkable that this pool of constraints is inconsistent given that the number of adjustable parameters is relatively high. We insist here on the fact that inspection of the constraint pool did not allow to resolve manually this inconsistency.

Finally, it appears that the hypotheses of interaction compositions on *top* are not well supported by experiments, and we propose to determine the necessarily observable compositions of the type i_c^r and $(\neg i_c^r)$. The rationale for limiting the compositions to basic types (signed interactions) is to provide easily interpretable results in terms of the interaction graph \mathcal{G} complemented with interaction signs (corresponding to the choice between activation and inhibition). This allows to determine for example whether there are unnecessary arcs in \mathcal{G} . On the other hand this restriction still allows to guide the user in the choice of hypotheses about interaction compositions. We conserve all the previous hypotheses except the ones about the interaction compositions on *top*. We consider a new set of these six interaction compositions for *top*: $ic_{top}^1 = i_{top}^1$, $ic_{top}^2 = i_{top}^2$, $ic_{top}^3 = i_{top}^3$, $ic_{top}^4 = \neg i_{top}^1$, $ic_{top}^5 = \neg i_{top}^2$, $ic_{top}^6 = \neg i_{top}^3$. Finally we challenge the observability constraints onto them to find which of them are necessary for these hypotheses. GNBox returns the property $a(ic_{top}^2) \vee o(ic_{top}^4)$ (observability property of (gyr, t_{gyr}^2, top) or observability property of $\neg(fis, t_{fis}^3, top)$)

in 4 seconds with a length of path $L = 10$, and the same formula in 60 seconds with a length of path $L = 100$. This indicates that any solution of all constraints (except the hypotheses of composition on *top*) has the property $o(ic_{top}^2)$ or the property $o(ic_{top}^4)$. This result provides an essential information to help the biologist to make additional hypotheses about interaction compositions.

Application to the gap-gene module of the segmentation of the *Drosophila melanogaster* embryo

In the first hours after fertilization, the embryo of the fly *Drosophila melanogaster* undergoes segmentation along the anteroposterior axis (head to tail). The embryo is partitioned into segments, each segment being made of cells characterized by specific levels of a set of proteins. Segmentation takes place in several successive stages controlled by distinct genetic modules. Here we focus on the gap-gene regulatory module.

The modeling and analysis of this network illustrates the expression of steady states in several segments of the embryo for the wild type and several mutant types, and the search for the minimum number of thresholds necessary to account for all the observations. The initial model is adapted from [11,12]. Although this model is not the most recent available, it is convenient for our purpose. The resolution of this query provides a set of minimal models (in terms of number of thresholds) consistent with a set of very diverse observations. The connection between the observations for all these models (one for each mutant type and for each segment) adds a new level of complexity.

The model, proposed in [11,12], controlling the gap-gene module contains seven genes: Giant denoted by *gt*, Hunchback zygotic denoted by *hb_z*, Hunchback maternal denoted by *hb_m*, Krüppel denoted by *kr*, Knirps denoted by *kni*, Bicoid denoted by *bcd*, and Caudal denoted by *cad*. The genes *bcd*, *hb_m* and *cad* are input genes: they influence other genes but are not influenced by any gene. Stocks of maternal mRNA and proteins are accumulated at specific places of the egg before fertilization. These molecules generate gradients along the anteroposterior axis. In the model these quantities are represented by input parameters (one for each chemical species and each region). The interaction graph \mathcal{G} between these genes is given in Figure 8 where input genes are represented by dotted circles filled in blue. Interactions and interaction compositions are given in Table 3. Moreover we consider the set of all additivity and observability constraints for all interaction compositions. The modeling in [11,12] takes into account four adjacent segments along the anteroposterior axis, denoted by *A*, *B*, *C* and *D*. Genetic experiments produced information on the concentration of the gap-gene proteins for the wild type, denoted by *wt*, and nine

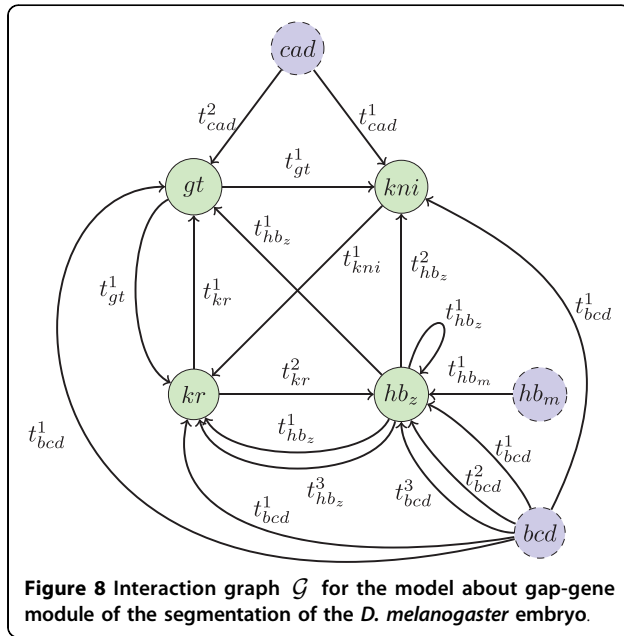


Table 3 Interactions and interaction compositions hypotheses for the model about gap-gene module of the segmentation of the *D. melanogaster* embryo

species	Interactions	interaction compositions
<i>gt</i>	$i_{gt}^1 = (hb_z, t_{hb_z}^1, gt)$ $i_{gt}^2 = (kr, t_{kr}^1, gt)$ $i_{gt}^3 = (bcd, t_{bcd}^1, gt)$ $i_{gt}^4 = (cad, t_{cad}^2, gt)$	$ic_{gt}^1 = \neg i_{gt}^1$ $ic_{gt}^2 = \neg i_{gt}^2$ $ic_{gt}^3 = i_{gt}^3$ $ic_{gt}^4 = i_{gt}^4$
<i>hb_z</i>	$i_{hb_z}^1 = (hb_z, t_{hb_z}^1, hb_z)$ $i_{hb_z}^2 = (kr, t_{kr}^2, hb_z)$ $i_{hb_z}^3 = (bcd, t_{bcd}^1, hb_z)$ $i_{hb_z}^4 = (bcd, t_{bcd}^2, hb_z)$ $i_{hb_z}^5 = (bcd, t_{bcd}^3, hb_z)$ $i_{hb_z}^6 = (hb_m, t_{hb_m}^1, hb_z)$	$ic_{hb_z}^1 = \neg i_{hb_z}^2$ $ic_{hb_z}^2 = i_{hb_z}^3$ $ic_{hb_z}^3 = i_{hb_z}^3 \wedge (i_{hb_z}^1 \vee i_{hb_z}^6)$ $ic_{hb_z}^4 = i_{hb_z}^4 \wedge (i_{hb_z}^1 \vee i_{hb_z}^6)$ $ic_{hb_z}^5 = i_{hb_z}^5 \wedge (i_{hb_z}^1 \vee i_{hb_z}^6)$
<i>kr</i>	$i_{kr}^1 = (gt, t_{gt}^1, kr)$ $i_{kr}^2 = (hb_z, t_{hb_z}^1, kr)$ $i_{kr}^3 = (hb_z, t_{hb_z}^3, kr)$ $i_{kr}^4 = (kni, t_{kni}^1, kr)$ $i_{kr}^5 = (bcd, t_{bcd}^1, kr)$	$ic_{kr}^1 = \neg i_{kr}^1$ $ic_{kr}^2 = i_{kr}^2 \wedge \neg i_{kr}^3$ $ic_{kr}^3 = \neg i_{kr}^4$ $ic_{kr}^4 = i_{kr}^5$
<i>kni</i>	$i_{kni}^1 = (gt, t_{gt}^1, kni)$ $i_{kni}^2 = (hb_z, t_{hb_z}^2, kni)$ $i_{kni}^3 = (bcd, t_{bcd}^1, kni)$ $i_{kni}^4 = (cad, t_{cad}^1, kni)$	$ic_{kni}^1 = \neg i_{kni}^1$ $ic_{kni}^2 = \neg i_{kni}^2$ $ic_{kni}^3 = i_{kni}^3$ $ic_{kni}^4 = i_{kni}^4$

mutants. The mutants are: knock-out (KO) on *gt* (the focal value of the *gt* component is 0 everywhere) denoted by *gt0*, KO on both *hb_z* and *hb_m* denoted by *hb0*, KO on *kr* denoted by *kr0*, KO on *kni* denoted by *kni0*, KO on *bcd* denoted by *bcd0*, KO on *hb_m* denoted by *hbm0*, KO on *cad* denoted by *cad0*, ectopic expression equal to 1 on *gt* (the focal value of the *gt* component is everywhere equal to 1) denoted by *gt1*, ectopic expression equal to 1 on *kni* denoted by *kni1*. We define a model $M^{R,T}$ for each segment $R \in \{A, B, C, D\}$ and each type $T \in \{wt, gt0, hb0, kr0, kni0, bcd0, cad0, hbm0, gt1, kni1\}$. For example, $M^{B,gt0}$ corresponds to the model of the mutant type *gt0* in segment *B*. The input parameters, discrete kinetic parameters and threshold parameters, between models are linked by introducing equality constraints between them, as explained in the section on the formalization of Thomas networks. Thus it would be redundant to impose constraints about interaction compositions for mutant types (the corresponding constraints for the wild type are sufficient). Obviously, these constraints lead to exactly the same threshold and discrete kinetic parameters for the four models associated to the four segments and each mutant.

The concentrations of the proteins produced by input genes *bcd*, *hb_m* and *cad* for each region *R* and each mutant type *T* are respectively denoted by $M_{bcd}^{R,T}$, $M_{hb_m}^{R,T}$ and $M_{cad}^{R,T}$. We impose in Constraint 4 that the inputs in the mutant types are equal to those of the wild type, except in the cases where some input genes themselves are mutated. This exception is due to the fact that the inputs come from the mother system, so only a mutation in this system can change the concentration value of the corresponding input.

Constraint 4

$$\bigwedge_R \bigwedge_{T \in \{gt0, kr0, kni0, gt1, kni1\}} M_{bcd}^{R,T} = M_{bcd}^{R,wt} \quad \wedge$$

$$M_{hb_m}^{R,T} = M_{hb_m}^{R,wt} \quad \wedge$$

$$M_{cad}^{R,T} = M_{cad}^{R,wt}$$

Moreover, we have inequality constraints between the thresholds for this model:

Constraint 5

$$\bigwedge_{p1, p2} (p1 < p2 \Rightarrow t_c^{p1} \leq t_c^{p2})$$

The observations relate to the existence of one steady state by mutant type and by segment with some properties between these states. The steady states are

represented by ordered lists of four protein concentrations: $S^{R,T} = \langle S_{gt}^{R,T}, S_{hb_z}^{R,T}, S_{kr}^{R,T}, S_{kni}^{R,T} \rangle$ for each region R and each mutant type T . The constraint associated to the observation of the steady state of each region R and each type T is:

Constraint 6

$$\bigwedge_{R,T} S^{R,T} = \left\langle S_{gt}^{R,T}, S_{hb_z}^{R,T}, S_{kr}^{R,T}, S_{kni}^{R,T} \right\rangle \wedge \text{path}(M^{R,T}, [S^{R,T}, S^{R,T}], 2)$$

The gradients of maternal origin mentioned above are used by the cell to derive positional information. To represent these gradients, the antero-posterior axis is partitioned into segments, each segment being identified by a combination of values of the input molecules bcd , hb_m and cad . We impose that the combinations of input quantities are different for all pairs of segments for the wild type:

Constraint 7

$$\bigwedge_{(R1,R2) \in \{A,B,C,D\}^2 | R1 \neq R2} (M_{bcd}^{R1,wt}, M_{hb}^{R1,wt}, M_{cad}^{R1,wt}) \neq (M_{bcd}^{R2,wt}, M_{hb_m}^{R2,wt}, M_{cad}^{R2,wt})$$

Table 4 shows the constraints between the concentrations of the steady states in regions A , B , C and D for each species and for each mutant; the bold font indicates a change of constraint compared to the wild type. All these constraints come from the interpretation of data in [12]. The first column of the table gives the mutant type, the second the genes, the third, fourth, fifth and sixth represent the steady state of the region A , B , C and D , respectively. Inequality symbols appear between the columns labeled A , B , C , D . They indicate constraints between concentrations of steady states of adjacent segments. Finally the two last columns give other constraints involving concentrations in segments that are not necessarily adjacent and comments about the differences compared to the wild type.

We note in the following C_{gap} the set of constraints associated to the existence of the steady states of the 4 regions for each of the 10 types (composed of constraints defining and linking the models $M^{R,T}$, Constraint 4, Constraint 5, Constraint 6, Constraint 7, and constraints in Table 4).

If we add to C_{gap} the constraint represented in Table 5 about the instantiation of steady states for all types, input parameters for the wild type according to the second table of [12], and the constraint $\wedge_c \wedge_p t_c^p = p$ we obtain a consistency in 11 seconds.

It appears in the second figure of [11] that there is no auto-interaction onto hb_z . In fact, after discussion with D. Thieffry, a synergy between Hunchback and Bicoid on the activation of Hunchback has been reported, and hb_m and hb_z are the same species in distinct compartments. This explain the interaction compositions onto hb_z with indexes 3, 4 and 5. If we do not consider this auto-interaction in C_{gap} by replacing the last three interaction compositions onto hb_z by $ic_{hb_z}^3 = i_{hb_z}^3 \wedge i_{hb_z}^6$, $ic_{hb_z}^4 = i_{hb_z}^4 \wedge i_{hb_z}^6$ and $ic_{hb_z}^5 = i_{hb_z}^5 \wedge i_{hb_z}^6$, and we still consider the constraint in Table 5 and the constraint $\wedge_c \wedge_p t_c^p = p$, we obtain an inconsistency in 7 seconds.

So we consider in the following the proposed model with this auto-interaction onto hb_z (in order to have a similar model to those in [11,12]). Obviously, C_{gap} alone (without the constraint represented in Table 5 and the constraint $\wedge_c \wedge_p t_c^p = p$) is consistent according to the first query.

In previous applications the thresholds t_c^p are instantiated and equal to p , p taking a value between 1 and max_c . This implies that the concentrations of c can take values between 0 and max_c . Insofar as the subdivision of the concentration space is only speculative, it is interesting to ask what is the smallest number of distinct thresholds necessary to get a model consistent with the observations. The extreme case would be the satisfaction of all observations and hypotheses with one threshold per component, i.e. with a Boolean model. It appears that C_{gap} plus $\wedge_c \wedge_p t_c^p = 1$ is inconsistent in 2 seconds. We can check easily this inconsistency: the constraint $t_{hb_z}^1 = t_{hb_z}^3$ is inconsistent with the observability constraint of ic_{kr}^2 because no state S satisfies ic_{kr}^2 i.e. the condition $S_{hb_z} \geq t_{hb_z}^1 \wedge S_{hb_z} < t_{hb_z}^3$.

To identify the minimum number of different thresholds needed to satisfy all the observations and hypotheses, we must build a query using the method of relaxation of constraints. But in this case, the relaxation takes place on the number of thresholds in $1..max_c$ for each component c .

To summarize, we challenge the hypotheses about the number of thresholds for all components. From a constraint point of view, this problem is specified and implemented in the following way. We introduce Boolean variables $B_{j,c}$ equivalent to "the number of thresholds for c is less or equal to j ", j being an integer in the interval $1..max_c - 1$. So we get six Boolean variables in our case: B_{1,hb_z} and B_{2,hb_z} for hb_z , $B_{1,kr}$ for kr , $B_{1,bcd}$ and $B_{2,bcd}$ for bcd , $B_{1,cad}$ for cad . Then GNBox searches for models consistent with the set of constraints defined above such that the number of these Boolean variables which are true is maximum. GNBox finds in 22 seconds that the only Boolean variables to be false are B_{2,hb_z} and $B_{2,cad}$. This indicates that hb_z must have at least 2 different values of thresholds, and cad must have at least 2 different values of thresholds.

Table 4 Constraints between stationary states and between input parameters for each region and each mutant type for the model about gap-gene module of the segmentation of the *D. melanogaster* embryo

Type	species	A	B	C	D	supp. constr.	Comments			
wt	bcd	$M_{bcd}^{A,wt}$	\geq	$M_{bcd}^{B,wt}$	\geq	$M_{bcd}^{C,wt}$	\geq	$M_{bcd}^{D,wt} = 0$		
	hb _m	$M_{hb_m}^{A,wt}$	\geq	$M_{hb_m}^{B,wt}$	\geq	$M_{hb_m}^{C,wt}$	\geq	$M_{hb_m}^{D,wt} = 0$		
	cad	$M_{cad}^{A,wt}$	\leq	$M_{cad}^{B,wt}$	\leq	$M_{cad}^{C,wt}$	\leq	$M_{cad}^{D,wt}$	$M_{cad}^{A,wt} = 0$	
	gt	$S_{gt}^{A,wt}$	$>$	$S_{gt}^{B,wt}$		$S_{gt}^{C,wt}$	$<$	$S_{gt}^{D,wt}$	$S_{hb_z}^{B,wt} > S_{hb_z}^{D,wt}$	
	hb _z	$S_{hb_z}^{A,wt}$	\geq	$S_{hb_z}^{B,wt}$	\geq	$S_{hb_z}^{C,wt}$	\geq	$S_{hb_z}^{D,wt}$		
	kr	$S_{kr}^{A,wt}$	$<$	$S_{kr}^{B,wt}$	\geq	$S_{kr}^{C,wt}$	\geq	$S_{kr}^{D,wt}$		
	kni	$S_{kni}^{A,wt}$	\leq	$S_{kni}^{B,wt}$	$<$	$S_{kni}^{C,wt}$	$>$	$S_{kni}^{D,wt}$		
gt0	gt	$S_{hb}^{A,gt0}$	\geq	$S_{hb}^{B,gt0}$	\geq	$S_{hb}^{C,gt0}$	\geq	$S_{hb}^{D,gt0}$	$S_{hb_z}^{B,gt0} > S_{hb_z}^{D,gt0}$	Knock-out
	hb _z	$S_{kr}^{A,gt0}$	$<$	$S_{kr}^{B,gt0}$	\geq	$S_{kr}^{C,gt0}$	\geq	$S_{kr}^{D,gt0}$		
	Kr	$S_{kni}^{A,gt0}$	\leq	$S_{kni}^{B,gt0}$	$<$	$S_{kni}^{C,gt0}$	\leq	$S_{kni}^{D,gt0}$		Kni expands into D
hb0	bcd	$M_{bcd}^{A,hb0}$	\geq	$M_{bcd}^{B,hb0}$	\geq	$M_{bcd}^{C,hb0}$	\geq	$M_{bcd}^{D,hb0}$		Knock-out
	hb _m									
	cad	$M_{cad}^{A,hb0}$	\leq	$M_{cad}^{B,hb0}$	\leq	$M_{cad}^{C,hb0}$	\leq	$M_{cad}^{D,hb0}$		
	gt								$\bigwedge_R S_{gt}^{R,hb0} \geq 1$	gt expand into BC
	hb _z									Knock-out
	kr								$\bigwedge_R S_{kr}^{R,hb0} = 0$	loss of kr into BC
	kni								$\bigwedge_R S_{kni}^{R,hb0} = 0$	loss of kni into BC
kr0	gt								$\bigwedge_R S_{gt}^{R,kr0} \geq 1$	gt expands into BC
	hb _z	$S_{hb_z}^{A,kr0}$	\geq	$S_{hb_z}^{B,kr0}$	\geq	$S_{hb_z}^{C,kr0}$	\geq	$S_{hb_z}^{D,kr0}$	$S_{hb_z}^{B,kr0} > S_{hb_z}^{D,kr0}$	
	kr									Knock-out
	kni								$\bigwedge_R S_{kni}^{R,kr0} = 0$	loss of kni into BC
kni0	gt	$S_{gt}^{A,kni0}$		$S_{gt}^{B,kni0}$		$S_{gt}^{C,kni0}$	$<$	$S_{gt}^{D,kni0}$		
	hb _z	$S_{hb_z}^{A,kni0}$	$>$	$S_{hb_z}^{B,kni0}$	\geq	$S_{hb_z}^{C,kni0}$	\geq	$S_{hb_z}^{D,kni0}$	$S_{hb_z}^{B,kni0} > S_{hb_z}^{C,kni0}$	
	kr	$S_{kr}^{A,kni0}$	\geq	$S_{kr}^{B,kni0}$	\geq	$S_{kr}^{C,kni0}$	\geq	$S_{kr}^{D,kni0}$	$S_{kr}^{C,kni0} \geq 1$	
	Kni		$<$							increase of kr into c knock-out
bcd0	bcd									Knock-out
	hb _m	$M_{hb_m}^{A,bcd0}$	\geq	$M_{hb_m}^{B,bcd0}$	\geq	$M_{hb_m}^{C,bcd0}$	\geq	$M_{hb_m}^{D,bcd0}$		
	cad	$M_{cad}^{A,bcd0}$	\leq	$M_{cad}^{B,bcd0}$	\leq	$M_{cad}^{C,bcd0}$	\leq	$M_{cad}^{D,bcd0}$		
	gt					$S_{gt}^{C,bcd0}$	$<$	$S_{gt}^{D,bcd0}$	$S_{gt}^{A,bcd0} = 0$	loss of gt into A
	hb _z								$\bigwedge_R S_{hb_z}^{R,bcd0} = 0$	loss of hb _z into ABC
	kr								$\bigwedge_R S_{kr}^{R,bcd0} = 0$	loss of kr into BC
	kni	$S_{kni}^{A,bcd0}$	\geq	$S_{kni}^{B,bcd0}$	\geq	$S_{kni}^{C,bcd0}$	$>$	$S_{kni}^{D,bcd0}$		kni expands into AB

Table 4 Constraints between stationary states and between input parameters for each region and each mutant type for the model about gap-gene module of the segmentation of the *D. melanogaster* embryo (Continued)

<i>hbm0</i>	<i>bcd</i>	$M_{bcd}^{A,hbm0}$	\geq	$M_{bcd}^{B,hbm0}$	\geq	$M_{bcd}^{C,hbm0}$	\geq	$M_{bcd}^{D,hbm0}$		
	<i>hb_m</i>									Knock-out
	<i>cad</i>	$M_{cad}^{A,hbm0}$	\leq	$M_{cad}^{B,hbm0}$	\leq	$M_{cad}^{C,hbm0}$	\leq	$M_{cad}^{D,hbm0}$		
	<i>gt</i>	$S_{gt}^{A,hbm0}$	$>$	$S_{gt}^{B,hbm0}$		$S_{gt}^{C,hbm0}$	$<$	$S_{gt}^{D,hbm0}$		
	<i>hb_z</i>	$S_{hb_z}^{A,hbm0}$	\geq	$S_{hb_z}^{B,hbm0}$	\geq	$S_{hb_z}^{C,hbm0}$	\geq	$S_{hb_z}^{D,hbm0}$	$S_{hb_z}^{B,hbm0} > S_{hb_z}^{D,hbm0}$	
	<i>kr</i>	$S_{kr}^{A,hbm0}$	$<$	$S_{kr}^{B,hbm0}$	\geq	$S_{kr}^{C,hbm0}$	\geq	$S_{kr}^{D,hbm0}$		
	<i>kni</i>	$S_{kni}^{A,hbm0}$	\geq	$S_{kni}^{B,hbm0}$	$<$	$S_{kni}^{C,hbm0}$	$>$	$S_{kni}^{D,hbm0}$		
<i>cad0</i>	<i>bcd</i>	$M_{bcd}^{A,cad0}$	\geq	$M_{bcd}^{B,cad0}$	\geq	$M_{bcd}^{C,cad0}$	\geq	$M_{bcd}^{D,cad0}$		
	<i>hb_m</i>	$M_{hb_m}^{A,cad0}$	\geq	$M_{hb_m}^{B,cad0}$	\geq	$M_{hb_m}^{C,cad0}$	\geq	$M_{hb_m}^{D,cad0}$		
	<i>cad</i>									
	<i>gt</i>	$S_{gt}^{A,cad0}$	$>$	$S_{gt}^{B,cad0}$				$S_{gt}^{D,cad0} = 0$		knock-out lass of <i>gt</i> into <i>D</i>
	<i>hb_z</i>	$S_{hb_z}^{A,cad0}$	\geq	$S_{hb_z}^{B,cad0}$	\geq	$S_{hb_z}^{C,cad0}$	\geq	$S_{hb_z}^{D,cad0}$	$S_{hb_z}^{B,cad0} > S_{hb_z}^{D,cad0}$	
	<i>kr</i>	$S_{kr}^{A,cad0}$	$<$	$S_{kr}^{B,cad0}$	\geq	$S_{kr}^{C,cad0}$	\geq	$S_{kr}^{D,cad0}$	$S_{kr}^{C,cad0} \geq 1$	increase of <i>kr</i> of into <i>C</i>
	<i>kni</i>							$\bigwedge_R S_{kni}^{R,cad0} = 0$		lass of <i>kni</i> into <i>C</i>
<i>gt1</i>	<i>gt</i>									ectopic expression
	<i>hb_z</i>	$S_{hb_z}^{A,gt1}$	\geq	$S_{hb_z}^{B,gt1}$	\geq	$S_{hb_z}^{C,gt1}$	\geq	$S_{hb_z}^{D,gt1}$	$S_{hb_z}^{B,gt1} > S_{hb_z}^{D,gt1}$	
	<i>kr</i>	$S_{kr}^{A,gt1}$	$<$	$S_{kr}^{B,gt1}$	\geq	$S_{kr}^{C,gt1}$	\geq	$S_{kr}^{D,gt1}$		
	<i>kni</i>							$\bigwedge_R S_{kni}^{R,gt1} = 0$		lass of <i>kni</i> into <i>C</i>
<i>Kni1</i>	<i>gt</i>	$S_{gt}^{A,kni1}$	\geq	$S_{gt}^{B,kni1}$	$>$	$S_{gt}^{C,kni1}$	$<$	$S_{gt}^{D,kni1}$		activation of <i>gt</i> into <i>B</i>
	<i>hb_z</i>	$S_{hb_z}^{A,kni1}$	\geq	$S_{hb_z}^{B,kni1}$	\geq	$S_{hb_z}^{C,kni1}$	\geq	$S_{hb_z}^{D,kni1}$		
	<i>Kr</i>					$S_{kr}^{C,kni1}$	\geq	$S_{kr}^{D,kni1}$	$S_{hb_z}^{B,kni1} = S_{hb_z}^{D,kni1}$	
	<i>kni</i>							$S_{kr}^{A,kni1} = S_{kr}^{B,kni1} = 0$		lass of <i>kr</i> into <i>B</i> ectopic expression

The last query gives, in 17 seconds, two possible instantiations of the t_C^p accepting C_{gap} and the minimal number of thresholds given by the previous query:

$$t_{bcd}^1 = 1, t_{bcd}^2 = 1, t_{bcd}^3 = 1, t_{hb_m}^1 = 1, t_{cad}^1 = 1, t_{cad}^2 = 2, t_{gt}^1 = 1, t_{hb_z}^1 = 1, t_{hb_z}^2 = 1 \text{ or } 2, t_{hb_z}^3 = 2, t_{kr}^1 = 1, t_{kr}^2 = 1, t_{kni}^1 = 1.$$

One remarks that the three thresholds of *hb_z* share only two values.

Conclusions

Our methodology is composed of two parts: (i) a declarative constraint-based approach; (ii) a formalism for the description of the dynamics of discrete networks. We have presented here applications involving gene regulatory networks whose behaviour is satisfactorily represented in the formalism of R. Thomas. But it is important to note that the methodology can be applied

to many other types of dynamical rules, such as Hopfield-like networks, Boolean networks with parallel, sequential or block-sequential updating. The only requirement is that the dynamical rules should be expressed as constraints on finite-domain variables. The potential domain of application of this methodology is thus much larger than just gene regulatory networks.

The Thomas' networks have largely been applied to the analysis of GRNs, for example those described in [7,13-15] or those described in [8,16,17] which use a very similar qualitative formalism.

Several modeling and simulation tools of biological regulatory networks (for example GINsim [18], BIOCHAM [19], GNA [20]) are used in combination with model checkers (NuSMV, CADP) and based on diverse

Table 5 Constraints of instantiation of stationary states according to the second table in [12] for each region and each mutant type for the model about gap-gene module of the segmentation of the *D. melanogaster* embryo

type	species	A	B	C	D
wt	bcd	3	2	1	0
	hb _m	1	1	0	0
	cad	0	0	1	2
	gt	1	0	0	1
	hb _z	3	2	1	0
	kr	0	2	1	0
	kni	0	0	1	0
gt0	gt	0	0	0	0
	hb _z	3	2	1	0
	kr	0	2	1	0
	kni	0	0	1	1
hb0	gt	1	1	1	1
	hb _z	0	0	0	0
	kr	0	0	0	0
	kni	0	0	0	0
kr0	gt	1	1	1	1
	hb _z	3	2	1	0
	Kr	0	0	0	0
	Kni	0	0	0	0
kni0	gt	1	0	0	1
	hb _z	3	2	1	0
	kr	0	2	2	0
	kni	0	0	0	0
bcd0	gt	0	0	0	1
	hb _z	0	0	0	0
	kr	0	0	0	0
	kni	1	1	1	0
hbm0	gt	1	0	0	1
	hb _z	3	2	1	0
	kr	0	2	1	0
	kni	0	0	1	0
cad0	gt	1	0	0	0
	hb _z	3	2	1	0
	kr	0	2	2	0
	kni	0	0	0	0
gt1	gt	1	1	1	1
	hb _z	3	2	1	0
	kr	0	1	1	0
	kni	1	1	1	1
kni1	gt	1	1	0	1
	hb _z	3	2	1	0
	kr	0	0	1	0
	kni	1	1	1	1

formalisms (logic, Petri nets, ODEs). The idea is to add to the simulation functionality a formal verification functionality to check, or optimize [21], the fitness between the simulated and the observed behaviours.

Three types of abstractions are available in BIOC-HAM, among which ordinary differential equations and Boolean networks. The inference of parameters is based on the technique of model-checking and the definition of a continuous degree of satisfaction of a temporal logic formula formalizing some observation on behaviour. This permits to find biochemical kinetic parameter values which are optimal with respect to a set of biological properties. Moreover it is possible to find the effect of parameter variations on the robustness of a behaviour specification [22]. Our work differs significantly in that it focuses to face the problem of incomplete knowledge to produce, by a constraint-based process, a class of models from which it is expected to design new experiments.

A steady state search module, including the so-called singular states, exists in GNA based on an integration of the SAT solver SAT4J [23]. This integration of a constraint approach avoids the generation of all the transitions to identify the steady states. But in contrast to our work aiming at providing general queries, [23] focus on the search of steady states, and only in the case of completely instantiated models (kinetic parameters instantiated and order of the thresholds predefined). The work in [24] search the same steady states with a CSP (Constraint Satisfaction Problem) formalization, the performances are worse than in [23]. In our work, we can write easily queries to identify steady states, and even cycles of length smaller than some predefined value. In addition in our case the kinetic parameters and the orders between thresholds can be only partially known. As explained here, other much more sophisticated queries are available, although in the current version we do not include singular states.

The formal approach proposed here modifies deeply the way to proceed in the building and in the exploration of genetic and biochemical networks, first by avoiding the usual trial-and-error procedure, and second by putting the emphasis on sets of solutions, rather than a single consistent solution arbitrarily chosen in a set. Last, the constraint approach lends to a unified description of network architecture and network behaviour, as both are described in terms on formal constraints. The knowledge available to initiate the modeling of a given phenomenon is generally sparse with respect to the complexity of the behaviour of the underlying networks.

It is thus essential to exploit consistently, efficiently, and in a joint manner, every bit of experimental information. The representation of knowledge in terms of constraints is a way to achieve, at least to some extent, this goal.

Our environment GNBox implements a wide panel of functionalities: simulations, consistency proof, relaxation in case of inconsistency, search for a minimal model, prediction of properties in case of consistency. This last functionality generates properties which are verified by all solutions of the constraint pool. In line with what we said above, note that such properties are really supported by data. This contrasts with the usual practice of using just one solution to make prediction, neglecting the existence of other solutions. Properties of the selected single model should not be considered as true predictions.

We have presented three biological applications illustrating the use of most of these functionalities. These applications involve networks containing about 5 species and 15 possible interactions, and with set of hypotheses and observations without systematic instantiation of threshold parameters, with a large range of types of behaviours. In the third application the queries involve several structurally related models in order to incorporate knowledge about wild-type and mutant behaviour, in four segments of the embryo. The set of constraints generates a dense network of dependencies between the variables. The performances of GNBox are good for the different types of queries presented in the three applications. The most computer-intensive queries are those involving paths. For such queries our approach is limited to networks of medium size.

The perspectives are governed by the biological problems. The methodologies and technologies employed must be chosen according to these problems. A first perspective is to prioritize biological experiments. For example, consider a situation in which the state of knowledge is such that the number of consistent instantiated models is still large, and it is possible to perform double knock-out experiments. In such situation it would be interesting to be able to determine the most informative choice of pairs of genes to target for knock-out, an informative experiment being one which will potentially add non redundant constraints and thus reduce the set of solutions. Another perspective is to refine the abstraction of the discrete behaviours: for example by taking into account the trajectories sliding along the thresholds [17,25], and taking into account the difference of delays of chemical reactions [26-28]. Another perspective is the extended repairing consistency techniques adding species, related to the problem of composing networks [29].

Additional material

Additional file 1: GNBox.

Acknowledgements

This work was performed at the TIMC-IMAG laboratory. Final editing of the manuscript was made while FC was at IRISA-INRIA. The work was supported by Microsoft Research through its PhD Scholarship Programme with a grant to FC. EF thanks IXXI (Institut des Systèmes Complexes Rhône-Alpes) for partial financial support. We thank Denis Thieffry for helpful clarifications on the *drosophila* embryo segmentation model.

Author details

¹Laboratoire TIMC-IMAG, UMR CNRS/UJF 5525, Domaine de la Merci, 38710 La Tronche, France. ²Laboratoire IRISA-INRIA centre de Rennes, Campus de Beaulieu, 35042 Rennes, France.

Authors' contributions

The methods and applications was mainly developed by FC on theoretical foundation and ideas provided by EF and LT. All authors equally wrote this manuscript and approved it.

Received: 22 January 2010 Accepted: 20 July 2010

Published: 20 July 2010

References

1. Thomas R, D'Ari R: *Biological Feedback* CRC Press 1990.
2. Thomas R, Kaufman M: **Multistationarity, the Basis of Cell Differentiation and Memory. II. Logical Analysis of Regulatory Networks in Term of Feedback Circuits.** *Chaos* 2001, **11**:180-195.
3. Corblin F, Bordeaux L, Fanchon E, Hamadi Y, Trilling L: **Connections and Integration with SAT Solvers: A Survey and a Case Study in Computational Biology.** *Hybrid Optimization: the 10 years of CPAIOR*, Springer 2010.
4. Kauffman S: **Metabolic stability and epigenesis in randomly constructed genetic nets.** *Journal of Theoretical Biology* 1969, **22**:437-467.
5. Demongeot J, Elena A, Sené S: **Robustness in regulatory networks: a multi-disciplinary approach.** *Acta Biotheoretica* 2008, **56**:27-49.
6. Mendoza L, Alvarez-Buylla E: **Dynamics of the Genetic Regulatory Network for Arabidopsis thaliana Flower Morphogenesis.** *Journal of Theoretical Biology* 1998, **193**(2):307-319.
7. Thieffry D, Thomas R: **Dynamical Behaviour of Biological Regulatory Networks -II. Immunity Control in Bacteriophage Lambda.** *Bulletin of Mathematical Biology* 1995, **57**:277-297.
8. Ropers D, de Jong H, Page M, Schneider D, Geiselmann J: **Qualitative simulation of the carbon starvation response in Escherichia coli.** *Biosystems* 2006, **84**(2):124-152.
9. Corblin F, Tripodi S, Fanchon E, Ropers D, Trilling L: **A declarative constraint-based method for analyzing discrete gene regulation networks.** *Biosystems* 2009, **98**(2):91-104.
10. Weinstein-Fischer D, Altuvia S: **Differential regulation of Escherichia coli topoisomerase I by Fis.** *Mol Microbiol* 2007, **63**(4):1131-1144.
11. Sánchez L, Thieffry D: **A Logical Analysis of the Drosophila Gap-gene System.** *J theor Biol* 2001, **211**:115-141.
12. Thieffry D, Sánchez L: **Alternative epigenetic states understood in terms of specific regulatory structures.** *Annals of the New York Academy of Sciences* 2002, **981**:135-153.
13. Sánchez L, van Helden J, Thieffry D: **Establishment of the dorso-ventral pattern during embryonic development of Drosophila melanogaster: a logical analysis.** *J Theor Biol* 1997, **187**:377-389.
14. Fanchon E, Corblin F, Trilling L, Hermant B, Gulino D: **Modeling the Molecular Network Controlling Adhesion Between Human Endothelial Cells: Inference and Simulation Using Constraint Logic Programming.** *LNCS, Computational Methods in Systems Biology* Danos V, Schachter V 2005, **3082**:104-118.

15. Guespin-Michel J, Bernot G, Comet JP, Mérieau A, Richard A, Hulen C, Polack B: **Epigenesis and dynamic similarity in two regulatory networks in *Pseudomonas aeruginosa*.** *Acta Biotheoretica* 2004, **52**(4):379-390.
16. de Jong H, Geiselmann J, Batt G, Hernandez C, Page M: **Qualitative Simulation of the initiation of sporulation in *Bacillus subtilis*.** *Bulletin of Mathematical Biology* 2004, **66**(2):261-299.
17. de Jong H, Gouzé JL, Hernandez C, Page M, Sari T, Geiselmann J: **Qualitative Simulation of Genetic Regulatory Networks Using Piecewise-Linear Models.** *Bulletin of Mathematical Biology* 2004, **66**(2):301-340.
18. Naldi A, Berenguier D, Fauré A, Lopez F, Thieffry D, Chaouiya C: **Logical modelling of regulatory networks with GINsim 2.3.** *Biosystems* 2009, **97**(2):134-139.
19. Calzone L, Fages F, Soliman S: **BIOCHAM: an environment for modeling biological systems and formalizing experimental knowledge.** *Bioinformatics* 2006, **22**(14):1805-1807.
20. de Jong H, Geiselmann J, Hernandez C, Page M: **Genetic Network Analyser: qualitative simulation of genetic regulatory networks.** *Bioinformatics* 2003, **19**(3):336-344.
21. Rizk A, Batt G, Fages F, Soliman S: **On a continuous degree of satisfaction of temporal logic formulae with applications to systems biology.** *Proc. of the Fourth International Conference on Computational Methods in Systems Biology (CMSB'08)*, LNCS SpringerHeiner M, Uhrmacher A 2008, **5307**:251-268.
22. Rizk A, Batt G, Fages F, Soliman S: **A general computational method for robustness analysis with applications to synthetic gene networks.** *Bioinformatics* 2009, **25**(12):169-178.
23. de Jong H, Page M: **Search for steady states of piecewise-linear differential equation models of genetic regulatory networks.** *IEEE/ACM Trans. Comput Biol Bioinform* 2008, **5**(02):208-222.
24. Devloo V, Hansen P, Labbé M: **Identification of All Steady States in Large Biological Systems by Logical Analysis.** *Bulletin of Mathematical Biology* 2003, **65** (6):1025-1051.
25. Corblin F, Fanchon E, Trilling L: **Modélisation de réseaux biologiques discrets en programmation logique par contraintes.** *TSI (Technique et Science Informatiques)* 2007, **26**:73-98.
26. Siebert H, Bockmayr A: **Temporal constraints in the logical analysis of regulatory networks.** *Theoretical Computer Science* 2008, **391**(3):258-275.
27. Ahmad J, Bernot G, Comet JP, Lime D, Roux O: **Hybrid modelling and dynamical analysis of gene regulatory networks with delays.** *ComPlexUs* 2006, **3**(4).
28. Batt G, Salah R, Maler O: **On Timed Models of Gene Networks.** *LNCS* 2007, **4763**:38-52.
29. Gössler G: **Compositional Reachability Analysis of Genetic Networks.** *Computational Methods in Systems Biology LNCS*, Springer Berlin/Heidelberg 2006, **4210**:212-226.

doi:10.1186/1471-2105-11-385

Cite this article as: Corblin et al.: Applications of a formal approach to decipher discrete genetic networks. *BMC Bioinformatics* 2010 **11**:385.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

