Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Theoretical Biology 229 (2004) 339–347
Journal of Theoretical Biology
www.elsevier.com/locate/jtbi

Application of formal methods to biological regulatory networks: extending Thomas' asynchronous logical approach with temporal logic

Gilles Bernot[a], Jean-Paul Comet[a,*], Adrien Richard[a], Janine Guespin[b]

[a] La.M.I., Laboratoire de Méthodes Informatiques, UMR 8042, CNRS & Université d'Évry, Boulevard François Mitterrand, 91025 Évry Cedex, France
[b] LMDF, Laboratoire de Microbiologie du Froid, GR en Biologie Intégrative et Modélisation, Université de Rouen, 76821 Rouen Cedex, France

Received 9 September 2003; received in revised form 26 March 2004; accepted 6 April 2004

Abstract

Based on the discrete definition of biological regulatory networks developed by René Thomas, we provide a computer science formal approach to treat temporal properties of biological regulatory networks, expressed in computational tree logic. It is then possible to build all the models satisfying a set of given temporal properties. Our approach is illustrated with the mucus production in *Pseudomonas aeruginosa*. This application of formal methods from computer science to biological regulatory networks should open the way to many other fruitful applications.

Keywords: Biological regulatory networks; Formal methods; Temporal logic; Model checking

1. Introduction

To elucidate the principles that govern biological complexity, computer modeling has to overcome ad hoc explanations in order to make emerge novel and abstract concepts (Huang, 2001). Computational *systems biology* (Wolkenhauer, 2001) tries to establish methods and techniques that enable us to understand biological systems as architectural systems, including their robustness, design and manipulation (Kitano, 2002a; Hasty et al., 2002). It means to understand: the structure of the systems (e.g. gene, metabolic or signal transduction networks), the dynamics of such systems, methods to control them, design and modify systems in order to cope with desired properties (Kitano, 2002b).

The abstraction offered by biological regulatory networks place the discussion at a biological level instead of a biochemical level only, that allows one to cover behaviors at an upper level. Biological regulatory networks modelize interactions between biological entities (mostly genes and proteins). One can refer to de Jong (2002) to have an overview of different approaches for genetic networks (notice that the following additional references of Demongeot et al. (1995, 2000, 2003) are relevant too). The static part of these models can be represented by a graph: vertices abstract the biological entities, and edges abstract their interactions. An inhibition is labeled by "–" (negative action) and an activation is labeled by "+" (positive action), see Fig. 4. More importantly it is then possible to study dynamic aspects by associating at a given time to each vertex a numerical value which describes the concentration of the corresponding entity. Temporal evolutions of these values define the dynamics of the system. The first approaches used differential equation systems to represent the dynamics, and to face the combinatorial explosure of the parameters, Thomas introduced at the end of the 1970s a boolean approach for regulatory networks (the expression of an entity is on or off) to capture the qualitative nature of the dynamics. He proved its usefulness in the context of immunity in

*Corresponding author. Tel.: +33-160873250; fax: +33-160873789.
E-mail addresses: bernot@lami.univ-evry.fr (G. Bernot), comet@lami.univ-evry.fr (J.-P. Comet), arichard@lami.univ-evry.fr (A. Richard), Janine.Guespin@univ-rouen.fr (J. Guespin).

0022-5193/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2004.04.003

bacteriophages (Thomas et al., 1976; Thomas, 1978). Later on, he generalized it to multivalued levels of concentration (the so called multivalued logic or "generalized logical approach" (Thomas et al., 1995)). Moreover, the vertices of Thomas' regulatory networks are abstracted into "variables" allowing the cohabitation of heterogeneous informations (e.g. adding environmental variables to genetic ones). He also introduced so called "logical" parameters which describe the weight of the interactions. Kaufman and co-workers (1985) also introduced a modeling with time delays for taking into account the time between the order to begin the synthesis (gene passes a threshold) and its execution (gene product has evolved) and similarly for the degradation. A more recent reference on this subject is that of Thomas and Kaufman (2001a) where the authors show how time delays can be viewed as a refinement of the model.

Let us remark as it will become evident in the sequel of the article that the first key idea of Thomas is the asynchronous updating technique defining the dynamics, which makes the difference with more classical approaches from computer science such as Petri nets. This key idea allows one to capture the essence of a continuous dynamics with a discrete approach.

Discrete and differential modelings are coherent: The Thomas discrete multivalued approach has been built as a discretization of continuous differential equation systems (Snoussi, 1989), and has been confronted to the more classical analysis in terms of differential equations (Kaufman et al., 1985; Kaufman and Thomas, 1987). Taking into account "singular states", Thomas and Snoussi showed that all steady states can be found via the discrete approach (Snoussi and Thomas, 1993). More recently, Thomas and Kaufman have shown that the discrete description provides a qualitative fit of the differential equations with a small number of possible combinations of values for the parameters (Thomas and Kaufman, 2001b). The discrete approach is not required to study regulatory networks, however it is often convenient at least because biological data are rarely quantitative.

Circuits in the graph are indeed the key concepts: A direct or indirect influence of a gene on itself corresponds to a closed oriented path which constitutes a feedback circuit. If the gene tends to favor (resp. decrease) its own production via this circuit, the feedback circuit is said positive (resp. negative). A feedback circuit is positive (resp. negative) iff its number of negative arrows is even (resp. odd). Feedback circuits are fundamental because they decide the existence of steady states of the dynamics. It has been noticed (Thomas, 1980; Thomas et al., 1995; Demongeot et al., 2000; Cinquin and Demongeot, 2002) that at least one positive regulatory circuit is necessary to generate multistationarity whereas at least one negative circuit is necessary to obtain a homeostasy or a stable oscillatory behavior.

Some formal methods have been already introduced to revisit the discrete asynchronous approach of Thomas. In Devloo et al. (2003), constraint programming has been used to detect all steady states in large regulatory networks. In Pérès and Comet (2003), we described a very preliminary work on the application of model checking to biological regulatory networks.

In this article we run the machinery of formal methods from computer science to revisit Thomas' discrete multivalued approach. Formal methods impose detailed definitions which are introduced in the following sections. Section 2 defines biological regulatory graphs which describe the interactions between biological entities. Section 3 introduces the parameters which pilot the behaviors of the system. Section 4 defines the dynamics. Being in the domain of formal methods, we inherit from computer science the whole collection of validation and verification tools. Model checking tools are particularly suited as described in Section 5.

We take as concrete running example the mucus production in *Pseudomonas aeruginosa* (Guespin-Michel and Kaufman, 2001). These bacteria are commonly present in the environment and secrete mucus only in lungs affected by cystic fibrosis. As it increases the respiratory deficiency of the patient, it is the major cause of mortality. Bacteria isolated from cystic fibrosis' lungs continue to grow in laboratory as mucous colonies for numerous generations (mucoid phenotype). A majority of these bacteria present a mutation:

- *Does it mean that the mutation is the cause of the passage to the mucoid state*?

A majority of biologists tend to follow this hypothesis. However, the regulatory network which controls the mucus production has been elucidated (Fig. 1) and contains a positive feedback circuit. This makes possible a dynamic with two stationary states which would allow,

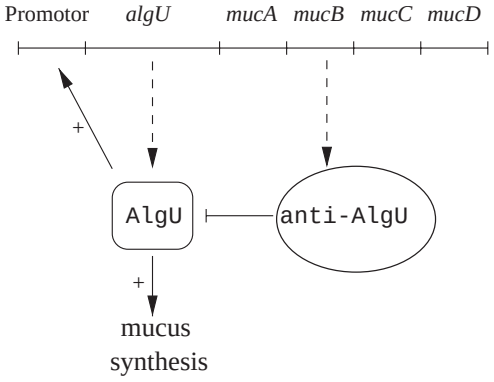

Fig. 1. Main regulatory genes of mucus production in *Pseudomonas aeruginosa*.

from a biological point of view, an epigenetic change (stable change of phenotype without mutation) from the non-mucoid state to the mucoid one.

- *Could the mucoidy be induced by an epigenetic phenomenon*?

  In this case the observed mutation (elimination of the anti-AlgU) could be favored later on because an inhibitor complex is produced, which is toxic for the bacteria.

Modeling can answer this last question which formally becomes: *is it possible to find at least one model of the wild bacteria, which is compatible with the known biological results and which has a multi-stationarity where one stable state regularly produces mucus and the other one does not*? In Section 5 model checking gives a positive answer. This epigenetic question has not only an academic interest because if this prediction is validated in vivo, it could lead to new therapeutic strategies.

We started from the following biological results. The main regulator of mucus production is the gene *algU*. It codes for the protein AlgU which positively regulates both an operon and all genes involved in the mucus synthesis. The operon is made up of 4 genes among which *algU* itself (positive autoregulation) and genes which code for an inhibitor complex (anti-AlgU) of the protein AlgU. This sketch of model (Fig. 1) is rather simple but it is sufficient to illustrate each definition introduced in the sequel and to show the interest of formal methods for biological systems.

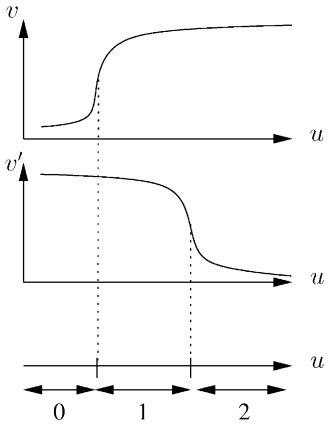

Fig. 2. Definition of the abstract concentration levels.

2. Biological regulatory graphs

Fig. 2 assumes that u is a variable (for example a gene product) which acts positively on v and negatively on v', each curve being the concentration of v (resp. v') with respect to the concentration of u, after a sufficient delay for u to act on v (resp. v'). Obviously, three regions are relevant in the different levels of concentration of u.

- In the first region u acts neither on v nor on v'.
- In the second region, u acts on v but it still does not act on v'.
- In the last region, u acts both on v and v'.

The sigmoid nature of the interactions shown in Fig. 2 is almost always verified and justifies this discretization of the concentration of u: three abstract levels (0, 1 and 2) emerge corresponding to the three previous regions and constitute the only relevant information from a qualitative point of view. More generally as shown in Fig. 3, if a variable acts on n variables, at most $n+1$ abstract regions are considered (from 0 to n, possibly less because two thresholds for two different target variables can be equal).

Now, to formally define a biological regulatory network we use a labeled directed graph. A vertex represents a variable (which can abstract a gene and its protein for instance) and has a boundary which is the maximal value of its discrete concentration level. A directed edge $(u \to v)$ is labeled with a threshold and a sign $+$ (resp. $-$) if u activates (resp. inhibits) v.

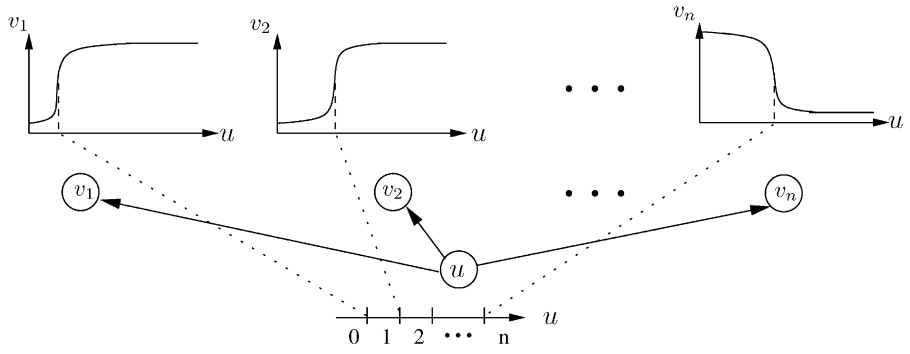

Fig. 3. Successors of u in the graph determine the abstract levels.

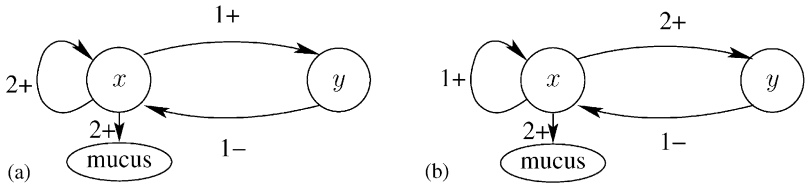

Fig. 4. Two possible regulatory graphs for mucus production in *Pseudomonas aeruginosa*.

Definition 1. A biological regulatory graph is a labeled directed graph $G = (V, E)$ where:

- each vertex v of V, called *variable*, is provided with a *boundary* $b_v \in \mathbb{N}^*$ less or equal to the out-degree of v in G, except when the out-degree is 0 where $b_v = 1$;
- each edge $(u \to v)$ of E is labeled with a couple (t, ϵ) where t, called *threshold*, is an integer between 1 and b_u and $\varepsilon \in \{-, +\}$.

The biological regulatory graph of Fig. 4, deduced from Fig. 1, contains two "classical" vertices x (which abstracts the gene *algU* and the protein `AlgU`) and y (which abstracts the gene *mucB* and the anti-`AlgU`) and an abstract "mucus" variable. The edges depict:

- $x \to x$: action of `AlgU` on its own promotor (auto induction),
- $y \to x$: inhibitor effect on `AlgU`,
- $x \to y$: positive influence of `AlgU` on its own inhibitors via the operon.

Since y acts on one variable, $b_y = 1$ and the threshold of edge $y \to x$ can only be 1. A question is to determine the thresholds of the other edges. It has been shown (Guespin-Michel and Kaufman, 2001) that mucus production occurs when x is over its second and last threshold ($x = 2$), but we do not know if x acts on y at a lower threshold than it acts on x or conversely. In other words, $b_x = 2$ and there are two possible biological regulatory graphs (Figs. 4(a) and (b)).

In Fig. 4, x belongs to a positive circuit: $x \to x$. It also belongs to a negative circuit via $x \to y \to x$, as well as y for the same reason. These two circuits compete on x (multi-stationarity against homeostasy).

3. Biological regulatory networks

In Figs. 2 and 3 we introduced the abstract concentration levels on the "horizontal" u axis. It remains to consider the "vertical" v axes: assuming that $u_1 \cdots u_n$ have an influence on v (entering arrows $u_i \to v$), toward which concentration level is v attracted? the set of all possible regulators of a variable being simply the set of its predecessors in the graph.

Notation. Let G be a biological regulatory graph and v be a variable, $G^{-1}(v)$ is the set of all predecessors of v, i.e. the set of variables u such that $(u \to v)$ is an edge of G.

But regulatory variables are not always active. At a given time, only some of them pass the threshold. Thus a variable v can be regulated by different subsets ω of its inhibitors/activators and we denote by $k_{v,\omega}$ the concentration level toward which v is attracted (Fig. 5). Biological regulatory networks are biological regulatory graph (Definition 1) together with those parameters $k_{v,\omega}$.

Definition 2. A biological regulatory network is a couple $\mathscr{R} = (G, \mathscr{K})$ where $G = (V, E)$ is a biological regulatory graph and $\mathscr{K} = \{k_{v,\omega}\}$ is a family of positive integers indexed by the set of couples (v, ω) such that

- v belongs to V,
- ω is a subset of $G^{-1}(v)$ and will be called a set of resources of v,
- $k_{v,\omega} \leqslant b_v$.

For example to turn the biological regulatory graphs of Fig. 4 into regulatory networks, six parameters have to be given $\mathscr{K} = \{k_{x,\{\}}, k_{x,\{y\}}, k_{x,\{x\}}, k_{x,\{x,y\}}, k_{y,\{\}}, k_{y,\{x\}}\}$. Because $b_x = 2$ and $b_y = 1$, each $k_{x,\ldots}$ can take the value 0, 1 or 2 and each $k_{y,\ldots}$ the value 0 or 1. So $3^4 \times 2^2$ different networks can be a priori associated with each graph of Fig. 4, which makes 648 different possible models. (Notice that by construction $k_{mucus,\{\}}$ is necessarily equal to 0 and $k_{mucus,\{x\}} = 1$.)

Unfortunately in general the parameters of $\mathscr{K}$ are not measurable in vivo. Consequently, additional properties deduced from biological experiments are needed to eliminate the models which do not satisfy them. This requires to study the dynamic behaviors of models.

4. Dynamics of biological regulatory networks

At a given time, each variable of a regulatory network has a unique concentration level. The knowledge of this concentration level for each variable is the *state* of the system.

Definition 3. A state of a biological regulatory network is a tuple $(n_{v_1}, \ldots, n_{v_p})$ where p is the number of variables, such that for each variable v_i, $n_{v_i} \in \mathbb{N}$ and $n_{v_i} \leqslant b_{v_i}$.

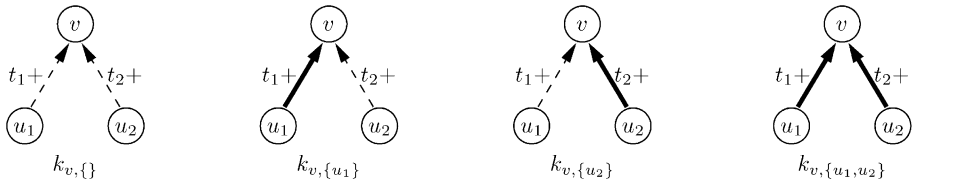

Fig. 5. Dashed arrows do not pass their threshold.

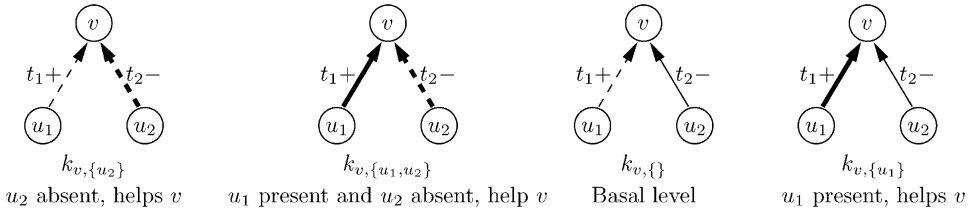

Fig. 6. Dashed arrows do not pass their threshold.

From a given state, the set of resources of a variable is deduced from the threshold of each edge.

Definition 4. Given a biological regulatory network, a state $(n_{v_1}, \ldots, n_{v_p})$ and an edge $(v_i \to v_j)$ labeled with (t, ε), the variable v_i is a *resource* of v_j iff:

- $n_{v_i} \geqslant t$ when $\varepsilon = +$,
- $n_{v_i} < t$ when $\varepsilon = -$.

This defines the set of resources ω of a variable v with respect to a state, as a subset of $G^{-1}(v)$.

Note the algebraic trick: a resource is either the presence of an activator or the *absence* of an inhibitor (whose concentration level does not reach the threshold, Fig. 6).

The parameters of $\mathscr{K}$ define in a simple way an automaton, called synchronous, which is an intermediate technical step to define the dynamics.

Definition 5. Let $\mathscr{R} = ((V, E), \mathscr{K})$ be a regulatory network, its *synchronous state graph* $\mathscr{S} = (S, T)$ is defined as follow:

- the set of vertices S contains all possible states, i.e. $\prod_{v \in V} [0, b_v]$.
- T is the set of edges of the form: $(n_{v_1}, \ldots, n_{v_p}) \to (k_{v_1,\omega_1}, \ldots, k_{v_p,\omega_p})$ where for all i, ω_i is the set of resources of v_i for the state $(n_{v_1}, \ldots, n_{v_p})$.

The out-degree of each vertex is exactly one in the synchronous state graph, thus it can be represented by a table which gives for each state the next state. Table 1 characterizes the synchronous state graphs, respectively, associated to Figs. 4(a) and (b). The indices of the parameters $k_{v,\omega}$ are uniquely determined by the column of the table for v, and by the thresholds of the

Table 1
State tables deduced from Figs. 4(a) and (b), respectively

State		Next state		State		Next state	
x	y	x	y	x	y	x	y
0	0	$k^a_{x,\{y\}}$	$k^a_{y,\{\}}$	0	0	$k^b_{x,\{y\}}$	$k^b_{y,\{\}}$
0	1	$k^a_{x,\{\}}$	$k^a_{y,\{\}}$	0	1	$k^b_{x,\{\}}$	$k^b_{y,\{\}}$
1	0	$k^a_{x,\{y\}}$	$k^a_{y,\{x\}}$	1	0	$k^b_{x,\{x,y\}}$	$k^b_{y,\{\}}$
1	1	$k^a_{x,\{\}}$	$k^a_{y,\{x\}}$	1	1	$k^b_{x,\{x\}}$	$k^b_{y,\{\}}$
2	0	$k^a_{x,\{x,y\}}$	$k^a_{y,\{x\}}$	2	0	$k^b_{x,\{x,y\}}$	$k^b_{y,\{x\}}$
2	1	$k^a_{x,\{x\}}$	$k^a_{y,\{x\}}$	2	1	$k^b_{x,\{x\}}$	$k^b_{y,\{x\}}$

underlying regulatory graph for ω (according to Definition 4). Each instantiation of the parameters in $\mathscr{K}$ defines an a priori different synchronous state graph.

An instantiation of $\mathscr{K}$ being given we can draw the synchronous state graph in a grid of dimension p: in Fig. 7 the mucus node has been ignored in order to get a 2-D grid.

Terminology: One calls *transition* an edge between two states of a state graph.

We build the dynamics of a regulatory network from the synchronous state graph according to two main ideas (Thomas and Kaufman, 2001b):

- A diagonal arrow in the synchronous state graph is a transition that changes *simultaneously* the concentration level of two or more variables. The probability that both variables pass through their respective thresholds at the same time is negligible in vivo, but we do not know which one will be passed first. Accordingly we replace any diagonal transition by the collection of the transitions which modify only one of the involved variables at a time. For example, $(1,0) \to (2,1)$ is replaced by the transitions

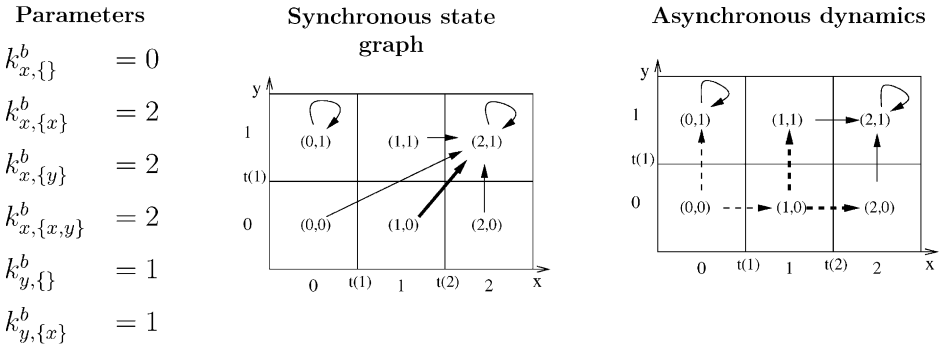

Fig. 7. Deducing dynamics from parameter values.

$(1,0)\to(2,0)$ and $(1,0)\to(1,1)$ in the asynchronous dynamics of Fig. 7 (bold arrows).

- An arrow of length greater or equal to 2 implies a variable which increases its concentration level abruptly and jumps several thresholds. Notice that the concentrations vary continuously in time, independently of whether they vary rapidly or not. The target of the synchronous transition is indeed only an attractor and real transitions should only address neighbor states. For example in Fig. 7, $(0,0)\to(2,1)$ gives rise to $(0,0)\to(1,0)$ instead of $(0,0)\to(2,0)$.

Definition 6. Let τ a transition $(n_{v_1},\ldots,n_{v_p})\to(n'_{v_1},\ldots,n'_{v_p})$. A *desynchronization* of τ is a transition of the form

$$(n_{v_1},\ldots,n_{v_{i-1}},n_{v_i},n_{v_{i+1}},\ldots,n_{v_p})\to(n_{v_1},\ldots,n_{v_{i-1}},n_{v_i}+\delta,n_{v_{i+1}},\ldots,n_{v_p}),$$

where:

- i is such that $n_{v_i}\neq n'_{v_i}$,
- $\delta=1$ if $n_{v_i}<n'_{v_i}$, else $\delta=-1$.

The dynamics of a regulatory network is defined by desynchronizing the synchronous state graph.

Definition 7. Let $\mathscr{R}$ be a regulatory network, its *asynchronous state graph* is defined as follows:

- the set of vertices is the set of states $\prod_{v\in V}[0,b_v]$,
- the set of transitions is the set of all desynchronizations of all transitions of the synchronous state graph of $\mathscr{R}$.

When several transitions start from the same state, they are concurrent and any of them can be randomly chosen. The attractors of the synchronous state graph remain the attractors of the asynchronous one, but the paths to them differ and can change the behavior from a given initial state. For example in Fig. 7, $(0,0)$ can reach $(0,1)$.

5. Computational tree logic and model checking

Parameters of $\mathscr{K}$ play a major role on the dynamics of the model. Unfortunately, most often they are not experimentally measurable. Indeed finding suitable classes of those parameters constitutes a major issue of the modeling activity. A particular class of continuous differential equation systems leads to some constraints on parameters (Snoussi and Thomas, 1993). More precisely if Thomas' approach is seen as a discretization of such a class of continuous differential equation systems, then parameters $k_{v,\omega}$ reflect a discretization of sums of ratios of positive constants (Snoussi and Thomas, 1993). In such a case, $\mathscr{K}$ has to verify the following constraints:

$$k_{v,\emptyset}=0 \quad \text{and} \quad \omega\subseteq\omega' \Rightarrow k_{v,\omega}\leqslant k_{v,\omega'}.$$

Nevertheless, it is possible to slacken these previous constraints to enlarge the set of models which can be described by this discrete formalism. For example, constitutive genes can have an expression different to the abstract level 0 (first constraint) depending of experimental conditions. Since these constraints are not required to have a coherent discrete description, they are often relaxed.

For the *Pseudomonas aeruginosa* example, it seems reasonable to take into account the previous constraints. Then 56 of the 648 initial parameter sets (28 for each regulatory graph of Fig. 4) satisfy them. They lead to 38 different asynchronous state graphs (16 for the first regulatory graph and 22 for the second).

To go further, biological knowledge or hypotheses about the behavior have to be used as indirect criteria to constrain $\mathscr{K}$. For example homeostasy (resp. multi-stationarity) is experimentally observable and, as mentioned in Section 1, it indicates that a negative (resp. positive) circuit is functional. Some necessary conditions for functionality of a circuit can constrain $\mathscr{K}$ (notion of characteristic states in Thomas et al. (1995)). For the running example, if the *mucoidy* could be explained by an epigenetic phenomenon then a multi-stationarity is necessary, and this leads to the functionality of the

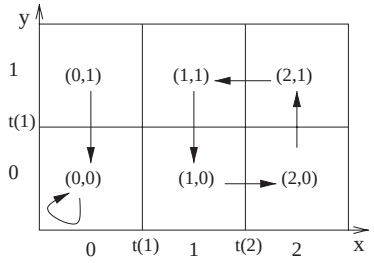

Fig. 8. An asynchronous state graph with two steady states. One steady state is singular as it is revealed by the cycle $(1,1)\to(1,0)\to(2,0)\to(2,1)\to(1,1)$.

unique positive circuit $x\to x$. Among the 56 previously mentioned sets of parameters, 19 satisfy this biological hypothesis (9 for the first regulatory graph and 10 for the second one). They lead to $7+10$ different asynchronous state graphs (for example, the one of Fig. 8).

To still go further, such static conditions must be reinforced by temporal properties formalizing biological knowledge or hypotheses. Since numerous models (families of parameters $\mathscr{K}$) have to be checked against those properties, a formal language (temporal logic (Emerson, 1990)) is needed to perform the checkings by computer. The temporal logic chosen here is CTL (**C**omputation **T**ree **L**ogic) because, according to the asynchronous state graphs of Definition 7, time has a tree structure.

Definition 8. A CTL formula on a regulatory network $\mathscr{R}$ is inductively defined by:

- atomic formulae are $\top$, $\bot$ or of the form $(v=n)$ where v is a variable of $\mathscr{R}$ and $n\in[0,b_v]$,
- if ϕ and ψ are formulae, then $(\neg\phi)$, $(\phi\wedge\psi)$, $(\phi\vee\psi)$, $(\phi\Rightarrow\psi)$, $AX\phi$, $EX\phi$, $A[\phi U\psi]$, $E[\phi U\psi]$, $AG\phi$, $EG\phi$, $AF\phi$, $EF\phi$ are formulae.

$\top$ is the always true formula; $\bot$ is the always false formula; $(v=n)$ is true iff the concentration level of the variable v is equal to n in the current state; $\neg,\wedge,\vee,\Rightarrow$ are the usual connectives (respectively *not*, *and*, *or*, *implication*). All the temporal connectives are pairs of symbols: the first element of the pair is A or E followed by X, F, G or U whose meanings are given in the next table.

A	for All paths choices	X	neXt state
E	for at least one path choice (Exist)	F	some Future state
		G	all future states (Globally)
		U	Until

For example $AX(y=1)$ means that in all next states accessible from the current state in the asynchronous state graph, the concentration level of y is 1. Note that this last formula is false in the asynchronous state graph of Fig. 7 if the initial state is $(0,0)$ or $(1,0)$ and it is true for all other initial states. $EG(x=0)$ means that there exists at least one path starting from the current state where the concentration level of x is constantly equal to 0. In Fig. 7 only states $(0,0)$ and $(0,1)$ satisfy the formula. $A[(x=0)U(x-1)]$ means that for any possible path from the current state there exists a time where $x=1$ and in between x remains equal to 0; this last formula is false in the asynchronous state graph of Fig. 7 if the current state is $(0,0)$ because the transition $(0,0)\to(0,1)$ leads to a path where x never reaches 1. And so on for other temporal connectives.

To test the epigenetic hypothesis described in Section 1 "*to find a model of the wild bacteria, which has a multi-stationarity where one stable state produces mucus and the other one does not*", we have proved first that for any model "*regularly produces mucus*" is equivalent to the fact that the concentration level of x is repeatedly equal to 2. Thus recurrence of mucus production can be expressed as follows:

$$(x=2)\Rightarrow AX\,AF(x=2), \tag{1}$$

where $AX\,AF(\phi)$ means that for all possible futures (excluding the present) ϕ will be satisfied at a given time. Moreover we know that the wild bacteria never produce mucus by themselves when starting from a basal state (second stable state):

$$(x=0)\Rightarrow AG(\neg(x=2)). \tag{2}$$

With respect to mucus production, $x=2$ induces the mucus production (and $x=0$ is the basal state). Consequently the AG statement says that mucus is never produced.

On the one hand, proposing a method independent of the example to formally express a biological hypothesis remains a difficult open problem. Here the major key to overcome the problem is our lemma about the relationship between $x=2$ and the mucus production. On the other hand, the CTL formulae being given, one can *automatically* extract the compatible models, i.e. the compatible families of parameters $\mathscr{K}$. For each state graph, *model checking* very efficiently computes all the states which satisfy a set of formulae (Huth and Ryan, 2000). If all the states satisfy the formulae, one says that the model satisfies them.

We have designed a software, `SMBioNet` (Selection of Models of Biological Networks), which allows one to select models of given regulatory graphs according to their temporal properties. The software takes as input biological regulatory graphs (with a graphical interface), CTL formulae and a set of functional loops. It gives as output all the models from the regulatory graphs which satisfy the formulae and make functional the given loops. Technically, `SMBioNet`:

- generates, from the graphs, all the biological regulatory networks;

- selects all models whose characteristic states of the specified loops cope with the constraints mentioned before;
- finally returns the models and associated asynchronous state graphs which satisfy the CTL formulae (using the SMV model checker (McMillan, 1993)).

Applied to the *Pseudomonas aeruginosa* problem, `SMBioNet` firstly automatically extracted the 19 models mentioned before out of the 648 possibilities, then using the two previous formulae, it finally selected 4 models leading to 4 different asynchronous state graphs for each regulatory graph (Fig. 4(a) and (b)). Consequently, `SMBioNet` gave us a positive answer to the epigenetic question: the set of remaining models is non-empty.

If *Pseudomonas aeruginosa* is actually compatible with one of these remaining models (whichever the "exact" model is, because they are observationally equivalent), it could open new therapeutics in prospects. We know that *Pseudomonas aeruginosa* satisfies Formula 2 (non-mucoid stationary state). Formula 1 constitutes the corner stone of the problem. Its logical structure suggests an experiment plan: pulse x up to saturation by an external signal, and *after the transitory phase* due to the pulse, check if the mucus production persists (Guespin-Michel and Kaufman, 2001). By the way, automatic tools dedicated to software testing can generate this experiment plan (the initial pulse comes from the left-hand side atomic formula ($x = 2$), and so on). We have also rigorously proved that the success of this experiment plan (prepared at LMDF in Rouen) is sufficient to validate the epigenetic hypothesis.

Assuming that the experiments are successful, it would give us another information on "satisfactory" models: the ability to switch from the normal to the mucoid state under a change of conditions. This knowledge can in general be expressed in a logical manner with CTL formulae which in turn can be used to reduce the number of remaining possible models. Here, the change of conditions (pulse of x) simply modifies the initial state to get $x = 2$. Thus formula 1 has already captured the knowledge that the new stationary state is mucoid and the experiment does not contribute to reduce the number of potential models. This comes from the simplicity of our example. In general every new experiment reveals new temporal properties which can be used to focus on a smaller set of satisfactory models, via model checking.

6. Conclusion and perspectives

We have defined a *formal* description of biological regulatory networks which allows a computer aided manipulation of the semantics of the discrete modeling of Thomas, this manipulation being proved correct by construction owing to the formalism. Our approach allows biology to take advantage of the whole corpus of formal methods from computer science. In particular temporal properties can be checked against models using CTL and model checking. Model checking is a first powerful tool offered by the formalization of biological regulatory networks. The cooperation of molecular biology and formal methods from computer science opens a large horizon of research perspectives. Let us mention for instance,

- To extract the specificities of the biological application domain in order to provide a user friendly syntax for temporal logics (patterns of formulae to express functionality, etc).
- Automatic generation of experiment plans.
- Preservation of properties when a regulatory network is embedded into another one, including the systematic treatment of knock-out mutants, identification of functional patterns (Shen-Orr et al., 2002) as well as the structuration of huge regulatory networks.
- Useful extensions of the Thomas' framework such as allowing variables which are both activator and inhibitor of the same target, taking into account time delays (Thomas and Kaufman, 2001b) between the beginning of the activation order on a variable and the synthesis of the product and conversely for the turn-off delays, offering a language to control transitions, taking into account populations of networks whose states are not synchronized, etc.

These constitute ongoing or future works of our genopole® research groups. Our aim is to link modelization and experiments together, by furnishing to biologists model structuration methods and model validation tools from current researches in theoretical computer science. The resulting formal models are not only a posteriori explanations of biological results, they are guides for biological experiments whose success will be in fine the discriminent criterion.

Acknowledgements

The authors thank genopole®-research in Evry (H. Pollard and P. Tambourin) for constant supports. We gratefully acknowledge the members of the genopole® working groups *observability* and G^3 for stimulating interactions. Comments of *D. Thieffry* at the beginning of this work have been very constructive. We also thank *V. Bassano* and *S. Pérès* for the prototypes they have developed and for helpful discussions. We are very grateful to the anonymous referee for very deep comments which have considerably improved the quality of this article. Let us finally mention that the biological experiments are made in Rouen in cooperation with a team of the university hospital center of

Grenoble in France, funded by the "*Association: vaincre la mucoviscidose*".

References

Cinquin, O., Demongeot, J., 2002. Positive and negative feedback: striking a balance between necessary antagonists. J. Theor. Biol. 216 (2), 229–241.

de Jong, H., 2002. Modeling and simulation of genetic regulatory systems: a literature review. J. Comput. Biol. 9 (1), 67–103.

Demongeot, J., Benaouda, D., Jezequel, C., 1995. Dynamical confinement in neural networks and cell cycle. Chaos 5 (1), 167–173.

Demongeot, J., Kaufman, M., Thomas, R., 2000. Positive feedback circuits and memory. C.R. Acad. Sci. III 323 (1), 69–79.

Demongeot, J., Aracena, J., Thuderoz, F., Baum, T., Cohen, O., 2003. Genetic regulation networks: circuits, regulons and attractors. C. R. Biol. 326 (2), 171–188.

Devloo, V., Hansen, P., Labbé, M., 2003. Identification of all steady states in large networks by logical analysis. Bull. Math. Biol. 65 (6), 1025–1051.

Emerson, E., 1990. In: Temporal and Modal Logic. In: van Leeuwen, J. (Ed.), Handbook of Theoretical Computer Science, Vol. B: Formal Models and Semantics. MIT Press, Cambridge, MA, pp. 995–1072.

Guespin-Michel, J., Kaufman, M., 2001. Positive feedback circuits and adaptive regulations in bacteria. Acta Biotheor. 49 (4), 207–218.

Hasty, J., McMillen, D., Collins, J., 2002. Engineered gene circuits. Nature 420 (6912), 224–230.

Huang, S., 2001. Genomics, complexity and drug discovery: insights from boolean network models of cellular regulation. Pharmacogenomics 2 (3), 203–222.

Huth, M., Ryan, M., 2000. Logic in Computer Science: Modelling and Reasoning about Systems. Cambridge University Press, Cambridge.

Kaufman, M., Thomas, R., 1987. Model analysis of the bases of multistationarity in the humoral immune response. J. Theor. Biol. 129 (2), 141–162.

Kaufman, M., Urbain, J., Thomas, R., 1985. Towards a logical analysis of the immune response. J. Theor. Biol. 114 (4), 527–561.

Kitano, H., 2002a. Computational systems biology. Nature 420 (6912), 206–210.

Kitano, H., 2002b. Looking beyond the details: a rise in system-oriented approaches in genetics and molecular biology. Curr. Genet. 41 (1), 1–10.

McMillan, K., 1993. Symbolic Model Checking. Kluwer Academic Publishers, Dordrecht.

Pérès, S., Comet, J.-P., 2003. Contribution of computation tree logic to biological regulatory networks: example from pseudomonas aeruginosa. In: Priami, C. (Ed.), Proceedings of the First International Workshop CMSB'2003, Lecture Notes in Computer Science, Vol. 2602. Springer, Berlin, pp. 47–56.

Shen-Orr, S., Milo, R., Mangan, S., Alon, U., 2002. Network motifs in the transcriptional regulation network of *Escherichia coli*. Nat. Genet. 31 (1), 64–68.

Snoussi, E., 1989. Qualitative dynamics of a piecewise-linear differential equations: a discrete mapping approach. Dynamics Stability Systems 4, 189–207.

Snoussi, E., Thomas, R., 1993. Logical identification of all steady states: the concept of feedback loop characteristic states. Bull. Math. Biol. 55 (5), 973–991.

Thomas, R., Gathoye, A., Lambert, L., 1976. A complex control circuit regulation of immunity in temperate bacteriophages. Eur. J. Biochem. 71 (1), 211–227.

Thomas, R., 1978. Logical analysis of systems comprising feedback loops. J. Theor. Biol. 73 (4), 631–656.

Thomas, R., 1980. On the Relation Between the Logical Structure of Systems and their Ability to Generate Multiple Steady States or Sustained Oscillations, Springer Series in Synergies, Vol. 9, Springer, Berlin, pp. 180–193.

Thomas, R., Kaufman, M., 2001a. Multistationarity, the basis of cell differentiation and memory. II. Logical analysis of regulatory networks in terms of feedback circuits. Chaos 11, 180–195.

Thomas, R., Kaufman, M., 2001b. Multistationarity, the basis of cell differentiation and memory. I. & II. Chaos 11, 170–195.

Thomas, R., Thieffry, D., Kaufman, M., 1995. Dynamical behavior of biological regulatory networks—I. Bull. Math. Biol. 57 (2), 247–276.

Wolkenhauer, O., 2001. Systems biology: the reincarnation of systems theory applied in biology? Brief Bioinform. 2 (3), 258–270.