Qualitative modelling and analysis of gene regulatory networks: application to the adaptation of Escherichia coli bacterium to carbon availability

Jamil Ahmad, Jérémie Bourdon, Damien Eveillard, Jonathan Fromentin, Olivier Roux, Christine Sinoquet

Qualitative modelling and analysis of gene regulatory networks: application to the adaptation of *Escherichia coli* bacterium to carbon availability

Jamil Ahmad[a], **Jérémie Bourdon**[b,c], **Damien Eveillard**[b], **Jonathan Fromentin**[a], **Olivier Roux**[a], **Christine Sinoquet**[b]

[a] IRCCyN, UMR C.N.R.S. 6597, Ecole Centrale de Nantes, 1 rue de la Noë, 44321 Nantes, France
[b] Lina, UMR C.N.R.S. 6241, Université de Nantes, 2 rue de la Houssinière, 44322 Nantes, France
[c] Centre INRIA Rennes Bretagne Atlantique, IRISA, campus de Beaulieu, F - 35 042 Rennes Cedex, France

— *Bioinformatics* —

LABORATOIRE D'INFORMATIQUE DE NANTES ATLANTIQUE

RESEARCH REPORT

N° hal-00359530

February 2009

UNIVERSITÉ DE NANTES — ECOLE DES MINES DE NANTES — CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Jamil Ahmad[a], **Jérémie Bourdon**[b,c], **Damien Eveillard**[b], **Jonathan Fromentin**[a], **Olivier Roux**[a], **Christine Sinoquet**[b]

Qualitative modelling and analysis of gene regulatory networks: application to the adaptation of Escherichia coli **bacterium to carbon availability**

32 p.

Les rapports de recherche du Laboratoire d'Informatique de Nantes-Atlantique
sont disponibles aux formats PostScript® et PDF® à l'URL :
http://www.sciences.univ-nantes.fr/lina/Vie/RR/rapports.html

Research reports from the Laboratoire d'Informatique de Nantes-Atlantique are available in PostScript® and PDF® formats at the URL:
http://www.sciences.univ-nantes.fr/lina/Vie/RR/rapports.html

Qualitative modelling and analysis of gene regulatory networks: application to the adaptation of *Escherichia coli* bacterium to carbon availability

Jamil Ahmad[a], Jérémie Bourdon[b,c], Damien Eveillard[b], Jonathan Fromentin[a], Olivier Roux[a], Christine Sinoquet[b]

Abstract

Attempts to model Gene Regulatory Networks (GRNs) have yielded very different approaches. Among others, variants of Thomas's asynchronous boolean approach have been proposed, to better fit the dynamics of biological systems: notably, genes were allowed to reach different discrete expression levels, depending on the states of other genes, called the regulators: thus, activations and inhibitions are triggered conditionally on the proper expression levels of these regulators. In contrast, some fine-grained propositions have focused on the molecular level as modelling the evolution of biological compound concentrations through differential equation systems. Both approaches are limited. The first one leads to an oversimplification of the system, whereas the second is incapable to tackle large GRNs. In this context, hybrid paradigms, that mix discrete and continuous features underlying distinct biological properties, achieve significant advances for investigating biological properties. One of these hybrid formalisms proposes to focus, within a GRN abstraction, on the time delay to pass from a gene expression level to the next. Until now, no research work has been carried out, which attempts to benefit from the modelling of a GRN by differential equations, converting it into a multi-valued logical formalism of Thomas, with the aim of performing biological applications. The present research work fills this gap by describing a whole pipelined process which supervises the following stages: (i) model conversion from a Piece-wise Affine Differential Equation (PADE) modelization scheme into a discrete model with attractors (and generation of the corresponune pour la journée portes ouvertes de PolyTech ?ding dynamical graph), (ii) on the basis of probabilistic criteria, extraction of subgraphs of interest from the former dynamical graph, (iii) conversion of the subgraphs into Parametric Linear Hybrid Automata, (iv) analysis of dynamical properties (*e.g.* cyclic behaviours) using hybrid model-checking techniques. The present work is the outcome of a methodological investigation launched to cope with the GRN responsible for the reaction of *Escherichia coli* bacterium to carbon starvation. As expected, we retrieve a remarkable cycle already exhibited by a previous analysis of the PADE model. Above all, hybrid model-checking enables us to discover additional insightful results, whose interpretations are in accordance with biological evidences.

Foreword

Due to their equally important complementary contributions, the authors would emphasize that the order chosen for the author list is the alphabetical order.

Introduction

A Gene Regulatory Network (GRN) is a collection of macromolecular compounds such as DNA and proteins, which functionally interact with each other in a cell. Some proteins, the transcription factors (TFs), serve only to activate genes and are therefore the main players in regulatory networks or cascades. By binding to the promoter region in the regulatory region of other genes, TFs turn the latter on, initiating the production of another protein, and so on. Some TFs are inhibitory. These interactions thereby govern the rates at which genes in the network are transcribed into mRNA...

In the simplest cases - that is when interactions do not involve more than two compounds at a time -, a GRN is typically described as a simple directed graph whose vertices are the components (for illustration, see Figure 2 (a)). The existence of a labelled directed edge between a pair of genes symbolizes an activation (+) or an inhibition (-) exerted by a gene over another gene through a protein production; besides, the label also mentions the expression level of the regulator gene for which the regulation (activation or inhibition) is triggered. Note that a non-inhibiting status is equivalent to an activating status, and symmetrically. Besides, a gene may contribute to activate another gene, together with other co-activators. A gene may also be the co-inhibitor of another gene. More generally, the co-regulation of a given gene is likely to involve activators as well as inhibitors. Since regulation is triggered depending on gene expression levels, the regulation of a given gene may involve various sets of co-regulator genes throughout the whole biological system's life. Hereafter, such set of genes will be called a *resource* for the regulated gene. In summary, given the current activating or inhibiting statuses of potentially co-regulating genes, a GRN determines the expression level of the gene under regulation, itself a potential regulator for other genes. In this regulatory context, investigating the respective behaviours of genes remains a key question.

Various models of GRNs have been developed to capture the behaviour of the system being modeled, and infer dynamical properties (see de (Jong, 2002) for a review). The following modelling techniques used include Boolean networks (Kauffman, 1993), Petri nets (Chaouiya, 2007), Bayesian networks (Hartemink *et al.*, 2001; Yu *et al.*, 2002), graphical Gaussian models (Markowetz *et al.*, 2005), Stochastic (Golightly *et al.*, 2006) and Process Calculi (Kuttler *et al.*, 2006). The most realistic dynamical models lie on differential equation systems dealing with protein productions that activate or repress genes. However, such modelling is not implementable for realistic biological systems, due to many unknown parameters. Thus, various alternative modelling approaches were proposed. Discretizing protein concentration by thresholds quickly appeared as an attractive lead. Henceforth, we will indifferently refer to protein concentration levels or gene expression thresholds. Two categories of approaches implement such a discretized approximation. On the one hand, qualitative methods based on Piecewise-

Affine Differential Equations (PADEs) showed relevant enough to overcome the lack of quantitative data on kinetic parameters and molecular concentrations and to fit biologists'expectations (Glass *et al.*, 1973; Snoussi, 1989; de Jong *et al.*, 2004; Batt *et al.*, 2005). On the other hand, an approach first proposed by Thomas combines discretization (both in terms of gene expression levels and time) with the attractor concept (Thomas *et al.*, 1990; Thomas, 1991; Snoussi *et al.*, 1993; Thomas *et al.*, 1995). The definition of this concept will be briefly recalled in the sequel. Time is viewed as proceeding in discrete steps. At each instant t, the current expression levels of the GRN's genes determine the genes' *attractors*, which are the thresholds towards which the genes'expression levels tend to evolve and which will therefore be assigned to genes at instant $t + 1$.

However, some processes, and among them, gene transcription, involve many biochemical reactions or may be delayed until the appropriate molecules are available, which can take time due to possible low concentrations of the latter in the cell. Now, the discrete model with attractors originally proposed implements instantaneous variations of the thresholds. In an ideal model based on discretization, transitions between expression level thresholds would be modelled as sigmoidal functions of the time. Due to unknown tuning parameters, this model is generally not implementable for realistic biological systems. An approximated model has thus been designed to cope with delays; it implements linear variations between genes'thresholds.

In this report, we tackle the problem of describing a realistic GRN through the approach of Thomas, extended with delays. The ultimate aim is identifying essential features of the dynamical behaviour of the GRN studied, using model-checking techniques. As a case study example, we analyse the GRN of the nutritional stress response in *Escherichia coli* bacterium. Though this GRN has been widely studied, the relation between the growth of *E. coli* and the availability of carbon source is still little understood.

In our approach, the discrete model is built from a formerly published PADE model (Ropers, 2006), thus benefitting from its parameter tuning. Besides, as the set of global states obtained as well as the transition graph are huge, our work is also novel in that it copes with this difficulty, implementing a complementary probabilistic approach: the latter is used to highlight subgraphs showing characterized states. Then, any such subgraph can be converted into a hybrid model with delays, for the purpose of behavioural property inference. Model-checking tools can be used to analyse these hybrid models.

We first describe the method implementing the conversion of a Piecewise-Affine Differential Equation model into a discrete model with attractors (Section 2). Nevertheless, the discrete model of a large GRN is not easily tractable for property inference implemented through model-checking techniques. Therefore, in Section 3, a method dedicated to the extraction of subgraphs of interest in the dynamical graph is proposed. This process is performed on the basis of a probabilistic rationale and identifies subgraphs characterized with remarkable states. Then, Section 4 focuses on the integration of delays into the discrete model, leading to an hybrid system. Throughout our exposition, the simplicist regulation system for bacterium *Pseudomonas aeruginosa*'s mucus production will be used for illustration. The outcome of our methodological approach is the processing scheme depicted in Figure 1. In Section 5, we apply the whole pipelined process in the case of the response of *Escherichia coli* bacterium to carbon availability.

Therein, we present and discuss insightful results obtained for this realistic case, originally the instigator case for the pipelined process design.

1 Conversion of a PADE model into a model with attractors

1.1 PADE model

We first recall how the concentration evolution of a protein regulated by a GRN can be modelled through a Piece-wise Affine Differential Equation (Snoussi, 1989; de Jong *et al.*, 2001). PADE modelling relies on discretization: for each protein i, its concentration is known to evolve within a domain discretized into an ordered set of thresholds $\theta_1, \theta_2...$

Definition 1 (Evolution of protein concentration)
Typically, the evolution of concentration x_i with time is expressed as: $\dot{x}_i = f_i(x) - \gamma_i x_i, \quad 1 \leq i \leq n, \quad x_i \geq 0,$
where $x = (x_1, \ldots, x_n)$ is a vector of n protein concentrations. The equation above relates the concentration modification rate $\dot{x}_i$ to a synthesis rate, $f_i(x)$, and a degradation rate, $\gamma_i x_i$.

Functions f_i express the dependence of x_i upon the concentrations of other constituents present in the cell. Such functions are derived from basic principles of chemical kinetics, including for example Michaelis-Menten enzymatic kinetics.

Notation 1 (Resource set)
In the following, $R(i)$ will denote the set of all resources likely to regulate gene i. A resource for gene i is itself a set of genes (possibly inclusing gene i) involved in the co-regulation of gene i.

Definition 2 (Description of regulation)
$f_i(x)$ expresses the synthesis rate of component i as a function of the concentrations of regulator genes in i^{th} gene's resources:
$f_i(x) = k_i + \sum_{r \in R(i)} k_{ir}\, b_{ir}(x), \quad k_i \in \mathbb{R}^{+*}, \quad k_{ir} \in \mathbb{R}^{+}, \quad b_{ir} \in \{0, 1\},$
where k_i and k_{ir} are kinetic parameters.

Switching the boolean parameter $b_{ir}(x)$ to 1 means that the corresponding resource r is active, that is each gene r_j belonging to the resource set r is either an activator or a non-inhibitor for gene i, depending on its concentration x_{r_j}. Switching $b_{ir}(x)$ from 0 to 1 and symmetrically relies on the satisfaction of constraints by the concentrations relative to the genes belonging to resource set r. In a discrete framework, such constraints are expressed through concentration thresholds.

Before we may further explain how entities b_{ir} describe regulator contributions, we need detail the concept of discretization. Such concentration thresholds aforecited are defined as follows:

Definition 3 (Discrete concentration thresholds)
θ_{j_α} denotes one of the thresholds between which the continuous variable x_j is likely to evolve.

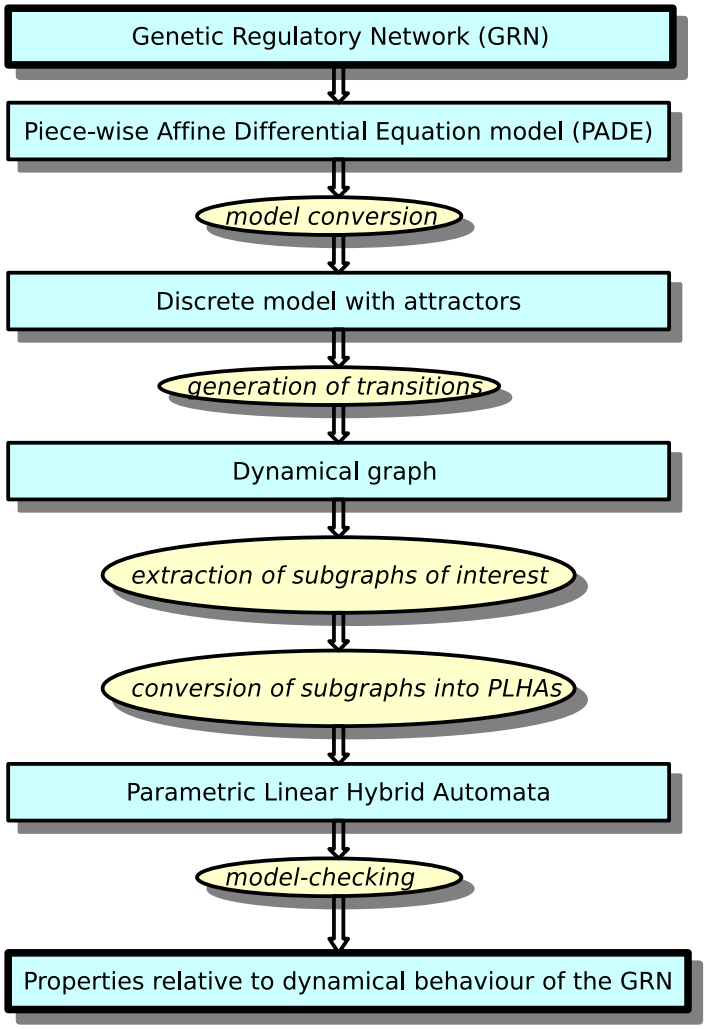


Figure 1: The pipelined process designed for the analysis of large GRNs.

For gene j characterized by τ_j thresholds, the following ordered relation is verified: $0 < \theta_{j_1} < \theta_{j_2} < \cdots < \theta_{j_{\tau_j}}$.

Any such set of thresholds defines a set of domains, further called local states, traversed by the system under study, when considering only gene j. More generally, this system evolves through global states, which refer to all possible combinations of local states associated with the genes in the system. Such previous concepts establish the notion of discrete dynamics of the system.

Then, the $b_{ir}(x)$ terms in Definition 2.2 will be tailored as functions of entities defined as follows:

Definition 4 (Step function)
Given r_j, a regulator gene belonging to resource r, and one of its τ_j thresholds $\theta_{r_{j\alpha}}$,

$$s^+(x_{r_j}, \theta_{r_{j\alpha}}) = \begin{cases} 1, & if \quad x_{r_j} \geq \theta_{r_{j\alpha}} \\ 0, & if \quad x_{r_j} < \theta_{r_{j\alpha}} \end{cases}$$

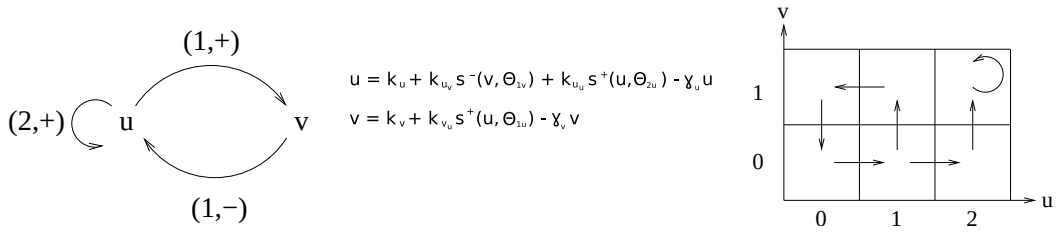
$$s^-(x_{r_j}, \theta_{r_{j\alpha}}) = 1 - s^+(x_{r_j}, \theta_{r_{j\alpha}}).$$

Finally, any co-regulation involving the genes of a resource set r may be modelled adapting b_{ir} as a combination of various step functions s^+ and s^-. The following grammar enumerates all possible combinations:

$$\begin{array}{ll} b_{ir} & ::= comb \\ comb & ::= s^+ \mid s^- \mid 1 - comb \mid comb\ comb. \end{array}$$

Through the b_{ir} coefficients, the activation or inhibition sigmoidal functions are approximated into piece-wise linear functions.

For a didactic exposition, we will illustrate the various concepts used throughout this article with the simple GRN involved in the mucus production of bacterium *P. aeruginosa*. Figure 2 (b) presents the PADE model corresponding to the GRN described in Figure 2 (a).



(a) (b) (c)

Figure 2: Regulation of the mucus production for *P. aeruginosa* bacterium (a) labelled GRN; (b) Piece-wise Affine Differential Equation model (PADE); (c) asynchronous discrete model. (a) The directed edge $u \rightarrow v$ labelled with $(1, +)$ means that u activates v as soon as u's expression level reaches the threshold value of 1. The directed edge $v \rightarrow u$ labelled with $(1, -)$ indicates that the inhibition of u by v is triggered as soon as v's expression level is 1. Note the positive feedback loop for u.

1.2 Discrete model with attractors

In the abstract semi-qualitative model of Thomas, each gene expression variation domain is discretized using appropriate thresholds. The knowledge of all such gene thresholds is the prerequisite for building the graph whose dynamical behaviours will be studied. Global states are directly inferred identifying all valid threshold combinations. In particular, biological knowledge allows discarding global states which do not exist: for example, antagonist components

can not show simultaneously high (resp. low) expression levels or concentrations. Once the valid global states of the dynamical graph are identified, its transitions have to be inferred. In the model inspired from that proposed by Thomas, the dynamical aspect is modelled through the attractor concept.

Definition 5 (Attractor)
In a given global state s, a gene u is associated to a specific attractor value, representing the expression level towards which this gene will tend to evolve, starting from s_u, its expression level in global state s. The evolution of gene u depends upon one or more other genes, together defining the resource set $r(u,s)$ for u in global state s. Therefore, the attractor value of gene u, $\mathcal{K}_{u,r(u,s)}$, is related to u's resource.

Table 1 recapitulates the three possible evolution tendencies for gene u.

$s_u < \mathcal{K}_{u,r(u,s)}$	The expression level of u tends to increase.
$s_u = \mathcal{K}_{u,r(u,s)}$	The expression level of u is steady.
$s_u > \mathcal{K}_{u,r(u,s)}$	The expression level of u tends to diminish.

Table 1: Determination of the tendency for gene u, depending on its expression level s_u and its attractor value $\mathcal{K}_{u,r(u,s)}$, in global state s. $r(u,s)$ denotes the resource set of gene u, in global state s.

Knowing these tendencies for all genes and for all global states is the key to infer the transitions of the dynamical graph. Central to any modelling paradigm using discretization is the concept of qualitative focal point.

Definition 6 (Qualitative focal point)
In global state s, with each gene u of the system evolving towards its attractor value $\mathcal{K}_{u,r(u,s)}$, the qualitative focal point is defined as the vector $(\mathcal{K}_{u_1,r(u_1,s)}, \cdots, \mathcal{K}_{u_n,r(u_n,s)})$. Any focal point is uniquely associated with an abstract region in the discretized hypercube of dimension n, where each dimension describes local state traversing for one of the n genes of the system.

A most difficult task remains in tuning attractor values: usually, instanciating attractor values for a given global state is an under-constrained problem. Biological knowledge together with Snoussi constraints prohibit some instanciations (Snoussi constraints specify that the addition of supplementary activating (resp. inhibiting) resources for a given gene obligatorily leads to the increase (resp. decrease) of its attractor value). Table 2 shows a possible instanciation for the GRN illustrated in Figure 2 (a). In the asynchronous model, best consonant with biological reality, a change of state is only allowed for at most one gene along each transition. As a result, if a global state s is its own successor, it is a steady global state whereas it possesses p successors if tendency to evolve is detected for p genes. Figure 2 (c) provides the asynchronous description derived from the tendencies of Table 2.

u	v	attractor for u	attractor for v	tendency for v	tendency for u
0	0	$\mathcal{K}_{u,\{v\}} = 2$	$\mathcal{K}_{v,\{\}} = 0$	$\nearrow$	$\rightarrow$
0	1	$\mathcal{K}_{u,\{\}} = 0$	$\mathcal{K}_{v,\{\}} = 0$	$\rightarrow$	$\searrow$
1	0	$\mathcal{K}_{u,\{v\}} = 2$	$\mathcal{K}_{v,\{u\}} = 1$	$\nearrow$	$\nearrow$
1	1	$\mathcal{K}_{u,\{\}} = 0$	$\mathcal{K}_{v,\{u\}} = 1$	$\searrow$	$\rightarrow$
2	0	$\mathcal{K}_{u,\{u,v\}} = 2$	$\mathcal{K}_{v,\{u\}} = 1$	$\rightarrow$	$\nearrow$
2	1	$\mathcal{K}_{u,\{u\}} = 2$	$\mathcal{K}_{v,\{u\}} = 1$	$\rightarrow$	$\rightarrow$
an instanciated model: $\mathcal{K}_{u,\{\}} = 0,\ \mathcal{K}_{u,\{v\}} = 2,\ \mathcal{K}_{u,\{u,v\}} = 2,\ \mathcal{K}_{v,\{\}} = 0,\ \mathcal{K}_{v,\{u\}} = 1$					

Table 2: A possible instanciation of attractors, for the GRN of Figure 2 (a). We explain the third line relative to global state $(s_u = 1,\ s_v = 0)$: since v is not inhibiting u ($s_v < 1$), v activates u as its only resource; u being in state 1, a consistent instanciation for $\mathcal{K}_{u,\{v\}}$ is thus the value of 2; the condition is required for u's activation of v ($s_u \geq 1$) and a coherent value for $\mathcal{K}_{v,\{u\}}$ is therefore 1. In conclusion, both gene expressions tend to increase.

1.3 Model conversion

The key to the conversion of a PADE model into a discrete model with attractors relies on the quasi-straightforward determination of such attractors from the differential equations, as well as a facility to instanciate them through the set of constraints associated with these equations. Indeed, the qualitative focal point of Thomas's formalism coincides with the abstract region (in the hypercube of dimension n) containing the steady state for the PADE system.

Proposition 1 (Conversion rule)
Referring to the PADE *related to gene i (definitions 2.1 and 2.2 combined),* $\dot{x}_i = k_i + \sum_{r \in R(i)} k_{ir}\, b_{ir}(x) - \gamma_i x_i, \quad 1 \leq i \leq n, \quad x_i \geq 0, \quad b_{ir} \in \{0, 1\}$*, we obtain the attractor value of gene i in global state s when $\dot{x}_i$ is equal to 0 (steady state) and $b_{i,r(i,s)}(x)$ is switched to 1 due to the activating regulation of resource $r(i, s)$:*

$$\mathcal{K}_{i,r(i,s)} = \mathscr{D}_i\left(\frac{k_i + \sum_{r \in R(i) \cap r(i,s)} k_{ir}}{\gamma_i}\right),$$

where the discretization function $\mathscr{D}_i$ converts the ratio into one of the τ_i θ_{i_α} thresholds associated with gene i.

Example 1

When applied to the case of P. aeruginosa's mucus production regulation (see Table 3), the conversion process exploits constraints relative to thresholds ((3) to (4)) as well as kinetic parameters ((5) to (8)).

(1) $\dot{u} = k_u + k_{u_v}\ s^-(v, \theta 1_v) + k_{u_u}\ s^+(u, \theta 2_u) - \gamma_u\ u$

(2) $\dot{v} = k_v + k_{v_u}\ s^+(u, \theta 1_u) - \gamma_v\ v$

(3) $0 \leq \theta 1_u < \theta 2_u \leq max_u$

(4) $0 \leq \theta 1_v \leq max_v$

(5) $0 \leq \frac{k_u}{\gamma_u} \leq \theta 1_u$

(6) $\theta 2_u \leq \frac{k_u + k_{u_v}}{\gamma_u} + \frac{k_u + k_{u_u}}{\gamma_u} + \frac{k_u + k_{u_v} + k_{u_u}}{\gamma_u} \leq max_u$

(7) $0 \leq \frac{k_v}{\gamma_v} \leq \theta 1_v$

(8) $\theta 1_v \leq \frac{k_v + k_{v_u}}{\gamma_v} \leq max_v$

(9) $\mathcal{K}_{u,\{\}} = \mathscr{D}_u(\frac{k_u}{\gamma_u})$

(10) $\mathcal{K}_{u,\{u\}} = \mathscr{D}_u(\frac{k_u + k_{u_u}}{\gamma_u})$

(11) $\mathcal{K}_{u,\{v\}} = \mathscr{D}_u(\frac{k_u + k_{u_v}}{\gamma_u})$

(12) $\mathcal{K}_{u,\{u,v\}} = \mathscr{D}_u(\frac{k_u + k_{u_u} + k_{u_v}}{\gamma_u})$

(13) $\mathcal{K}_{v,\{\}} = \mathscr{D}_v(\frac{k_v}{\gamma_v})$

(14) $\mathcal{K}_{v,\{u\}} = \mathscr{D}_v(\frac{k_v + k_{v_u}}{\gamma_v})$

Table 3: Identification of resources and tuning of attractors from the PADE of Figure 2 (b). Attractors are easily identified from equations (1) and (2): in addition to $\mathcal{K}_{u,\{v\}}$, $\mathcal{K}_{u,\{u\}}$ and $\mathcal{K}_{v,\{u\}}$, attractors corresponding to the absence of resource are $\mathcal{K}_{u,\{\}}$ and $\mathcal{K}_{v,\{\}}$. Moreover, attractor $\mathcal{K}_{u,\{u,v\}}$ has to be created. It follows from equations (3) to (8) and from Snoussi constraints ($\mathcal{K}_{u,\{\}} \leq \mathcal{K}_{u,\{v\}} \leq \mathcal{K}_{u,\{u,\ v\}}$, $\mathcal{K}_{u,\{\}} \leq \mathcal{K}_{u,\{u\}} \leq \mathcal{K}_{u,\{u,v\}}$ and $\mathcal{K}_{v,\{\}} \leq \mathcal{K}_{v,\{u\}}$) that one of the possible instanciations is the one deduced in Table 2. $\mathscr{D}_u$ and $\mathscr{D}_v$ are discretization functions used to convert ratios into the appropriate concentration thresholds.

2 Extraction of subgraphs of interest

We implemented a coloration method designed to highlight the most interesting states of the dynamical graph. This method relies on a probabilistic rationale.

Turning the dynamical state graph initialy obtained into a Markov chain is straightforward. For each transition originating from a given global state i ($1 \leq i \leq N$), a probability is computed as the inverse of the outter degree of state i. Formerly, the transition matrix M of the Markov chain associated to the dynamical graph denoted $G = (V, E)$ satisfies

$$\forall i, j \in V, M_{j,i} = \frac{[\![i \rightarrow j \in E]\!]}{\#\{k,\ i \rightarrow k \in E\}},$$

where $[\![B]\!] = 1$ if property B is true and 0 otherwise (Iverson's notation) and $\#\{k,\ i \rightarrow k \in E\}$ is the outter degree of state i.

Next, we define the steady-state probability $\mathbb{P}^\star$ as

$$\mathbb{P}^\star = \lim_{\ell \rightarrow \infty} \frac{1}{N} \sum_{i=0}^{\ell} M^i F,$$

where F is the vector of initial probabilities (in the sequel, this vector is set as $F_i = 1/|V|$, $1 \leq i \leq N$.). Here, the sum ensures the convergence to a unique probability distribution, even in the case of a non irreducible or periodic Markov chain.

We use the steady-state probability $\mathbb{P}^\star$ to highlight vertices in G (*i.e.* the global states of the dynamical graph). Relying on steady-state probabilities is justified by their being closely related to the number of times the different states are traversed in infinite random trajectories. Consequently, the higher is such a probability, the more important would be the associated state, with regard to the system's behaviour.

As infinite trajectories do not make sense in a biological context, it is more relevant instead to focus on finite trajectories. We define the vector $\mathbb{P}^\ell$ of ℓ-finite state probabilities to be

$$\mathbb{P}^\ell = \frac{1}{\ell} \sum_{i=0}^{\ell} M^i F,$$

where F is the vector of initial probabilities.

The ℓ-finite state probability $\mathbb{P}^\ell[i]$ is proportional to the mean number of times a given state i is traversed.

Notice that we thus provide a way to colorize the dynamical state graph by assigning to each state i a colour value proportional to $\mathbb{P}^\ell[i]$. For long trajectories, when ℓ is approximately the number of states in the graph, the states supposedly most crucial to the biological system's behaviour are emphasized. In an automated approach, we use vector $\mathbb{P}^\ell$ to prune the dynamical graph by extracting the induced subgraphs composed of states i such that $\mathbb{P}^\ell[i] > 2/N$. Each subgraph identified consists of states reached at most twice in long trajectories. In the case of

E. coli response to carbon deprivation, this cut off threshold of $2/N$ ensures that the subgraphs obtained are tractable for any further analysis.

3 Extending the discrete model paradigm with delays: the hybrid model

3.1 Clocks and delays

The evolution of the expression of a given gene is a continuous non-linear process (see Figure 3). This fact is not taken into account in the discrete modelling formalism of Thomas, where gene expression evolves from one level to another level in a discrete fashion (see Figure 3 (b)). In the field of biological modelling, paradigms have been proposed to simulate continuous temporal evolution (Bernot *et al.*, 2004; Adelaïde *et al.*, 2004; Siebert *et al.*, 2006). The refinement of discrete modelling by a more enhanced formalism of hybrid modelling has been proposed (Ahmad *et al.*, 2007), in which the sigmoid-like evolution is no more approximated by a discrete step but by a piece-wise linear curve instead (Figure 3 (c)). Since we now consider that the *delay* needed for a gene to evolve from expression level a to $a+1$ or $a-1$ is not null, we have to deal with additional concepts, namely *time intervals* and *clocks*.

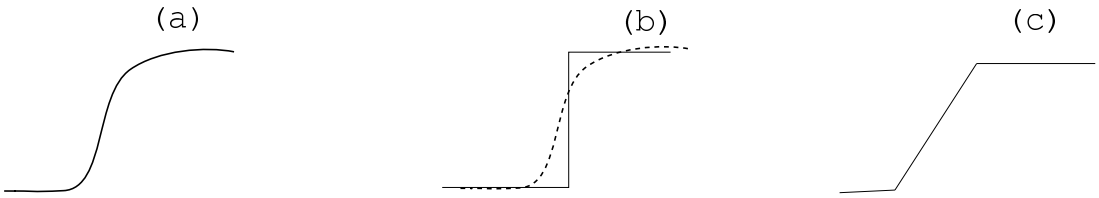


Figure 3: A sigmoid relation (a) and its discrete (b) and piece-wise linear (c) approximations.

The widely-spread timed automaton formalism provides a formal framework to describe hybrid systems (Alur *et al.*, 1994). In this framework, any global state of the system modeled is described by a discrete spacial location (in our case, the vector of current gene expression values) and a vector of continuous variables, called clocks. Any gene u is associated to a clock (denoted h_u). Evolving synchronously with time, clock intervals therefore superimpose a representation of continuous system's dynamics on the already defined discrete dynamics. The clocks act as transition guards and are reset to 0 when the system passes from one discrete location to another one. The more general class of Linear Hybrid Automata (LHA) is the appropriate framework allowing the definition of time interval associated to a clock (Henzinger *et al.*, 1995). For any clock, its current value measures the time elapsed since the most recent change occurred for the system, in the discrete space of gene expressions. Thus, if the system consists of n genes, an LHA formalism superimposes a temporal hypercube of dimension n to the discrete global state space. For illustration, in dimension 2, a global state is now associated to a rectangular temporal region bounded by four delays (see Figure 4). The delay for gene u to increase up to next discrete

level is a real parameter depending on u's current discrete state ($d_u{}^+ > 0$); symmetricaly, the delay to decrease down to next discrete level is $\mid d_u{}^- \mid$ ($d_u{}^- < 0$).

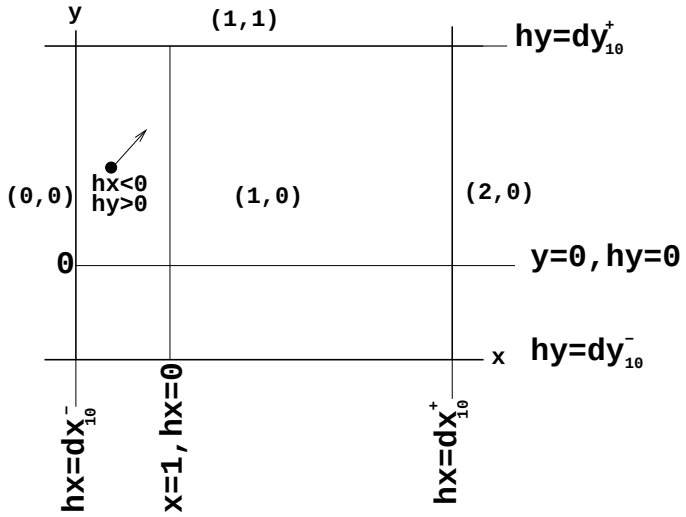


Figure 4: Hybrid model - temporal regions and delays -. Here, global state $(x, y) = (1, 0)$ is associated to delays $d_{x_{10}}^+$, $d_{x_{10}}^-$, $d_{y_{10}}^+$ and $d_{y_{10}}^-$.

To model time elapsing, we use a subclass of the LHA formalism, which associates to each gene u a clock rate, $\dot{h_u}$, in the restricted set $\{-1, 0, 1\}$. Rates -1, 0 and 1 respectively signify that gene expression level is decreasing, staying at the same level or increasing. Any clock rate $\dot{h_u}$ related to gene u indicates the evolution tendency for this gene, with respect to the current global state. Prior to any analysis of the dynamical behaviour of the modeled system, each such clock rate must be tuned. In contrast with the asynchronous discrete model of Thomas, tuning the clock rates now requires looking several steps ahead in the dynamical graph, in order to capture the whole actual tendency. For instance, in Figure 2 (c), if one confines to a depth of 2 to examine next transitions, when starting from state $(0, 1)$, as u decreases and v increases, $\dot{h_u}$ and $\dot{h_v}$ are respectively set to 1 and -1 (see (Ahmad *et al.*, 2007) for details).

Delays being considered as parameters, such a model will be called a Parametric Linear Hybrid Automaton (PLHA) in the sequel. Now all concepts have been unformally introduced and illustrated, next subsection will rigorously define PLHAs together with their semantics.

3.2 Parametric Linear Hybrid Automata

We remind the reader that derivative $\dot{x}$ denotes the evolution rate for protein concentration x while $\dot{h_x}$ is the evolution rate of the clock h_x associated with variable x.

Notation 2
Let X and P be respectively a set of real variables and a set of parameters. An atomic constraint

is a formula of the form $x \bowtie c$, for $x \in X$, $c \in \mathbb{Q} \cup P$ and $\bowtie \in \{<, \leq, \geq, >\}$. We denote $\mathcal{C}(X, P)$ the set of constraints *over a set of variables X and parameters P, which consists of conjunctions of atomic constraints. Given a constraint g, we let $\mathcal{V}(g)$ be the set of variables that appear in g. We let $\mathcal{C}^{=}(X, P)$ (resp. $\mathcal{C}^{\leq}(X, P)$, $\mathcal{C}^{\geq}(X, P)$) be the set of constraints using only $=$ (resp. $\leq$, $\geq$).*

Definition 7 (PLHA**)**
A PLHA *is a tuple $(L, \ell_0, X, P, E, Inv, Dif)$ defined as follows:*

- *L is a finite set of locations*
- *$\ell_0 \in L$ is the initial location*
- *P is a finite set of delay parameters*
- *X is a finite set of clocks*
- *$E \subseteq L \times \mathcal{C}^{=}(X, P) \times 2^X \times L$ is a finite set of* edges*, $e = (\ell, g, R, \ell') \in E$ represents an edge from location ℓ to location ℓ', associated with the guard g and the reset set $R \subseteq X$ (we require that $\mathcal{V}(g) \subseteq R$)*
- *$Inv : L \to \mathcal{C}^{\leq}(X, P) \cup \mathcal{C}^{\geq}(X, P)$ assigns an invariant to any location*
- *$Dif : L \times X \to \{-1, 0, 1\}$ maps each pair (ℓ, x) to an evolution rate.*

The semantics of a PLHA is a timed transition system. It is defined according to the time domain $\mathbb{T}$. We let $\mathbb{T}^* = \mathbb{T} \setminus \{0\}$.

Definition 8 (Semantics of a PLHA**)**
Let γ be a valuation for the parameters P. The $(\mathbb{T}, \gamma)$–semantics of a parametric LHA *$H = (L, \ell_0, X, P, E, Inv, Dif)$ is defined as a timed transition system $S_H = (S, s_0, \mathbb{T}, \to)$ where: (1) $S = \{(\ell, \nu) \mid \ell \in L \text{ and } \nu \models Inv(\ell)\}$; (2) $s_0 = (\ell_0, \nu_0)$ with $\nu_0(x) = 0$ for every $x \in X$; (3) the relation $\to \subseteq S \times \mathbb{T} \times S$ is defined for $t \in \mathbb{T}$ as:*

- *discrete transitions: $(\ell, \nu) \xrightarrow{0} (\ell', \nu')$ iff $\exists (\ell, g, R, \ell') \in E$ such that $\gamma(\nu) = true$, $\nu'(x) = 0$ if $x \in R$ and $\nu'(x) = \nu(x)$ if $x \notin R$.*
- *continuous transitions: For $t \in \mathbb{T}^*$, $(\ell, \nu) \xrightarrow{t} (\ell', \nu')$ iff $\ell' = \ell$, $\nu'(x) = \nu(x) + Dif(\ell, x) \times t$, and for every $t' \in [0, t]$, $(\nu(x) + Dif(\ell, x) \times t') \models Inv(\ell)$.*

The semantics of a PLHA implements two types of transitions: discrete and continuous. Invariants and guards are constraints set on subsets of clocks. Invariants specify the conditions under which the system is allowed to stay in the current state, while time elapses. A *discrete transition* is an instantaneous transition that occurs between two discrete locations. It is fired when the associated guard is satisfied. *Continuous transitions* account for elapsing of time in a discrete location until the associated invariant condition is no more satisfied. A continuous transition allows the updating of the clocks in any time interval $[0, t]$, according to the evolution rates specified for the clocks and provided that the invariant conditions are still verified. We refer the reader to appendix 1 for the formal definition of the semantics of PLHAs.

Example 2 (PLHA)
The Parametric Linear Hybrid Automaton of the example of P. aeruginosa (see Figure 2) is shown in Figure 5. Here, the delays are represented by the notation $d^{\alpha}_{i,\ell}$, where α denotes the delay sign ($+$ for activation and $-$ for inhibition) of a gene i in a location ℓ. This automaton has six locations. The locations are labelled with the invariant conditions while the discrete transitions are labelled with guards and clock resets.

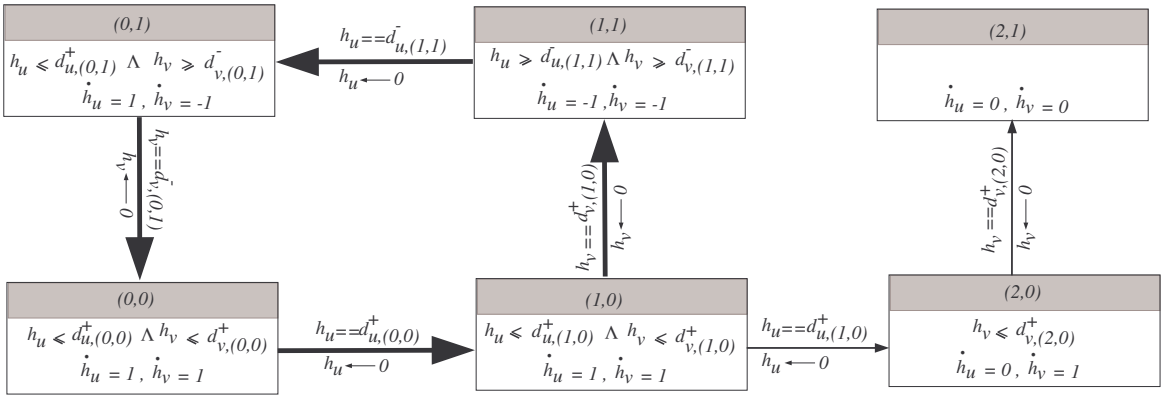


Figure 5: Hybrid model for *P. aeruginosa* mucus production.

3.3 Automatic symbolic analysis of a PLHA through HyTech model-checker

HyTech is the model-checker chosen in our study (Henzinger *et al.*, 1997). It is adapted to hybrid systems: it has the ability to manage parameters through synthesizing constraints relative to these parameters, thus satisfying necessary conditions for the existence of the behaviours analysed.

Definition 9 (Trajectories and cycles)
A trajectory is a sequence of states related by discrete and continuous transitions. A cycle is a trajectory that starts in a given location and returns to this same location further on.

In the hybrid model of a GRN, we respectively denote $\varphi(t)$ for $t \in \mathbb{R}_{\geq 0}$ and $\mathbb{S}$ the sequence of points of a trajectory and the set of all points in its state space.

Definition 10 (Invariance kernel)
A trajectory $\varphi(t)$ is viable in $\mathbb{S}$ if $\varphi(t) \in \mathbb{S}$ for all $t \geq 0$. A subset K of $\mathbb{S}$ is said to be invariant if for any point $p \in K$, a trajectory starting in p is viable in K. An invariance kernel *K is the largest invariant subset of $\mathbb{S}$.*

For illustration, the set of constraints displayed in Table 4 characterizes the invariance kernel of the example relative to *P. aeruginosa* (see Figure 2). For the sake of simplicity, we only deal

with few delay parameters, assuming that all d_{ij}^{α} are equal, whatever the actual value of j, and similarly for all d_{ij}^{α}, whatever the actual value of i.

$$
\begin{aligned}
& d_{x_0}^{+} + d_{x_1}^{-} + d_{y_1}^{-} \leq d_{y_0}^{+} \\
\wedge\ & d_{y_0}^{+} + d_{x_1}^{-} \leq d_{x_0}^{+} + 2d_{x_1}^{+} + d_{y_1}^{-} \\
\wedge\ & d_{x_1}^{-} \leq d_{x_0}^{+} + d_{y_0}^{+} + d_{y_1}^{-} \\
\wedge\ & d_{y_0}^{+} + d_{y_1}^{-} \leq d_{x_0}^{+} + d_{x_1}^{-} \\
\wedge\ & d_{x_0}^{+} + d_{y_1}^{-} \leq d_{y_0}^{+} + d_{x_1}^{-}
\end{aligned}
$$

Table 4: Delay constraints characterizing the invariance kernel of *P. aeruginosa*.

4 The pipelined process applied to the analysis of the reaction of $E.$ *coli* to carbon availability

We recall the reader that the pipeline process implemented schedules the following tasks: (i) conversion from a PADE model to a model with attractors, (ii) identification of the corresponding transition graph, (iii) identification of induced subgraphs of interest in the former graph, implemented through a probabilistic approach, (iv) modelization of subgraphs in the framework of PLHA formalism, (v) analysis of characterized dynamical behaviours through $HyTech$ model-checker. Note that third step actually provides a visualization tool able to point out subgraphs containing global states of interest.

It must be emphasized that our contribution is the first example of an application of timed-model checking techniques on the case of *E. coli* regulation related to carbon availability. Indeed, the former works relative to this GRN did not take into account the concept of delays (Batt *et al.* 2005).

The implementation of the protocol previously described provides multiple significant results in the case of *E. coli* response to carbon availability.

4.1 The PADE model of the carbon starvation response in *Escherichia Coli*

The growth of bacterial populations is related to the quantity of nutrients present in their environment. Nutrient availability entails an exponential increase of the prokaryotic biomass whereas nutritional stress induces growth deceleration or even growth stop. Thus, bacterial populations are subject to transitions between two states denoted as exponential and stationary phases re-

spectively. The switch between these two phases is crucial to bacterial survival and is controlled by a GRN that integrates various environmental signals.

The GRN controlling the response to carbon deprivation has been widely studied *E. Coli*, in the past decades. In contrast to most studies focusing on only one or a few components of this network, Ropers and co-authors'recent contribution implemented the modelling of concentration evolution for six key global regulators of this network (Ropers *et al.*, 2006). This model relates the behaviours of five genes (crp, cya, fis, gyr_{AB}, $topA$) and two supplementary "signals" such as the carbon starvation information and the quantity of stable RNAs. The reader interested in details about the biological hypotheses used for describing the genetic interactions is referred to (Ropers *et al.*, 2006).

The PADE model of Ropers and co-workers was shown to fit to typical features describing the transition between bacterial growth phases. We therefore admit that this model was validated and we used a slightly simplified version as a starting point to establish a more refined modelling approach based on attractors and delays. The simplified version of the PADE model adapted from Ropers and co-workers'model is shown in Table 5. Herein, we present the simplified equations together with their associated constraints. As the variable x_{rrn} corresponding to stable RNAs had no influence on others variables, it was discarded from our own PADE model. In addition, we dismissed three thresholds, $\theta 3_{crp}$, $\theta 3_{cya}$ and $\theta 5_{fis}$, which appeared to be useless. Thus, contraints applying to $\theta 3_{crp}$ now apply to $\theta 2_{crp}$; similarly, $\theta 2_{cya}$ is constrained as was $\theta 3_{cya}$; finally, parameter inequalities relative to the former $\theta 5_{fis}$ now apply to $\theta 4_{fis}$.

4.1.1 Conversion of the PADE model relative to *E. coli* response to nutrient availability into a discrete model with attractors

Benefitting from a previous PADE modelling of *E. coli* response to carbon availability, we are thus able to skip the tedious task of identifying attractors *ab initio*. Moreover, the instanciation is facilitated by the set of constraints associated with the PADE model. Indeed, provided that we understand how to relate the kinetic parameters and the degradation rate of the PADE system with attractor values, the instanciation process will be significantly simplified. Table 6 focuses on an excerpt of Ropers and co-authors'model (in its simplified version).

The equation in line (1) models the variation of x_{topA}, that is the concentration of topoisomerase. For the sake of simplicity, Ropers and co-authors considered that a single promoter is involved in the expression of $topA$ gene, whereas there are indeed five promoters involved in its expression. The expression of this gene is also controlled by antagonistic agents: $topA$ is activated by a low level of fis; in contrast, it is activated by a high level of gyr_{AB}. Two different thresholds have been considered in the simplified version, $\theta 1_{topA}$ and $\theta 2_{topA}$. Stimulation of $topA$ promoter by its resource $\{gyr_{AB}, \overline{fis}, \overline{topA}\}$, where the first gene is activating and the two others are not inhibiting, entails maximal production of $topA$. It follows that $\theta 2_{topA} < \frac{k_{topA}}{\gamma_{topA}} < max_{topA}$.

$\dot{u}_s = 0$
$\dot{x}_{crp} = k1_{crp} + k2_{crp}\ s^-(x_{fis}, \theta2_{fis})\ s^+(x_{cya}, \theta1_{cya})\ s^+(u_s, \theta_s) + k3_{crp}\ s^-(x_{fis}, \theta1_{fis}) - \gamma_{crp}\ x_{crp}$ $0 < \theta1_{crp} < \theta2_{crp} < max_{crp}$ $\theta1_{crp} < \frac{k1_{crp}}{\gamma_{crp}} < \theta2_{crp}$ $\theta1_{crp} < \frac{k1_{crp}+k2_{crp}}{\gamma_{crp}} < \theta2_{crp}$ $\theta2_{crp} < \frac{k1_{crp}+k3_{crp}}{\gamma_{crp}} < max_{crp}$ $\theta2_{crp} < \frac{k1_{crp}+k2_{crp}+k3_{crp}}{\gamma_{crp}} < max_{crp}$
$\dot{x}_{cya} = k1_{cya} + k2_{cya}\ (1 - s^+(x_{crp}, \theta2_{crp})\ s^+(x_{cya}, \theta2_{cya})\ s^+(u_s, \theta_s)) - \gamma_{cya}\ x_{cya}$ $0 < \theta1_{cya} < \theta2_{cya} < max_{cya}$ $\theta1_{cya} < \frac{k1_{cya}}{\gamma_{cya}} < \theta2_{cya}$ $\theta2_{cya} < \frac{k1_{cya}+k2_{cya}}{\gamma_{cya}} < max_{cya}$
$\dot{x}_{fis} = k1_{fis}\ (1 - s^+(x_{crp}, \theta1_{crp})\ s^+(x_{cya}, \theta1_{cya})\ s^+(u_s, \theta_s))\ s^-(x_{fis}, \theta4_{fis})$ $+k2_{fis}\ s^+(x_{gyr_{AB}}, \theta1_{gyr_{AB}})\ s^-(x_{topA}, \theta2_{topA})\ s^-(x_{fis}, \theta4_{fis})\ (1 - s^+(x_{crp}, \theta1_{crp})\ s^+(x_{cya}, \theta1_{cya})\ s^+(u_s, \theta_s)) - \gamma_{fis}x_{fis}$ $0 < \theta1_{fis} < \theta2_{fis} < \theta3_{fis} < \theta4_{fis} < max_{fis}$ $\theta1_{fis} < \frac{k1_{fis}}{\gamma_{fis}} < \theta2_{fis}$ $\theta4_{fis} < \frac{k1_{fis}+k2_{fis}}{\gamma_{fis}} < max_{fis}$
$\dot{x}_{gyr_{AB}} = k_{gyr_{AB}}(1 - s^+(x_{gyr_{AB}}, \theta2_{gyr_{AB}})\ s^-(x_{topA}, \theta1_{topA}))\ s^-(x_{fis}, \theta3_{fis}) - \gamma_{gyr_{AB}}\ x_{gyr_{AB}}$ $0 < \theta1_{gyr_{AB}} < \theta2_{gyr_{AB}} < max_{gyr_{AB}}$ $\theta2_{gyr_{AB}} < \frac{k_{gyr_{AB}}}{\gamma_{gyr_{AB}}} < max_{gyr_{AB}}$
$\dot{x}_{topA} = k_{topA}\ s^+(x_{gyr_{AB}}, \theta2_{gyr_{AB}})\ s^-(x_{topA}, \theta1_{topA})\ s^-(x_{fis}, \theta3_{fis}) - \gamma_{topA}\ x_{topA}$ $0 < \theta1_{topA} < \theta2_{topA} < max_{topA}$ $\theta2_{topA} < \frac{k_{topA}}{\gamma_{topA}} < max_{topA}$

Table 5: Equations and associated constraints depicting the simplified model adapted from Ropers and co-authors, to simulate the response to carbon deprivation in *Escherichia coli*. The five variables correspond to protein concentrations: x_{crp} (CRP), x_{cya} (Cya), x_{fis} (Fis), $x_{gyr_{AB}}$ (GyrAB), x_{topA} (TopA).

Applying this process to each equation in the PADE system of Table 5, we finally obtain a discrete model with instanciated attractor values. As explained in subsection 1.2, the construction of the dynamical graph is now straightforward. However, as foreseeable for such a complex GRN as *E.coli* reponse to nutrient availability, before behavioural property inference may be performed through model-checking techniques, a simplification stage is required. For example, the dynamical graph corresponding to *E. coli* response to nutrient availability contains such a high number N of vertices (*i.e.* states) as 810.

(1) $\dot{x}_{topA} = k_{topA}\ s^+(x_{gyr_{AB}}, \theta 2_{gyr_{AB}})\ s^-(x_{topA}, \theta 1_{topA})\ s^-(x_{fis}, \theta 3_{fis}) - \gamma_{topA}\ x_{topA}$

(2) $0 < \theta 1_{topA} < \theta 2_{topA} < max_{topA}$

(3) $\theta 2_{topA} < \frac{k_{topA}}{\gamma_{topA}} < max_{topA}$

(4) $\mathcal{K}_{topA,\{gyr_{AB},\overline{fis},\overline{topA}\}} = \mathscr{D}_{topA}(\frac{k_{topA}}{\gamma_{topA}})$, $\mathcal{K}_{topA,\{\}} = \mathscr{D}_{topA}(\frac{0}{\gamma_{topA}})$

(5) $\mathcal{K}_{topA,\{\}} = 0$

(6) $\theta 1_{topA} = 1,\ \theta 2_{topA} = 2$

(7) $\mathcal{K}_{topA,\{gyr_{AB},\overline{fis},\overline{topA}\}} = 2$

Table 6: Identification of resources and tuning of attractors for the response to carbone starvation in *E. coli*. The differential equation of Ropers'model (1) allows the identification of the two attractors concerned (4). $\mathscr{D}_{topA}$ is a function used to obtain an integer attractor value (discretization). Attractor $\mathcal{K}_{topA,\{\}}$ (5), corresponding to the case when no resource is available, is trivially instanciated with the value of 0; together with conversion rule (3), Ropers's constraints (2) induce the instanciation of concentration thresholds (6); finally, a value of 2 is an instanciation of $\mathcal{K}_{topA,\{gyr_{AB},\overline{fis},\overline{topA}\}}$ attractor's value consistent with (3), (4) and (6) constraints.

4.2 The initial dynamical graph

The entire transition graph contains 810 global states and 3827 transitions. However, the dynamics of the exponential phase and that of the stationary phase are to be studied separately. Indeed, our purpose here is not to focus on transitions switching from one phase to the other one. The graph describing the dynamics of the stationary phase consists of 405 global states and 1523 transitions. The graph corresponding to the exponential phase contains 405 states and 1494 transitions. After conversion from the PADE model into a discrete model with attractors, we dismissed some states known to be never encountered ($crp = 0$). In this report, we chose to concentrate on the exponential phase. The reduced graph describing the dynamics of the exponential phase consists of 108 global states.

Incidentally, we checked that some specific properties reported in the literature hold for the model inferred. For example, the crp/fis antagonism ($crp = 2\ and\ fis = 0$) is verified as expected. Besides, it has been checked that DNA supercoiling is absent from *every cycle* belonging to the graph characterizing the exponential phase: $fis = 0 \implies topA > gyr_{AB}$. Indeed, two mechanisms were described by Travers *et al.* to explain how the nucleoid-associated protein FIS

modulates the topology of DNA in a growth-phase dependent manner, to counteract excessive levels of superhelicity (Travers *et al.*, 2001). First, the binding of FIS to DNA constrains negative superhelicity to low levels; second, a reduction in the expression and effectiveness of DNA gyrase achieves the same result. Conversely, high fis expression levels do themselves require a high negative superhelical density.

4.3 Extraction of a characterized cycle

When applying the "coloration" process to the graph related to exponential phase, we identify the subgraph depicted in Figure 6. This subgraph is outstandingly dense in states of interest (*i.e. potentially frequently encountered states in long trajectories*) and therefore contains several qualitative cycles, among which we recognize a cycle well-known in *E. coli* response to carbon availability:

$\mathtt{012\underline{10}0} \rightarrow \mathtt{012\underline{11}0} \rightarrow \mathtt{012\underline{12}0} \rightarrow \mathtt{012\underline{22}0} \rightarrow \mathtt{012\underline{32}0} \rightarrow \mathtt{012\underline{42}0} \rightarrow \mathtt{012\underline{41}0} \rightarrow$
$\mathtt{012\underline{40}0} \rightarrow \mathtt{012\underline{30}0} \rightarrow \mathtt{012\underline{20}0} \rightarrow \mathtt{012\underline{10}0}$

(the six values respectively correspond to crp, cya, fis, gyr_{AB}, $topA$ and rrn). Interestingly, it happens that this cycle corresponds to the one identified by Ropers and co-workers (Ropers *et al.*, 2006), displayed in Figure 7, except that only 4 levels are considered for fis.

Moreover, $HyTech$ model-checking techniques enable us to capture this cycle through the analysis of *invariance kernel* in the hybrid model built for the subgraph of Figure 6. As stated in the study based on PADE modelling (Ropers *et al.*, 2006), we show that the system behaviour is likely to end running into the qualitative cycle aforementioned. At this stage, it is remarkable that a graph pruning probabilistic process combined with hybrid modelling on the one hand and PADE modelling on the other hand meet to reveal the very same qualitative cycle.

Before commentating on the results obtained through $HyTech$ analysis, we have to define the so-called *full period* (denoted $\pi(u)$) as the sum of all delays for a gene u to pass sequentially and successively, once through each of all its expression levels.

It should be noticed that the real time for a gene to run along this route (if it actually takes place) may be greater than the full period since it may include lazy stages, *i.e.* some time intervals where there is neither increase nor decrease.

4.3.1 Identification of temporal constraints

Restraining to fis and gyr_{AB}, the cycle aforementioned is merely expressed as

$$\overset{++}{1}0 \rightarrow \overset{++}{1}1 \rightarrow \overset{+-}{1}2 \rightarrow \overset{+-}{2}2 \rightarrow \overset{+-}{3}2 \rightarrow \overset{--}{4}2 \rightarrow \overset{--}{4}1 \rightarrow \overset{-=}{4}0 \rightarrow \overset{-+}{3}0 \rightarrow \overset{-+}{2}0 \rightarrow \overset{++}{1}0$$

where symbols $+$, $-$ and $=$ indicate the evolution tendency for each gene.

The analysis with $HyTech$ provides two kinds of results relative to this peculiar cycle. On the one hand, constraints are identified, which determine a cyclic behaviour (inequalities

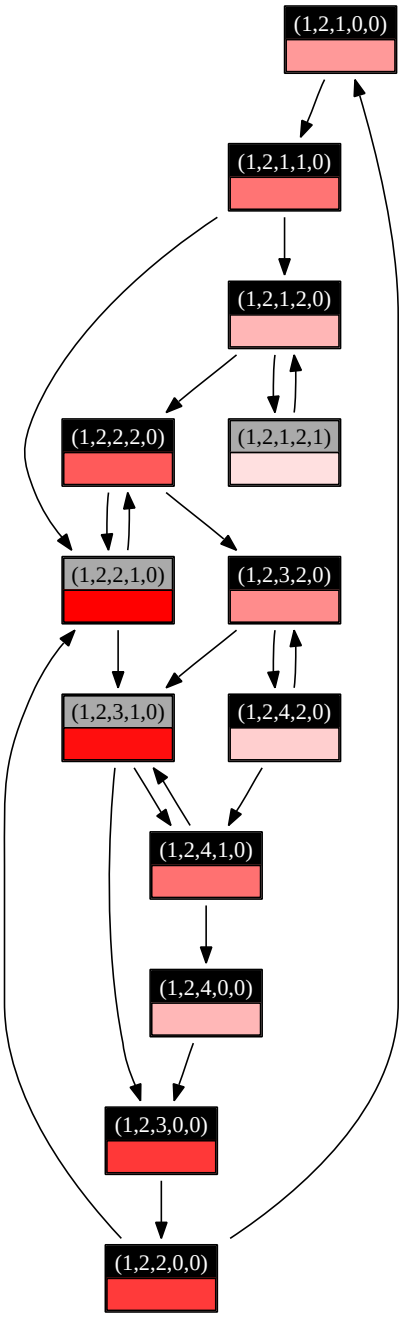


Figure 6: A subgraph showing cycles in the exponential phase, for *E. coli* bacterium. The global states are represented in the same manner as in Figure 5. The five values respectively correspond to crp, cya, fis, gyr_{AB}, and $topA$. The states of Ropers and co-workers'cycle are highlighted as dark rectangles.

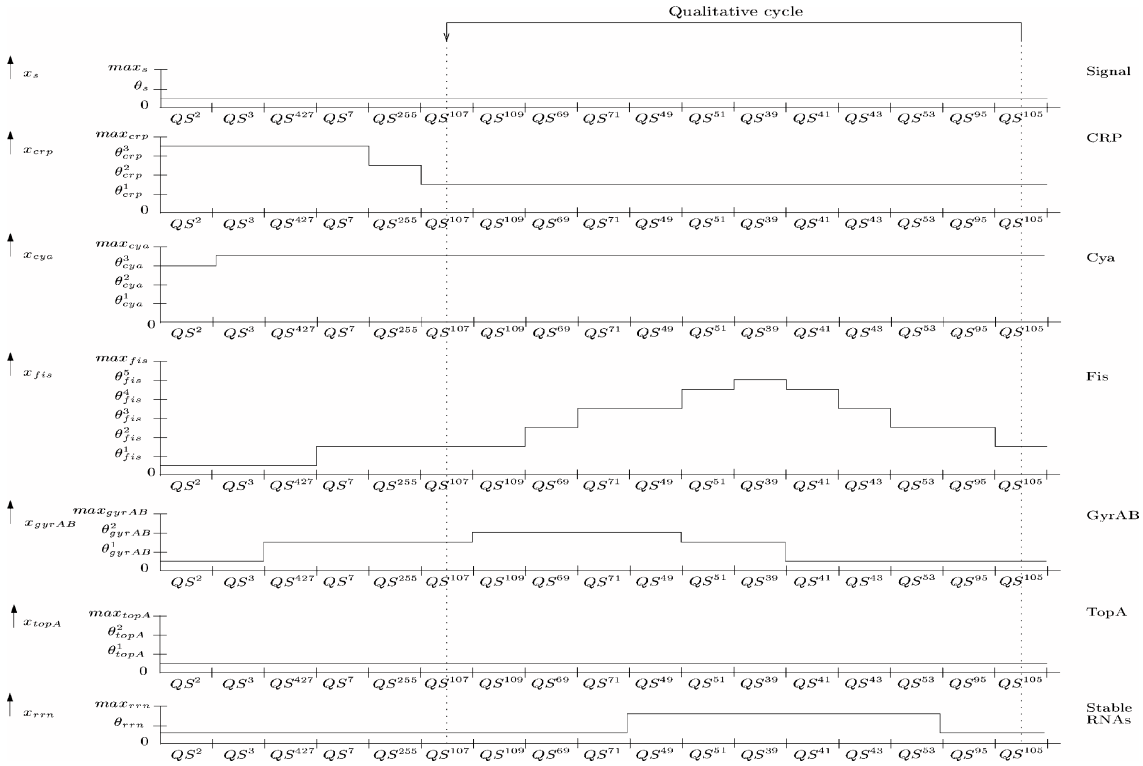


Figure 7: Qualitative cycle of *E. Coli* associated with the series of phases denoted $\mathtt{Q}_\mathtt{s}^{107}, \mathtt{Q}_\mathtt{s}^{109}, \mathtt{Q}_\mathtt{s}^{69}, \mathtt{Q}_\mathtt{s}^{71}, \mathtt{Q}_\mathtt{s}^{49}, \mathtt{Q}_\mathtt{s}^{51}, \mathtt{Q}_\mathtt{s}^{39}, \mathtt{Q}_\mathtt{s}^{41}, \mathtt{Q}_\mathtt{s}^{49}, \mathtt{Q}_\mathtt{s}^{43}, \mathtt{Q}_\mathtt{s}^{53}, \mathtt{Q}_\mathtt{s}^{95}, \mathtt{Q}_\mathtt{s}^{105}$ © (Ropers, 2006).

registered (1) to (3), in Table 7). On the other hand, we exhibit a relation between the length L of this cycle and the delays associated with genes (equalities 4 (a) and 4 (b), in Table 7).

4.3.2 Interpretation and relevance with regard to biological evidences

- First, it follows directly from (3) that $\pi(gyr_{AB}) \leq \pi(fis)$.

  This inequality is explained by the fact that there exists a phase in the cycle where gyr_{AB} is *lazy* (*i.e.* it stays at the same expression level). This may be observed in the phase $\overline{\overline{40}}$ of the cycle, corresponding to the state $\mathtt{Q}_\mathtt{s}^{41}$ in figure 7. Thus, the qualitative cycle in which *E. coli* bacterium is involved during the exponential phase following a carbon supply is possible when gyr_{AB} qualitative period is smaller than that of fis: therefore, the path of transitions through minimum to maximum qualitative state and back is traversed faster for gyr_{AB} than for fis. This remark implies that a slight increase of gyr_{AB}'s period might not allow the bacterium to stay in the exponential phase. gyr_{AB} is closely related to DNA supercoiling. Therefore, slowing down the gyr_{AB} cycle running would entail diminishing DNA supercoiling reactivity. Independent studies have shown the great

$$d^+_{fis_1} + d^+_{fis_2} + d^+_{fis_3} + |d^-_{fis_3}| + |d^-_{fis_2}| \leq d^+_{gyr_{AB_0}} + d^+_{gyr_{AB_1}} + |d^-_{gyr_{AB_2}}|$$

$$d^+_{gyr_{AB_0}} + d^+_{gyr_{AB_0}} \leq d^+_{fis_1} + |d^-_{fis_2}| + |d^-_{fis_3}|$$

$$d^+_{gyr_{AB_0}} + d^+_{gyr_{AB_1}} + |d^-_{gyr_{AB_2}}| + |d^-_{gyr_{AB_1}}| \leq d^+_{fis_1} + d^+_{fis_2} + d^+_{fis_3} + |d^-_{fis_4}| + |d^-_{fis_3}| + |d^-_{fis_2}|$$

$$L = d^+_{fis_1} + d^+_{fis_2} + d^+_{fis_3} + |d^-_{fis_4}| + |d^-_{fis_3}| + |d^-_{fis_2}|$$

$$L = d^+_{gyr_{AB_0}} + d^+_{fis_1} + d^+_{fis_2} + d^+_{fis_3} + |d^-_{fis_4}|.$$

Table 7: Identification of temporal constraints associated with the existence of the cycle highlighted in Figure 6.

impact of the DNA-superhelicity on the bacterial gene activity (see (Hatfield *et al.*, 2002)) for a review). Basal expressions of genes are low when chromosomal superhelical density is low, and conversely. Because of the necessity to react to environmental variation for survival, low bacterial activity acts as a trigger for switching to the stationary phase. This remark confirms the temporal constraint mentioned above as a biological insight. D. Ropers and co-workers depicted this qualitative cycle as an unexpected result. However, our investigation of temporal properties associated with this cycle points out insights that are relevant with biological evidences about DNA supercoiling.

- Second, we deduce from (4) (a) that $L = \pi(fis)$ (and hence $L \geq \pi(gyr_{AB})$.)

  Thus, we are able to prove that the cycle length is exactly the full period of fis. This result is consistent with the fact that there is no lazy phase for fis. Moreover, this observation implies that fis plays the major role in the qualitative cycle in which the bacterium is kept during the exponential phase. Therefore, an experimental calibration of fis temporal properties might shed light on this bacterial model. Furthermore, point ii. shows that the reactivity of DNA supercoiling mentioned above is related to the delay taken by fis to complete its qualitative period. Again, the temporal properties deduced from the qualitative model reinforce the biological relevance of the model.

- Finally, (4) (a) and (4) (b) entail that $d^+_{gyr_{AB_1}} = |d^-_{fis_3}| + |d^-_{fis_2}|$.

  Equality (iii) indicates that, in the sequence $\mathtt{Q}_\mathtt{s}^{43} - \mathtt{Q}_\mathtt{s}^{53} - \mathtt{Q}_\mathtt{s}^{95}$ of figure 7, while fis decreases from level `3` to level `1` (within the time delay $|d^-_{fis_3}| + |d^-_{fis_2}|$), in the same time, gyr_{AB} increases from level `0` to level `1` (within the delay $d^+_{gyr_{AB_0}}$), which means that, in the $\mathtt{Q}_\mathtt{s}^{105}$ phase, gyr_{AB} should be at level `1`, as it is in phase $\mathtt{Q}_\mathtt{s}^{107}$. This property is particularly hard to verify by experimental means. However, the nice consequence of this

observation is that the so-called “Qualitative cycle” of D. Ropers in Figure 7 is actually a cycle. In this case, analyzing the temporal properties associated with the qualitative model reinforces previous computational investigations.

5 Conclusion

In this document, we have presented a complete process devoted to infer behavioural properties of realistic GRNs. As a conclusion to former research works, some of the co-authors concluded that hybrid modelling including linear delays as an approximation constitutes a valuable refinement with respect to the initial model of Thomas (Ahmad *et al.*, 2007). It was announced that the modelization of a GRN related to *E. coli* was under investigation at the same time. The work reported here dealt with the methodological analysis led to tackle the case of *E. coli*’s response to carbon starvation, in particular.

As predictable, the difficulties encountered during our study lied in the high dimension of the associated discrete dynamical graph. A first trick consisted in benefitting from the tuning of a former published model, itself settled on solid biological grounds, to avoid tedious identification of resource sets and facilitate the instanciation of attractor values. On the *E. coli* benchmark, we have shown that it is possible to convert a PADE model into a model with attractors. Then, a graph coloration method based on probabilistic reasoning allowed us to focus on subgraphs dense in presumed states of interest. Applying such a coloration method to provide subgraphs tractable by such model-checkers as $HyTech$ might be an attractive solution to tackle the analysis of large GRNs.

As a remarkable result, not only did the coloration method described point out a cycle already reported in biological literature, model-checking performed on the hybrid model also captured this cycle. Besides, our approach allowed to refine the temporal constraints that are necessary to reach particular qualitative transitions, such those of interest observed by Ropers and collaborators. Thus, beyond simple verification performance, interesting relations between delays have been inferred through our formalism. They enable further investigations that lead on to future experiments or novel biological insights about the mechanisms responsible for specific dynamical behaviours.

Finally, the methodological investigation conducted on *E. coli* system constitutes a first valuable contribution to show the relevance of pipelining different methods to tackle large biological system analysis.

Acknowledgement

This project was supported by the AtlanSTIC Research Cluster CNRS FR2819.

Authors' contributions

CS, OR and JB initiated the collaboration between two laboratories of the AtlanSTIC Research Cluster. All co-authors participated in the design of the study and contributed to the methodological investigation. CS induced the application to the analysis of a realistic (large) GRN. DE selected the model of *E. coli* response to carbon availability, on the basis of previous studies. JF and DE carried out the conversion of the PADE model of *E. coli* response to carbon availability into a discrete model with attractors. JB designed and ran the probabilistic analysis of the dynamical graph, to produce tractable subgraphs. JA provided a hybrid model for one of the subgraphs of interest and analysed it through $HyTech$ model-checker. Results obtained through model-checking were thoroughly analysed and commentated by OR, JA and JF. DE brought his biological expertise to interpret results. All co-authors brought their contribution in writing the manuscript and CS integrated these various contributions in the manuscript.

References

Adélaïde, M., Sutre, G., 2004. Parametric analysis and abstraction of genetic regulatory networks. Proc. 2nd Workshop on concurrent models in molecular biology, BioCONCUR'04, Electronic Notes in Theor. Comp. Sci. Amsterdam, Elsevier.

Ahmad, J., Bernot, G., Comet, J.-P., Lime, D., Roux, O., 2007. Hybrid modelling and dynamical analysis of gene regulatory networks with delays. ComPlexUs, Karger Publisher. 3(4), 231–251.

Alur, R., Dill, D.L., 1994. A theory of timed automata. Theor. Comput. Sci. 126, 183–235.

Batt, G., Ropers, D., de Jong, H., Geiselmann, J., Mateescu, R., Page, M., Schneider, D., 2005. Validation of qualitative models of genetic regulatory networks by model checking: analysis of the nutritional stress response in *Escherichia coli*. Bioinformatics. 21(Suppl 1), i19–i28.

Bernot, G., Comet, J.-P., Richard, A., Guespin, J., 2004. Application of formal methods to biological regulatory networks: extending Thomas' asynchronous logical approach with temporal logic. J. Theor. Biol. 229(3), 339–347.

Chaouiya C., 2007. Petri net modelling of biological networks. Brief. Bioinform. 8(4), 210–9.

de Jong, H., 2002. Modeling and simulation of genetic regulatory systems: a literature review. J. Comput. Biol. 9(1), 67–103, doi: 10.1089/10665270252833208.

de Jong, H., Gouzé, J.L., Hernandez, C., Page, M., Sari, T., Geiselmann, J., 2004. Qualitative simulation of genetic regulatory networks using piecewise-linear models. Bull. Math. Biol. 66(2), 301–340.

de Jong, H., Page, M., Hernandez, C., Geiselmann, J., 2001. Qualitative simulation of genetic regulatory networks : method and application. Proc. of the Seventeenth International Joint Conference on Artificial Intelligence, IJCAI'01, B. Nebel (ed.), Morgan Kaufmann, San

Mateo, CA. 67–73.

Glass, L., Kauffman, S.A., 1973. The logical analysis of continuous non linear biochemical control networks. J. Theor. Biol. 1(39), 103–129.

Golightly, A., Wilkinson, D.J., 2006. Bayesian sequential inference for stochastic kinetic biochemical network models. J. Comput. Biol. 13(3), 838–851.

Hartemink, A.J., Gifford, D.K., Jaakkola, T.S. Young, R.A., 2001. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. Pac. Symp. Biocomput. 422–433.

Hatfield, G. W., Benham, C.J., 2002. DNA topology-mediated control of global gene expression in Escherichia coli. Annu. Rev. Genet. 36, 175–203, doi:10.1146/annurev.genet.36.032902.111815.

Henzinger, T.A., Ho, P.-H., 1995. Algorithmic analysis of nonlinear hybrid systems. CAV: Computer-Aided Verification, Lecture Notes in Computer Science 939, Springer, 225–238.

Henzinger, T.A., Ho, P.-H., Wong-Toi, H., 1997. HYTECH: a model checker for hybrid systems. International Journal on Software Tools for Technology Transfer. 1 (1–2), 110–122.

Kauffman, S.A., 1993. Origins of order: self-organization and selection in evolution. Oxford University Press. Technical monograph. ISBN 0-19-507951-5.

Kuttler, C., Niehren, J. 2006. Gene regulation in the Pi calculus: simulating cooperativity at the Lambda switch. LNCS, 4230, 24–55. doi: 10.1007/11905455.

Markowetz, F., Grossmann, S., Spang, R., 2005. Probabilistic soft interventions in conditional Gaussian networks. Proc. Tenth International Workshop on Artificial Intelligence and Statistics (AISTATS'05), R. Cowell and Z. Ghahramani (eds.).

Ropers, D., de Jong, H., Page, M., Schneider, D., Geiselmann, J., 2006. Qualitative simulation of the carbon starvation response in *Escherichia coli*. Biosystems. 2(84), 124–152, doi:10.1016/j.biosystems.2005.10.005.

Siebert, H., Bockmayr, A., 2006. Incorporating time delays into the logical analysis of gene regulatory networks. Computational Methods in Systems Biology, CMSB'06, Corrado Priami, Trento, Italy, Lecture Notes in Computer Science, Springer. 4210, 169-183.

Siebert, H., Bockmayr, A., 2008. Temporal constraints in the logical analysis of regulatory networks. Theor. Comput. Sci. 391(3), 258–275.

Snoussi, E.H., 1989. Qualitative dynamics of a piecewise-linear differential equations : a discrete mapping approach. Dynamics and stability of Systems. 4(3 & 4), 189–207.

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

Thomas, R., 1991. Regulatory networks seen as asynchronous automata : a logical description. J. Theor. Biol. 153, 1–23.

Thomas, R., d'Ari, R., 1990. Biological Feedback. CRC Press.

Thomas, R., Thieffry, D., Kaufman, M., 1995. Dynamical behaviour of biological regulatory networks: I. Biological role of feedback loops and practical use of the concept of the loop-characteristic state. Bull. Math. Biol. 57(2), 247–276.

Travers, A., Schneider, R., Muskhelishvili, G., 2001. DNA supercoiling and transcription in *Escherichia coli*: The FIS connection. Biochimie. 2, 1213-217, doi:10.1016/S0300-9084(00)01217-7.

Yu, J., Smith, V., Wang, P., Hartemink, A., Jarvis, E., 2002. Using Bayesian network inference algorithms to recover molecular genetic regulatory networks. International Conference on Systems Biology 2002 (ICSB02), December.

Qualitative modelling and analysis of gene regulatory networks: application to the adaptation of *Escherichia coli* bacterium to carbon availability

Jamil Ahmad[a], **Jérémie Bourdon**[b,c], **Damien Eveillard**[b], **Jonathan Fromentin**[a], **Olivier Roux**[a], **Christine Sinoquet**[b]

Abstract

Attempts to model Gene Regulatory Networks (GRNs) have yielded very different approaches. Among others, variants of Thomas's asynchronous boolean approach have been proposed, to better fit the dynamics of biological systems: notably, genes were allowed to reach different discrete expression levels, depending on the states of other genes, called the regulators: thus, activations and inhibitions are triggered conditionally on the proper expression levels of these regulators. In contrast, some fine-grained propositions have focused on the molecular level as modelling the evolution of biological compound concentrations through differential equation systems. Both approaches are limited. The first one leads to an oversimplification of the system, whereas the second is incapable to tackle large GRNs. In this context, hybrid paradigms, that mix discrete and continuous features underlying distinct biological properties, achieve significant advances for investigating biological properties. One of these hybrid formalisms proposes to focus, within a GRN abstraction, on the time delay to pass from a gene expression level to the next. Until now, no research work has been carried out, which attempts to benefit from the modelling of a GRN by differential equations, converting it into a multi-valued logical formalism of Thomas, with the aim of performing biological applications. The present research work fills this gap by describing a whole pipelined process which supervises the following stages: (i) model conversion from a Piece-wise Affine Differential Equation (PADE) modelization scheme into a discrete model with attractors (and generation of the corresponune pour la journée portes ouvertes de PolyTech ?ding dynamical graph), (ii) on the basis of probabilistic criteria, extraction of subgraphs of interest from the former dynamical graph, (iii) conversion of the subgraphs into Parametric Linear Hybrid Automata, (iv) analysis of dynamical properties (*e.g.* cyclic behaviours) using hybrid model-checking techniques. The present work is the outcome of a methodological investigation launched to cope with the GRN responsible for the reaction of *Escherichia coli* bacterium to carbon starvation. As expected, we retrieve a remarkable cycle already exhibited by a previous analysis of the PADE model. Above all, hybrid model-checking enables us to discover additional insightful results, whose interpretations are in accordance with biological evidences.

LINA, Université de Nantes
2, rue de la Houssinière
B.P. 92208 — F-44322 NANTES CEDEX 3