BIOINFORMATICS

Vol. 26 ECCB 2010, pages i603–i610
doi:10.1093/bioinformatics/btq387

Efficient parameter search for qualitative models of regulatory networks using symbolic model checking

Gregory Batt[1,*], Michel Page[2,3], Irene Cantone[4], Gregor Goessler[2], Pedro Monteiro[2,5] and Hidde de Jong[2]

[1]INRIA Paris - Rocquencourt, Le Chesnay, [2]INRIA Grenoble - Rhône-Alpes, Montbonnot, [3]IAE, Université Pierre Mendès France, Grenoble, France, [4]Clinical Sciences Center, Imperial College, London, UK and [5]INESC/Instituto Superior Técnico, Lisbon, Portugal

ABSTRACT

Motivation: Investigating the relation between the structure and behavior of complex biological networks often involves posing the question if the hypothesized structure of a regulatory network is consistent with the observed behavior, or if a proposed structure can generate a desired behavior.

Results: The above questions can be cast into a parameter search problem for qualitative models of regulatory networks. We develop a method based on symbolic model checking that avoids enumerating all possible parametrizations, and show that this method performs well on real biological problems, using the IRMA synthetic network and benchmark datasets. We test the consistency between IRMA and time-series expression profiles, and search for parameter modifications that would make the external control of the system behavior more robust.

Availability: GNA and the IRMA model are available at http://ibis.inrialpes.fr/

Contact: gregory.batt@inria.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

A central problem in the analysis of biological regulatory networks concerns the relation between their structure and dynamics. This problem can be narrowed down to the following two questions: (a) Is a hypothesized structure of the network consistent with the observed behavior? (b) Can a proposed structure generate a desired behavior?

Qualitative models of regulatory networks, such as (synchronous or asynchronous) Boolean models and piecewise-affine differential equation (PADE) models, have been proven useful for addressing the above questions. The models are coarse-grained, in the sense that they do not explicitly specify the biochemical mechanisms. However, they include the logic of gene regulation and allow different expression levels of the genes to be distinguished. They are interesting in their own right, as a way to capture in a simple manner the complex dynamics of a large regulatory network (Chaves *et al.*, 2009; Fauré *et al.*, 2006; Monteiro *et al.*, 2008; Saez-Rodriguez *et al.*, 2009). They can also be used as a first step to orient the development of more detailed quantitative ODE models.

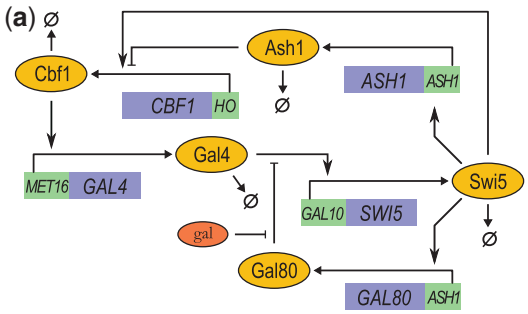
Qualitative models bring specific advantages when studying the relation between structure and dynamics. In order to answer questions (a) and (b), one has to search the parameter space to check if for some parameter values the network is consistent with the data or can attain a desired control objective. In qualitative models, the number of different parametrizations is finite and the number of possible values for each parameter is usually rather low. This makes parameter search easier to handle than in quantitative models, where exhaustive search of the continuous parameter space is in general not feasible. Moreover, qualitative models are concerned with trends rather than with precise quantitative values, which corresponds to the nature of much of the available biological data (Cantone *et al.*, 2009).

Nevertheless, the parametrization of qualitative models remains a complex problem. For most models of networks of biological interest the state and parameter spaces are too large to exhaustively test all combinations of parameter values. The aim of this article is to address this search problem for PADE models by treating it in the context of formal verification and symbolic model checking (Clarke *et al.*, 1999; Fisher and Henzinger, 2007).

Our contributions are twofold. On the methodological side, we develop a method that in comparison with our previous work (Batt *et al.*, 2005) makes it possible to efficiently analyze large and possibly incompletely parametrized PADE models. This is achieved by a *symbolic encoding* of the model structure, constraints on parameter values and transition rules describing the qualitative dynamics of the system. We can thus take full advantage of symbolic model checkers for testing the consistency of the network structure with dynamic properties expressed in temporal logics. The computer tool GNA has been extended to export the symbolic encoding of PADE models in the NuSMV language (Cimatti *et al.*, 2002). In comparison with related work (Barnat *et al.*, 2009; Bernot *et al.*, 2004; Corblin *et al.*, 2009; Fromentin *et al.*, 2007), our method applies to incompletely instead of fully parametrized models, provides more precise results and the encoding is efficient without (strongly) simplifying the PADE dynamics.

On the application side, we show that the *method performs well on real problems*, by means of the IRMA synthetic network and benchmark experimental datasets (Cantone *et al.*, 2009). More precisely, we are able to find parameter values for which the network satisfies temporal-logic properties describing observed expression profiles, both on the level of individual and averaged time series. The method is selective in the sense that only a small part of the parameter space is found to be compatible with the observations. Analysis of these parameter values reveals that biologically relevant constraints have been identified. Moreover, we make suggestions to improve the robustness of the external control of the IRMA behavior by proposing a rewiring of the network.

*To whom correspondence should be addressed.



(b)

$$\dot{x}_{Gal4} = \kappa^0_{Gal4} + \kappa_{Gal4}\, s^+(x_{Cbf1}, \theta_{Cbf1}) - \gamma_{Gal4}\, x_{Gal4}$$
$$\dot{x}_{Swi5} = \kappa^0_{Swi5} + \kappa_{Swi5}\, s^+(x_{Gal4}, \theta_{Gal4})\,(1 - s^+(x_{Gal80}, \theta_{Gal80}) s^-(u_{gal}, \theta_{gal})) - \gamma_{Swi5}\, x_{Swi5}$$
$$\dot{x}_{Ash1} = \kappa^0_{Ash1} + \kappa_{Ash1}\, s^+(x_{Swi5}, \theta^a_{Swi5}) - \gamma_{Ash1}\, x_{Ash1}$$
$$\dot{x}_{Cbf1} = \kappa^1_{Cbf1}\, s^+(x_{Swi5}, \theta^c_{Swi5}) + \kappa^2_{Cbf1}\, s^+(x_{Swi5}, \theta^c_{Swi5}) s^-(x_{Ash1}, \theta_{Ash1}) - \gamma_{Cbf1}\, x_{Cbf1}$$
$$\dot{x}_{Gal80} = \kappa^0_{Gal80} + \kappa_{Gal80}\, s^+(x_{Swi5}, \theta^g_{Swi5}) - \gamma_{Gal80}\, x_{Gal80}$$

Fig. 1. Synthetic IRMA network in yeast. (**a**) Schematic representation of the network constructed in Cantone *et al*. (2009). The green and blue boxes are promoter and genes, and the yellow and red ovals are proteins and metabolites. (**b**) PADE model of IRMA, with state variables x, protein synthesis constants κ, decay constants γ and thresholds θ. The input variable u_{gal} refers to the presence of galactose ($\dot{u}_{gal} = 0$). The subscripts $_{Gal4}$, $_{Swi5}$, $_{Ash1}$, $_{Cbf1}$, $_{Gal80}$ refer to the proteins.

2 QUALITATIVE MODEL OF IRMA NETWORK

2.1 IRMA network

IRMA is a synthetic network constructed in yeast and proposed as a benchmark for modeling and identification approaches (Cantone *et al*., 2009). The network consists of five well-characterized genes that have been chosen such that different kinds of interactions are included, notably transcription regulation and protein–protein interactions. The endogenous copies of the genes were deleted to reduce crosstalk of IRMA with the regulatory networks of the host cell. In order to further isolate the synthetic network from its cellular environment, the genes belong to distinct, non-redundant pathways.

The structure of the IRMA network is shown in Figure 1a. The expression of the *CBF1* gene is under the control of the *HO* promoter, which is positively regulated by Swi5 and negatively regulated by Ash1. *CBF1* encodes the transcription factor Cbf1 that activates expression of the *GAL4* gene. The *GAL10* promoter is activated by Gal4, but only in the absence of Gal80 or in the presence of galactose. Gal80 binds to the Gal4 activation domain, but galactose releases this inhibition of transcription. The *GAL10* promoter controls the expression of *SWI5*, whose product not only activates the above-mentioned *HO* promoter, but also the *ASH1* promoter controlling the expression of the *GAL80* and *ASH1* genes.

The network contains one positive (Swi5/Cbf1/Gal4/Swi5) and two negative (Swi5/Gal80/Swi5; Swi5/Ash1/Cbf1/Gal4/Swi5) feedback loops. Negative feedback loops are a necessary condition for the occurrence of oscillations (Thomas and d'Ari, 1990), while the addition of positive feedback is believed to increase the robustness of the oscillations (Tsai *et al*., 2008). Consequently, for suitable parameter values IRMA might function as a synthetic oscillator.

2.2 Measurements of IRMA dynamics

The behavior of the network has been monitored in response to two different perturbations (Cantone *et al*., 2009): shifting cells from glucose to galactose medium (switch-on experiments), and from galactose to glucose medium (switch-off experiments). The terms 'switch-on' ('switch-off') refer to the activation (inhibition) of *SWI5* expression during growth on galactose (glucose). For these two perturbations, the temporal evolution of the expression of *all* the genes in the network was monitored by qRT-PCR with good time resolution.

Figure 2a represents the expression of all genes, averaged over five (switch-on) or four (switch-off) independent experiments. In the switch-off experiments (galactose to glucose), the transcription of all genes is shut off. In the switch-on experiments, a seemingly oscillatory behavior is present with Swi5 peaks at 40 and 180 min, and Swi5, Cbf1 and Ash1 expressed at moderate to high levels (Cantone *et al*., 2009).

The analysis of the individual time series reveals that in some cases the gene expression profiles are indeed similar, at least qualitatively, whereas in other cases notable differences exist (e.g. the oscillatory behavior is not present in all switch-on time series, see Fig. 2c). In the latter case, averaged expression levels may be a misleading representation of the network behavior.

2.3 PADE model of IRMA network

We built a qualitative model of the IRMA dynamics using PADE models of genetic regulatory networks. PADE models, originally introduced in Glass and Kauffman (1973), provide a coarse-grained picture of the network dynamics. They have the following general form:

$$\dot{x}_i = f_i(x) \triangleq \sum_{l \in L_i} \kappa^l_i\, b^l_i(x) - \gamma_i x_i,\ i \in [1, n] \qquad (1)$$

where $x \in \Omega \subset \mathbb{R}^n_{\geq 0}$ represents a vector of n protein (or RNA) concentrations. The synthesis rate is composed of a sum of synthesis constants κ^l_i, each modulated by a regulation function $b^l_i(x) \in \{0, 1\}$, with l in an index set L_i. A regulation function is an algebraic expression of step functions $s^+(x_j, \theta_j)$ or $s^-(x_j, \theta_j)$ which formalizes the regulatory logic of gene expression. θ_j is a so-called threshold for the concentration x_j. The step function $s^+(x_j, \theta_j)$ evaluates to 1 if $x_j > \theta_j$, and to 0 if $x_j < \theta_j$, thus capturing the switch-like character of gene regulation ($s^-(x_j, \theta_j) = 1 - s^+(x_j, \theta_j)$). The degradation of a gene product is a first-order term, with a degradation constant γ_i.

In the case of IRMA, we define five variables, each corresponding to the total concentration of a protein, and an input variable denoting the concentration of galactose. Notice that the measurements of the network dynamics concern mRNA and not protein levels. We assume that the variations in mRNA and protein levels are the same, even though this may not always be the case. A similar approximation is made in Cantone *et al*. (2009), where protein and mRNA levels are assumed to be proportional.

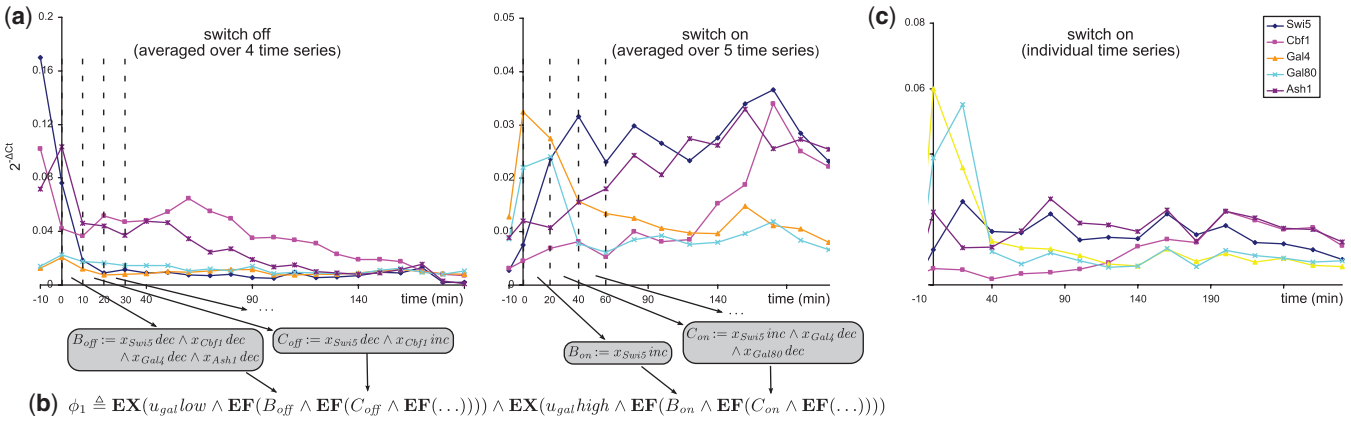


Fig. 2. Dynamic behavior of the IRMA network in response to medium shift perturbations. (**a**) Temporal profiles of averaged gene expression measured with qRT-PCR during switch-off (left) and switch-on (right) experiments (data from Cantone *et al*., 2009). (**b**) Temporal logic encoding of the switch-off and switch-on behaviors. The operator $\mathbf{EF}\phi$ expresses the possibility to reach a future state satisfying ϕ, whereas the operator $\mathbf{EX}\phi$ is used to require the existence of an initial state satisfying ϕ. $u_{gal}low$ and $u_{gal}high$ denote the absence and presence of galactose, respectively. See Clarke *et al*. (1999) for more details on the temporal logic CTL. Only changes greater than 5×10^{-3} units are considered significant. (**c**) Temporal gene expression profile in an individual switch-on experiment showing a switch-off-like behavior.

The PADE model of the IRMA network is shown in Figure 1b. Consider the equation for the protein Gal4. κ^0_{Gal4} is its basal synthesis rate, and $\kappa^0_{Gal4}+\kappa_{Gal4}$ its maximal synthesis rate when the *GAL4* activator Cbf1 is present (i.e. $x_{Cbf1} > \theta_{Cbf1}$). Swi5 is regulated in a more complex way. The expression of its gene is activated by Gal4, but only when not both Gal80 is present and galactose absent (which would lead to Gal4 inactivation by Gal80). The step-function expression in Figure 1b mathematically describes this condition. The IRMA PADE model is described in more detail in Section 1 of the Supplementary Material.

The model resembles the ODE model in Cantone *et al*. (2009), but notably approximates the Hill-type kinetic rate laws by step functions. It thus makes the implicit assumption that important qualitative dynamical properties of the network are intimately connected with the network structure and the regulatory logic, independently from the details of the kinetic mechanisms and precise parameter values. Several studies have shown this assumption to be valid in a number of model systems (Chaves *et al*., 2009; Davidich and Bornholdt, 2008), although care should be exercised in deciding exactly when modeling approximations are valid (Polynikis *et al*., 2009).

To investigate for the possible existence of unknown interactions between the synthetic network and the host, we would like to test by means of the PADE model if the network structure and the regulatory logic alone can fully account for the trends in the gene expression profiles observed in Cantone *et al*. (2009). Because the addition of galactose does not always lead to an effective activation of the IRMA genes, we also search for parameter modifications that would render the network response to galactose more robust.

3 SEARCH OF PARAMETER SPACE USING SYMBOLIC MODEL CHECKING

3.1 Qualitative analysis of PADE models

The advantage of PADE models is that the qualitative dynamics of high-dimensional systems are relatively easy to analyze, using only the total *order* on parameter values rather than exact numerical values (Batt *et al*., 2008; Edwards and Glass, 2006). The main difficulty lies in treating the discontinuities in the right-hand side of the differential equations, at the threshold values of the step functions. Following Gouzé and Sari (2002), the use of differential inclusions based on Filippov solutions has been proposed in Batt *et al*. (2008) and implemented in the computer tool GNA (Batt *et al*., 2005). Here, we recast this analysis in a form that underlies the symbolic encoding of the dynamics below.

The key to our reformulation of the qualitative analysis of the PADE dynamics is the extension of step functions s^+ to interval-valued functions S^+, where

$$S^+(x_j,\theta_j) = \begin{cases} [0,0] \text{ if } x_j < \theta_j \\ [0,1] \text{ if } x_j = \theta_j \\ [1,1] \text{ if } x_j > \theta_j \end{cases} \tag{2}$$

Because the step functions are not defined at their thresholds, we conservatively assume that they can take any value between 0 and 1 [see Chaves *et al*. (2009) for a similar idea]. When replacing the step functions by their extensions, the regulation functions $b^l_i(x)$ become *interval-valued* functions $B^l_i : \mathbb{R}^n_{\geq 0} \to \{[0,0],[0,1],[1,1]\}$, and Equation (1) generalizes to the following differential inclusion using interval arithmetic (Moore, 1979):

$$\dot{x}_i \in F_i(x) \triangleq \sum_{l \in L_i} \kappa^l_i B^l_i(x) - \gamma_i x_i,\ i \in [1,n] \tag{3}$$

The solutions of (3) are for practical purposes the same as the solutions of the differential inclusions defined in Batt *et al*. (2008) (see Section 2 of the Supplementary Material).

The starting point for our qualitative analysis is the introduction of a rectangular partition $\mathcal{D}$ of the state space Ω. This partition is a rectangular grid defined by the threshold parameters $\Theta_i = \{\theta^j_i \mid j \in J_i\}$, where J_i is an index set, and the so-called focal parameters $\Lambda_i = \{\sum_{l \in B} \kappa^l_i / \gamma_i \mid B \subseteq L_i\}$, $i \in [1,n]$. Focal parameters are steady-state concentrations towards which the system locally converges in a monotonic way (Glass and Kauffman, 1973). For Gal4,

we have $\Theta_{Gal4}=\{\theta_{Gal4}\}$ and $\Lambda_{Gal4}=\{0,\kappa^0_{Gal4}/\gamma_{Gal4},(\kappa^0_{Gal4}+\kappa_{Gal4})/\gamma_{Gal4}\}$.

Interestingly, the partition has the property that in each domain $D\in\mathcal{D}$, the protein production rates are identical: for all $x,y\in D$, it holds that $B^l_i(x)=B^l_i(y)\triangleq B^l_i(D)$. As a consequence, the derivatives of the concentration variables have a *unique sign pattern*: for all $x,y\in D$, it holds that $\text{sign}(F_i(x))=\text{sign}(F_i(y))\subseteq\{-1,0,1\}$, where $\text{sign}(A)\triangleq\{\text{sign}(a)\,|\,a\in A\}$ denotes the signs of the elements in A (Batt *et al*., 2008). Notice that this property is not obtained for less fine-grained partitions used in related work (Barnat *et al*., 2009; Bernot *et al*., 2004; Chaves *et al*., 2009; Corblin *et al*., 2009; Fauré *et al*., 2006; Fromentin *et al*., 2007). It will be found critical for the search of parametrized models of IRMA that satisfy the time-series data.

The above considerations motivate a discrete abstraction, resulting in a *state transition graph*. In this graph, the states are the domains $D\in\mathcal{D}$, and there is a transition from a domain D to another domain D', if there exists a solution of the differential inclusion (3) that starts in D and reaches D', without leaving $D\cup D'$. The state transition graph defines the qualitative dynamics of the system, in the sense that paths in this graph describe how the qualitative state of the system evolves over time (Batt *et al*., 2008).

In Batt *et al*. (2008), three different types of transitions are defined: *internal*, from a domain D to itself; *dimension-increasing*, from a domain D to another, higher dimensional domain D' ($D\subseteq\partial D'$); and *dimension-decreasing*, from a domain D to a lower dimensional domain D' ($D'\subseteq\partial D$), where ∂D denotes the boundary of D in its supporting hyperplane. We reformulate here the transition rules using the interval extensions of the regulation functions. We introduce an interval-valued function $F_i:\mathcal{D}\times\mathcal{D}\to 2^{\mathbb{R}}$, where $F_i(D,D')=\sum_{l\in L_i}\kappa^l_i B^l_i(D)-\gamma_i D'_i$, for $D,D'\in\mathcal{D}$. $F_i(D,D')$ represents the flow in D infinitely close to D'. In order to evaluate $F_i(D,D')$, we use interval arithmetic (Moore, 1979). For instance, in a domain in which $x_{Swi5}>\theta^c_{Swi5}$ and $x_{Ash1}=\theta_{Ash1}$, we have $S^+(x_{Swi5},\theta^c_{Swi5})=[1,1]$ and $S^-(x_{Ash1},\theta_{Ash1})=[0,1]$, so that the differential inclusion for x_{Cbf1} becomes $[\kappa^1_{Cbf1}-\gamma_{Cbf1}x_{Cbf1},\kappa^1_{Cbf1}+\kappa^2_{Cbf1}-\gamma_{Cbf1}x_{Cbf1}]$. We obtain the following transition rule:

Proposition 1 (Dimension-increasing transition). *Let $D,D'\in\mathcal{D}$ and $D\subseteq\partial D'$, that is, D lies in the boundary of D'. $D\to D'$ is a dimension-increasing transition iff*

(1) *$\forall i\in[1,n]$, such that D_i and D'_i coincide with a value in $\Theta_i\cup\Lambda_i$, it holds that $0\in F_i(D',D)$, and*

(2) *$\forall i\in[1,n]$, such that $D_i\neq D'_i$, it holds that $\exists\alpha>0$ such that $\alpha\in F_i(D',D)(D'_i-D_i)$*

Condition 1 guarantees that solutions can remain in domains located in threshold and focal planes, while Condition 2 expresses that the direction of the flow in the domains $(F_i(D',D))$ agrees with their relative position (D'_i-D_i). The proof of the rule and the rules for other types of transition can be found in Section 3 of the Supplementary Material.

It can be shown that exact parameter values are *not* needed for the analysis of the qualitative dynamics of a PADE model: it is sufficient to know the *ordering of the threshold and focal parameters* (Batt *et al*., 2008). This comes from the fact that the sign of F_i, and hence the transitions and the state transition graph, are invariant for regions of the parameter space defined by a total order on $\Theta_i\cup\Lambda_i$. We call each such total order a *parametrization* of the PADE model.

3.2 A model-checking approach for parameter search

For large graphs like that obtained for IRMA (which has about 50 000 states), verifying the compatibility of the network structure with an observed or desired behavioral property is impossible to do by hand. This has motivated the use of model-checking tools (e.g. Barnat *et al*., 2009; Batt *et al*., 2005; Bernot *et al*., 2004; Fisher and Henzinger, 2007). For PADE models, each state in the graph is described by atomic propositions whose truth values are preserved under the discrete abstraction, such as the above-mentioned derivative sign patterns. The atomic propositions are used to formulate properties in a *temporal-logic formula* ϕ and *model checkers* automatically test if the state transition graph T satisfies the formula ($T\models\phi$).

Because the number of possible parametrizations and the size of state transition graphs rapidly grow with the number of genes, the naive approach consisting in enumerating all parametrizations of a PADE model, and for each of these generating the state transition graph and testing $T\models\phi$, is only feasible for the simplest networks. We therefore propose an alternative approach, based on the *symbolic* encoding of the above search problem, without explicitly generating the possible parametrizations of the PADE models and the corresponding state transition graphs. This enables one to exploit the capability of symbolic model checkers to efficiently manipulate *implicit* descriptions of the state and parameter space.

3.3 Symbolic encoding of PADE model and dynamics

We summarize the main features of the encoding. We particularly focus on the discretization of the state space, which connects the symbolic encoding to the mathematical analysis of PADE models, and the use of the discretization for the computation of $F_i(D',D)$, which is essential for determining state transitions.

We call $\mathcal{C}$ a discretization function that maps $D\in\mathcal{D}$ to a set of unique integer coordinates, and $\mathcal{C}(D)=\mathcal{C}(D_1)\times\cdots\times\mathcal{C}(D_n)$. Let m_i be the number of non-zero parameters in $\Theta_i\cup\Lambda_i$, $i\in[1,n]$. Then $\mathcal{C}(D_i)\in\{0,1,\ldots,2m_i+1\}$, and more specifically, $\mathcal{C}(D_i)\in\{0,2,\ldots,2m_i\}$ if D_i coincides with a threshold or focal plane, and $\mathcal{C}(D_i)\in\{1,3,\ldots,2m_i+1\}$ otherwise. More generally, $\mathcal{C}(S)=\{\mathcal{C}(D)\,|\,D\subseteq S\}$, for any set of domains S. Obviously, $\mathcal{C}$ can also be used for the discretization of parameter values. Given the following total order on the threshold and focal parameters of variable x_{Gal4}, $0<\kappa^0_{Gal4}/\gamma_{Gal4}<\theta_{Gal4}<(\kappa^0_{Gal4}+\kappa_{Gal4})/\gamma_{Gal4}$, we find $\mathcal{C}(0)=0$ (by definition), $\mathcal{C}(\kappa^0_{Gal4}/\gamma_{Gal4})=2$, $\mathcal{C}(\theta_{Gal4})=4$ and $\mathcal{C}(\kappa^0_{Gal4}+\kappa_{Gal4})/\gamma_{Gal4})=6$.

The above discretization motivates the introduction of symbolic variables $\hat{D}_i$, $\hat{D}'_i$, $\hat{\theta}^j_i$, $\hat{\lambda}^j_i$ encoding $\mathcal{C}(D_i)$, $\mathcal{C}(D'_i)$, $\mathcal{C}(\theta^j_i)$, $\mathcal{C}(\lambda^j_i)$, respectively, with $\theta^j_i\in\Theta_i$ and $\lambda^j_i\in\Lambda_i$. The different conditions in Proposition 1 can be expressed in terms of these variables. For instance, $\text{sign}(D'_i-D_i)$ becomes $\text{sign}(\hat{D}'_i-\hat{D}_i)$. In the case of $F_i(D',D)$, multiplication by $1/\gamma_i$ does not change the sign, but gives the more convenient expression

$$F_i(D,D')/\gamma_i=\sum_{l\in L_i}(\kappa^l_i/\gamma_i)B^l_i(D)-D'_i \qquad (4)$$

The first term in the right-hand side is simply an interval whose upper and lower bounds are focal parameters, determined by the regulation functions $B_i^l(D)$. By redefining the step functions in terms of the symbolic variables:

$$S^+(D_j,\theta_j)=\begin{cases}[0,0] & \text{iff } \hat{D}_j<\hat{\theta}_j\\ [0,1] & \text{iff } \hat{D}_j=\hat{\theta}_j\\ [1,1] & \text{iff } \hat{D}_j>\hat{\theta}_j\end{cases} \quad (5)$$

each $B_i^l(D)$ can be simply computed using interval arithmetic. This allows the interval bounds of $\sum_{l\in L_i}(\kappa_i^l/\gamma_i)B_i^l(D)$ to be computed, which are simply given by variables $\hat{\lambda}_i^j$. Subtracting $\hat{D}_i'$ allows the sign of $F_i(D',D)$ and thus the conditions for a transition $D\to D'$ to be evaluated.

The specification of transitions in a symbolic way is the main stumble block for the efficient encoding of the PADE dynamics, especially when D is located on a threshold plane. In our previous work (Batt *et al*., 2008), the computation of transitions required the enumeration of an exponential number of domains surrounding D (Barnat *et al*., 2009). The interval-based formulation proposed here allows (the sign of) $F_i(D,D')$ to be computed in one stroke.

The implementation in a model checker such as NuSMV (Cimatti *et al*., 2002) is straightforward with the above encoding. We apply invariant constraints on the symbolic variables to exclude all valuations of $\hat{D}_i$, $\hat{D}_i'$, $\hat{\theta}_i^j$, $\hat{\lambda}_i^j$ that do not correspond to a valid transition from D to D'. We apply three types of invariants. The first ones constrain parameters to remain constant. The second ones constrain D and D' to be neighbors in the state space (e.g. $D\subseteq\partial D'$ for dimension-increasing transitions). The last ones constrain the relative position of D_i and D_i' and the parameter order as stated in the transitions conditions. For comparison with experimental data, we also need to know the variations of concentrations of gene products in each state. These correspond to the derivative sign pattern, $\text{sign}(F_i(D,D'))$.

The initial states of our symbolic description include each possible parametrization, that is, all possible values for $\hat{\theta}_i^j$ and $\hat{\lambda}_i^j$, and transition towards all states D. In CTL, a temporal logic property ϕ holds if all initial states satisfy ϕ. Therefore, by testing whether $\neg\phi$ holds, we verify the absence of a parametrization satisfying ϕ. A counterexample to $\neg\phi$ thus directly returns a valid parametrization. The current version 8 of GNA has been extended with export functionalities to generate the symbolic encoding of PADE models in the NuSMV language.

4 VALIDATION: CONSISTENCY OF IRMA NETWORK WITH EXPERIMENTAL DATA

4.1 Temporal-logic encoding of observations

Even when genetic constructs are tested separately and assembled with care, it is not obvious that a synthetic network will function in its cellular context as initially planned. Here, we test the consistency between the IRMA network and the experimental data by expressing that for each condition, switch-on and switch-off, there must exist an initial state of the system and a path starting from this state along which the gene expression changes correspond to the observed time-series data. For example, for the switch-off time-series we encode that there exists an initial state where in absence of galactose the expression of *SWI5*, *CBF1*, *GAL4* and *ASH1* decreases (in the interval [0,10] min), and from which a state can be reached where the expression of *SWI5* decreases and the expression of *CBF1* increases (in the interval [10,20] min), etc. The generation of this property ϕ_1 from the experimental data leads to the temporal-logic formula shown in Figure 2b. To disregard small fluctuations due to biological and experimental noise, we considered changes of magnitude less than 5×10^{-3} units not significant. Moreover, we ignore in our specification the very first measurements (in the interval [−10,0]), just before shifting cells to a new medium, as they probably reflect network-independent effects (Cantone *et al*., 2009).

The data presented in Cantone *et al*. (2009) for switch-on and switch-off conditions are the average of 5 and 4 individual experiments, respectively. As noticed in Section 2.2, considering the averaged gene expression profile may be misleading. Asking for consistency between our model and the result of each individual experiment might therefore be more appropriate. This leads us to define a second property ϕ_2 similar to ϕ_1 but requiring the existence of nine paths in the graph, one for each of the observed behaviors in the nine individual experiments. Although the information we extract from the experimental data only concerns trends in gene product levels, the accumulation of these simple observations leads to fairly complex constraints. Property ϕ_2 involves nearly 160 constraints on derivative signs.

4.2 Testing consistency of network with observations

We use our symbolic encoding of the PADE dynamics to test $\neg\phi_1$. NuSMV returns false, meaning that a parametrization satisfying the averaged time-series data exists (Section 3.3). The result was obtained in 49 s on a laptop (PC, 2.2 GHz, 1 core, 2 GB RAM), with an additional 100 s to provide the counterexample (Table 1). When analyzing the corresponding parametrization, the thresholds are mostly greater than the focal parameter for basal expression and smaller than the focal parameter for upregulated expression, e.g. $\kappa^0_{Ash1}/\gamma_{Ash1}<\theta_{Ash1}<(\kappa^0_{Ash1}+\kappa_{Ash1})/\gamma_{Ash1}$. This is not surprising as the focal parameters correspond to the lowest and highest possible expression levels. The threshold at which Ash1 controls *CBF1* expression is expected to lie between the two extremes. The only exception is Gal80, for which it holds $(\kappa^0_{Gal80}+\kappa_{Gal80})/\gamma_{Gal80}<\theta_{Gal80}$. According to this constraint, Gal80 plays no role in the system, since it cannot exceed the threshold concentration above which it inhibits Swi5. This is interesting because it suggests that the switch-off behavior may occur even without any inhibition by Gal80, and consequently, in a galactose-independent manner.

The dynamic properties of the PADE model can be analyzed in more detail by means of GNA. This shows the existence of an asymptotically stable steady state corresponding to switch-off conditions, with low Swi5, Gal4, Cbf1, Ash1 and Gal80 concentrations. In addition, GNA finds strongly connected components (SCCs) consistent with the observed damped oscillations in galactose media. However, the attractors co-exist irrespectively of the presence or absence of galactose, revealing that galactose does not necessarily drive the system to a single attractor for this particular parametrization.

We also tested whether the above parametrization is consistent with time-series data from the individual experiments. The model checker shows that it does not satisfy the more constraining property ϕ_2. However, we do find another parametrization for which ϕ_2 holds.

Table 1. Summary of parametrizations found by checking the consistency of the IRMA structure with the observed and desired behaviors, expressed as temporal-logic properties ϕ_1, ϕ_2 and ϕ_3. The table shows the parametrization returned when testing the truth value of the property on the symbolically encoded PADE model and gene expression profiles (left) and summarizes all parametrizations satisfying the properties (right).

	Symbolic state space and symbolic parameter space		Symbolic state space and fully parametrized models	
Property	Existence of parametrization	Parametrization[a]	Number of parametrizations	Parametrization[a]
ϕ_1: averaged time-series	Yes (49 s)	$\frac{\kappa^0_{Swi5}}{\gamma_{Swi5}} < \theta^g_{Swi5} < \theta^c_{Swi5} < \theta^a_{Swi5} < \frac{\kappa^0_{Swi5}+\kappa_{Swi5}}{\gamma_{Swi5}}$ $\wedge\ \frac{\kappa^0_{Gal80}}{\gamma_{Gal80}} < \frac{\kappa^0_{Gal80}+\kappa_{Gal80}}{\gamma_{Gal80}} < \theta_{Gal80}$	64 (885 s)	See Section 4 of Supplementary Material
ϕ_2: individual time-series	Yes (131 s)	$\frac{\kappa^0_{Swi5}}{\gamma_{Swi5}} < \theta^c_{Swi5} < \theta^a_{Swi5} < \theta^g_{Swi5} < \frac{\kappa^0_{Swi5}+\kappa_{Swi5}}{\gamma_{Swi5}}$ $\wedge\ \frac{\kappa^0_{Gal80}}{\gamma_{Gal80}} < \theta_{Gal80} < \frac{\kappa^0_{Gal80}+\kappa_{Gal80}}{\gamma_{Gal80}}$	4 (2021 s)	$\frac{\kappa^0_{Swi5}}{\gamma_{Swi5}} < \theta^c_{Swi5} < (\theta^a_{Swi5}, \theta^g_{Swi5}) < \frac{\kappa^0_{Swi5}+\kappa_{Swi5}}{\gamma_{Swi5}}$ $\wedge\ (\frac{\kappa^0_{Gal80}}{\gamma_{Gal80}}, \theta_{Gal80}) < \frac{\kappa^0_{Gal80}+\kappa_{Gal80}}{\gamma_{Gal80}}$
ϕ_3: single attractor	Yes (126 s)	$\theta^c_{Swi5} < \frac{\kappa^0_{Swi5}}{\gamma_{Swi5}} < \theta^g_{Swi5} < \theta^a_{Swi5} < \frac{\kappa^0_{Swi5}+\kappa_{Swi5}}{\gamma_{Swi5}}$ $\wedge\ \theta_{Gal80} < \frac{\kappa^0_{Gal80}}{\gamma_{Gal80}} < \frac{\kappa^0_{Gal80}+\kappa_{Gal80}}{\gamma_{Gal80}}$	7 (1300 s)	$\theta^c_{Swi5} < \frac{\kappa^0_{Swi5}}{\gamma_{Swi5}} < \theta^a_{Swi5} < \frac{\kappa^0_{Swi5}+\kappa_{Swi5}}{\gamma_{Swi5}}$ $\wedge\ \theta_{Gal80} < \frac{\kappa^0_{Gal80}+\kappa_{Gal80}}{\gamma_{Gal80}}$ $\wedge\ (\theta^g_{Swi5} < \frac{\kappa^0_{Swi5}}{\gamma_{Swi5}} \vee \theta_{Gal80} < \frac{\kappa^0_{Gal80}}{\gamma_{Gal80}})$

[a]All parametrizations shown additionally include $[\kappa^1_{Cbf1}/\gamma_{Cbf1} < \theta_{Cbf1} < (\kappa^1_{Cbf1}+\kappa^2_{Cbf1})/\gamma_{Cbf1}] \wedge [\kappa^0_{Gal4}/\gamma_{Gal4} < \theta_{Gal4} < (\kappa^0_{Gal4}+\kappa_{Gal4})/\gamma_{Gal4}] \wedge [\kappa^0_{Ash1}/\gamma_{Ash1} < \theta_{Ash1} < (\kappa^0_{Ash1}+\kappa_{Ash1})/\gamma_{Ash1}]$.

In this case, all thresholds are situated between the basal and upregulated focal parameters.

4.3 Detailed analysis of valid parameter set

Our consistency tests only confirm that a parametrization exists for which the structure of the network is consistent with the observed behavior. However, it does not say if this is trivially the case (for most parametrizations) or if the properties are selective (for only a few parametrizations). To investigate this we exhaustively generated all parametrizations, and tested for each of them properties ϕ_1 and ϕ_2. Although the total number of parameter orderings is fairly large, the exhaustive analysis is still manageable for networks of this size.

Out of the 4860 completely parametrized PADE models, we found that only a surprisingly small subset is consistent with the observations. For the averaged time series, only 64 parametrizations are consistent, while for the individual time series this subset is further reduced to 4 (Table 1). The properties extracted from the data are thus quite selective.

The results for individual time series indicate that to be consistent with the experimental data, the activation threshold of *CBF1* by Swi5 (θ^c_{Swi5}), must be smaller than the activation thresholds of *ASH1* and *GAL80* by Swi5 (θ^a_{Swi5} and θ^g_{Swi5}). Interestingly, this result is corroborated by independent measurements of promoter activities, which show that the activation threshold for the *ASH1* promoter, controlling *ASH1* and *GAL80* expression, is nearly twice as high as the one for the *HO* promoter controlling *CBF1* expression (Table S1 of Cantone *et al*., 2009).

A second finding is that the dynamics of the system is consistent with the experimental data even if $\theta_{Gal80} < \kappa^0_{Gal80}/\gamma_{Gal80}$, that is when *GAL80* is constitutively expressed above its inhibition threshold. This indicates that an effective regulation of *GAL80* expression by Swi5 is of little importance for the functioning of the network. Indeed, it was found that *GAL80* is not much responsive to changes in Swi5 availability: Cantone *et al*. observed that a 6-fold increase of *SWI5* expression leads to only a negligible (1.08-fold) increase in *GAL80* expression levels (Fig. 4A in Cantone *et al*., 2009).

5 RE-ENGINEERING: IMPROVING EXTERNAL CONTROL BY GALACTOSE

In one experiment at least, the addition of galactose does not significantly change the system's behavior: a switch-off-like response is observed in switch-on conditions. To obtain a more robust external control of the system, we would like to ensure that the addition of galactose drives the system out of the low-Swi5 state.

5.1 Temporal-logic specification of design objective

We start by specifying that in switch-off conditions the Swi5 concentration must eventually remain low, that is, equal to its basal expression level $\kappa^0_{Swi5}/\gamma_{Swi5}$. This is expressed in CTL as **AFAG** $x_{Swi5}\,low$. In switch-on conditions, an oscillatory behavior in the concentration of Swi5 is expected. It can be formulated by means of the formula $\mathbf{AGAF}(x_{Swi5}\,inc \wedge \mathbf{AF}\ x_{Swi5}\,dec)$, requiring that an increase in x_{Swi5} is observed infinitely often and necessarily followed by a decrease in x_{Swi5}. In addition to these two basic requirements, we impose that in presence of galactose, the Swi5 concentration cannot indefinitely stay low: $u_{gal}\,high \rightarrow \mathbf{AF}\neg x_{Swi5}\,low$. We prefix these specifications so as to express the possibility (**EX**) to reach the appropriate attractor from at least one initial state, and the necessity (**AX**) to leave the switch-off steady state for all initial states in switch-on conditions:

$$\begin{aligned}\phi_3 \triangleq\ & \mathbf{EX}(u_{gal}\,high \wedge \mathbf{AGAF}(x_{Swi5}\,inc \wedge \mathbf{AF}\ x_{Swi5}\,dec))\\ & \wedge \mathbf{EX}(u_{gal}\,low \wedge \mathbf{AFAG}\ x_{Swi5}\,low)\\ & \wedge \mathbf{AX}(u_{gal}\,high \rightarrow \mathbf{AF}\neg x_{Swi5}\,low)\end{aligned}$$

5.2 Parametrizations consistent with design objective

Using symbolic model checking, we test the feasibility of ϕ_3. In about 2 min, we find a valid parametrization (Table 1). For this parametrization, in the presence of galactose GNA finds two terminal SCCs attracting the major part of the state space, and notably the switch-off state. In the absence of galactose, although SCCs are present, they are non-terminal and one can show that a unique stable steady state with all genes off (i.e. corresponding to switch-off conditions) is eventually always reached.

Recall that one of the time series in the switch-on conditions contradicts our specification. It is consequently not surprising that none of the parametrizations consistent with the experimental data satisfies ϕ_3. We searched for all valid parametrizations and found that only 7 out of 4860 are consistent with our specification (Table 1).

A first surprising feature is that $\theta^c_{Swi5} < \kappa^0_{Swi5}/\gamma_{Swi5}$: Swi5 must always activate *CBF1*. Stated differently, this constraint simply suggests to remove the regulation of *CBF1* by Swi5. This can be explained by a qualitative analysis of the system dynamics. In the presence of galactose, we expect oscillations for Swi5. However, the presence of Swi5 is required for the expression of *CBF1* since the *HO* promoter functions like an AND gate: *HO* is on if and only if Swi5 is present *and* Ash1 is absent. So, if Swi5 is not permanently present, Cbf1 and then Gal4 might disappear, causing the system to converge to the switch-off state.

A second surprising feature is that the regulation of *GAL80* by Swi5 should not be effective. Indeed $\theta^g_{Swi5} < \kappa^0_{Swi5}/\gamma_{Swi5}$ or $\theta_{Gal80} < \kappa^0_{Gal80}/\gamma_{Gal80}$ means that either the *GAL80* promoter is always activated, or that the Gal80 concentration is always sufficient to repress *SWI5*. As above, this suggests to remove an interaction, namely the regulation of *GAL80* by Swi5. Interestingly, the demand for increased external control of the system leads us to a simplified design in which two out of the three feedback loops (Swi5/Cbf1/Gal4/Swi5 and Swi5/Gal80/Swi5) are removed.

6 DISCUSSION

We propose a method for efficient search of the parameter space of qualitative models of regulatory networks, to investigate the relation between structural and behavioral properties of these systems.

On the methodological side, the main novelty is that we develop a *symbolic encoding* of the dynamics of PADE models, enabling the use of highly efficient model-checking tools for analyzing *incompletely parametrized models*. The symbolic encoding avoids explicit state space generation and the enumeration of possible parametrizations. We demonstrate that the proposed approach scales up to relatively complex synthetic networks. Although developed for PADE models, the main ideas underlying the approach carry over to logical models (Thomas and d'Ari, 1990).

On the biological side, we show the *practical relevance* of the approach by means of an application to the IRMA network. The parameter constraints we obtained are precise, have a clear biological interpretation, and are consistent with independent experimental observations. Even when considering complex dynamical properties, the search of the parameter space takes at most a few minutes. Our results seem to confirm the intended separation of IRMA from the host network, and suggest that to obtain a more robust response to the addition of galactose, an effective rewiring of the network would be needed.

In comparison with traditional quantitative approaches, the results we obtain are quite general, since they do not depend on specific molecular mechanisms or parameter values. Moreover, the analysis is exhaustive in the sense that the entire parameter space is scanned. These two features are particularly interesting for 'negative results', such as showing that a given design is not likely to show a desired behavior. In contrast, quantitative ODE models like those developed in Cantone *et al*. (2009) do not predict a range of possible behaviors but rather single out one likely behavior with quantitative precision. Qualitative and quantitative approaches provide complementary information on system dynamics.

In comparison with other analysis and verification methods developed for similar modeling formalisms (Barnat *et al*., 2009; Bernot *et al*., 2004; Corblin *et al*., 2009; Fromentin *et al*., 2007), our approach is original in two respects. First, it applies to incompletely parametrized models and can handle any dynamical property expressible in temporal logics supported by the model checker. Second, we reason at a finer abstraction level, in that we take into account dynamics on the thresholds and work with a partition of the state space preserving derivative sign patterns. The latter feature is particularly well-suited for the comparison of model predictions with time-series data in IRMA.

An interesting direction for further research is to consider more general problems in which not only parameters but also regulation functions are incompletely specified. This would make a connection with work on the reverse engineering of Boolean models (Martin *et al*., 2007; Perkins *et al*., 2004).

ACKNOWLEDGEMENTS

We would like to thank Delphine Ropers, Maria Pia Cosma and Diego di Bernardo for helpful discussions and contributions.

Funding: The European Commission COBIOS FP6-2005-NEST-PATH-COM/043379; the French ANR Calamar ANR-08-SYSC-003.

Conflict of Interest: none declared.

REFERENCES

Barnat,J. *et al*. (2009) On algorithmic analysis of transcriptional regulation by LTL model checking. *Theor. Comput. Sci.*, **410**, 3128–3148.

Batt,G. *et al*. (2008) Symbolic reachability analysis of genetic regulatory networks using discrete abstractions. *Automatica*, **44**, 982–989.

Batt,G. *et al*. (2005) Validation of qualitative models of genetic regulatory networks by model checking. *Bioinformatics*, **21** (Suppl. 1), i19–i28.

Bernot,G. *et al*. (2004) Application of formal methods to biological regulatory networks. *J. Theor. Biol.*, **229**, 339–348.

Cantone,I. *et al*. (2009) A yeast synthetic network for *in vivo* assessment of reverse-engineering and modeling approaches. *Cell*, **137**, 172–181.

Chaves,M. *et al*. (2009) Geometry and topology of parameter space: investigating measures of robustness in regulatory networks. *J. Math. Biol.*, **59**, 315–358.

Cimatti,A. *et al*. (2002) NuSMV2: an opensource tool for symbolic model checking. In *CAV'02*, Vol. 2404 of *LNCS*. Springer, pp. 359–364.

Clarke,E.M. *et al*. (1999) *Model Checking*. MIT Press, Cambridge, USA.

Corblin,F. *et al*. (2009) A declarative constraint-based method for analyzing discrete genetic regulatory networks. *Biosystems*, **98**, 91–104.

Davidich,M. and Bornholdt,S. (2008) The transition from differential equations to Boolean networks: a case study in simplifying a regulatory network model. *J. Theor. Biol.*, **255**, 269–277.

Edwards,R. and Glass,L. (2006) A calculus for relating the dynamics and structure of complex biological networks. In Berry,R.S. and Jortner,J. (eds) *Adventures in Chemical Physics*, Vol. 132. Wiley, Hobeken, USA, pp. 151–178.

Fauré,A. *et al*. (2006) Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle. *Bioinformatics*, **22**, e124–e131.

Fisher,J. and Henzinger,T.A. (2007) Executable cell biology. *Nat. Biotechnol.*, **25**, 1239–1250.

Fromentin,J. *et al*. (2007) Analysing gene regulatory networks by both constraint programming and model-checking. In *IEEE EMBC07*, pp. 4595–4598.

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

Gouzé,J.-L. and Sari,T. (2002) A class of piecewise linear differential equations arising in biological models. *Dyn. Syst.*, **17**, 299–316.

Martin,S. *et al*. (2007) Boolean dynamics of genetic regulatory networks inferred from microarray time series data. *Bioinformatics*, **23**, 866–874.

Monteiro,P.T. *et al*. (2008) Temporal logic patterns for querying dynamic models of cellular interaction networks. *Bioinformatics*, **24**, i227–i233.

Moore,R.E. (1979) *Methods and Applications of Interval Analysis*. SIAM, Philadelphia, USA.

Perkins,T.J. *et al*. (2004) Inferring models of gene expression dynamics. *J.Theor. Biol.*, **230**, 289–299.

Polynikis,A. *et al*. (2009) Comparing different ODE modelling approaches for gene regulatory networks. *J. Theor. Biol.*, **261**, 511–530.

Saez-Rodriguez,J. *et al*. (2009) Discrete logic modelling as a means to link protein signalling networks with functional analysis of mammalian signal transduction. *Mol. Syst. Biol.*, **5**, 331.

Thomas,R. and d'Ari,R. (1990) *Biological Feedback*. CRC Press, Boca Raton, USA.

Tsai,T.Y.-C. *et al*. (2008) Robust, tunable biological oscillations from interlinked positive and negative feedback loops. *Science*, **321**, 126–129.