Reconstruction of switching thresholds in piecewise-affine models of genetic regulatory networks

Samuel Drulhe, Giancarlo Ferrari-Trecate, Hidde de Jong, Alain Viari

Reconstruction of Switching Thresholds in Piecewise-Affine Models of Genetic Regulatory Networks

S. Drulhe[1], G. Ferrari-Trecate[2,3], H. de Jong[1], and A. Viari[1]

[1] INRIA Rhône-Alpes, 655 avenue de l'Europe, Montbonnot,
38334 Saint Ismier Cedex, France
{Samuel.Drulhe, Hidde.de-Jong, Alain.Viari}@inrialpes.fr
[2] INRIA, Domaine de Voluceau, Rocquencourt - B.P.105,
78153, Le Chesnay Cedex, France
Giancarlo.Ferrari-Trecate@inria.fr
[3] Dipartimento di Informatica e Sistemistica, Università degli Studi di Pavia,
Via Ferrata 1, 27100 Pavia, Italy

Abstract. Recent advances of experimental techniques in biology have led to the production of enormous amounts of data on the dynamics of genetic regulatory networks. In this paper, we present an approach for the identification of PieceWise-Affine (PWA) models of genetic regulatory networks from experimental data, focusing on the reconstruction of switching thresholds associated with regulatory interactions. In particular, our method takes into account geometric constraints specific to models of genetic regulatory networks. We show the feasibility of our approach by the reconstruction of switching thresholds in a PWA model of the carbon starvation response in the bacterium *Escherichia coli*.

1 Introduction

Recent advances of experimental techniques in biology have led to the production of enormous amounts of data on the dynamics of cellular processes. Prominent examples of such techniques are DNA microarrays and gene reporter systems, which allow gene expression to be measured with varying degrees of precision and throughput. One of the major challenges in biology today consists in the analysis and interpretation of these data, with a view to identifying the networks of interactions between genes, proteins, and small molecules that regulate the observed processes. The mapping of these *genetic regulatory networks* is a key issue for understanding the functioning of a cell and for designing interventions of biotechnological or biomedical relevance.

The problem of identifying genetic regulatory networks from gene expression data has attracted much attention over the last ten years. Most approaches are based on the use of *linear* models (*e.g.*, [1, 2, 3]), for which powerful identification algorithms exist. However, given that the underlying biological processes are usually strongly nonlinear, the models are valid only near an equilibrium

J. Hespanha and A. Tiwari (Eds.): HSCC 2006, LNCS 3927, pp. 184–199, 2006.

point (see [4] for an exception). While there have been some approaches based on *nonlinear* models of genetic regulatory networks, the practical applicability of these models is often compromised by the intrinsic mathematical and computational difficulty of nonlinear system identification. Not surprisingly, most authors have therefore focused on specific classes of nonlinear models, with restrictions that reduce the number of parameters and simplify the mathematical form (*e.g.*, [5, 6]).

Another class of models that seems to strike a good compromise between the advantages and disadvantages of linear and nonlinear models are the *PieceWise-Affine (PWA)* models of genetic regulatory networks introduced by Glass and Kauffman in the 1970s [7]. The study of these models and their generalizations has been an active research area in both mathematical biology and hybrid systems theory (*e.g.*, [8, 9, 10, 11, 12, 13]). Notwithstanding their simple mathematical form, PWA systems capture essential aspects of gene regulation, as demonstrated by several modeling studies of regulatory networks of biological interest [12, 14]. Moreover, powerful techniques for the identification of PWA systems have been developed in the field of hybrid systems (see [15] and the references therein), which might be profitably applied to the reconstruction of genetic regulatory networks from experimental data.

Although the available hybrid identification algorithms provide a good starting point, they are generic in nature and therefore not well-adapted to a number of constraints specific to PWA models of genetic regulatory networks. First of all, the state space regions associated with modes of the system are hyperrectangular, as they are defined by switching thresholds of the concentration variables. Second, there exist strong dependencies between the modes of the system, as a consequence of the coordinated control of gene expression. Third, the aim of the system identification process is not to generate a single model, but *all* models with a minimal number of regulatory interactions that are consistent with the experimental data.

The aim of our paper is to make a first step towards the adaptation of existing algorithms for the identification of PWA models so as to take into account the above constraints. In particular, we focus on a crucial stage of the identification process: the estimation of the switching thresholds that partition the state space into hyperrectangular regions. We introduce an algorithm that, given gene expression time-series data classified according to the regulatory modes, produces all minimal sets of switching thresholds. We thus assume here that the preliminary problem of detecting mode switches in time-series data has been solved [15], although we are of course well aware that the underlying classification algorithms will probably have to be tailored to gene expression data as well. In order to illustrate the feasibility of our approach, we apply the threshold reconstruction algorithm to a PWA model of the carbon starvation response in *Escherichia coli* [8, 14]. The gene expression data has been obtained by simulation, while adjusting the noise level and the sampling frequency to the real data that will ultimately be available to us. The work presented in this paper is complementary to the approach of Perkins and colleagues [16], who focus on

the reconstruction of the regulatory modes once the switching thresholds of the system are known.

In the next two sections, we will review PWA models of genetic regulatory networks and discuss the use of hybrid identification techniques for their reconstruction. In Sections 4 to 6 we introduce the notions of cut and multicut, formulate the switching threshold reconstructing problem in terms of these concepts, and introduce a so-called multicut algorithm that, under suitable assumptions, reconstructs minimal sets of switching thresholds from gene expression data. Section 7 presents the results of the multicut algorithm in the context of the *E. coli* carbon starvation model. In the final section we summarize our contributions and indicate directions for further research.

2 Piecewise-Affine Models of Genetic Regulatory Networks

A variety of model formalisms have been proposed to describe the dynamics of genetic regulatory networks (see [17] for a review). One particularly well-adapted to the currently available experimental data is the following class of *PWA differential equations* [7]:

$$\dot{x} = h(x) = f(x) - g(x)x, \tag{1}$$

where $x = [x_1, \ldots, x_n]' \in \Omega \subset \mathbb{R}^n_{\geq 0}$ is a vector of cellular protein concentrations, $f = [f_1, \ldots, f_n]'$, $g = \mathrm{diag}\,(g_1, \ldots, g_n)$, and Ω is a bounded, n-dimensional hyperrectangle. In (1), the rate of change of each protein concentration x_i is the difference of the rate of synthesis $f_i(x)$ and the rate of degradation $g_i(x)x_i$. The map f_i is defined as a sum of terms having the general form $\kappa_i^l\, b_i^l(x)$, where $\kappa_i^l > 0$ is a rate parameter and $b_i^l(x) : \Omega \to \{0, 1\}$ a piecewise-constant function defined in terms of the scalar step functions s^+ and s^- defined as

$$s^+(x_i, \theta_i) = \begin{cases} 1 \text{ if } x_i > \theta_i \\ 0 \text{ if } x_i < \theta_i \end{cases} \quad \textit{and} \quad s^-(x_i, \theta_i) = 1 - s^+(x_i, \theta_i), \tag{2}$$

with $\theta_i > 0$ a constant denoting a threshold concentration for x_i. The step functions are reasonable approximations of sigmoid functions, which represent the switch-like character of the interactions found in gene regulation. The map g_i, which expresses regulation of protein degradation, is defined analogously, except that it is required to be strictly positive. Examples of PWA models of genetic networks are given in [8, 10].

We now show how model (1) can be recast into a standard PWA system. Consider the union of threshold hyperplanes $\Theta = \cup_{i\in\{1,\ldots,n\}, l_i\in\{1,\ldots,p_i\}}\{x \in \Omega : x_i = \theta_i^{l_i}\}$, where p_i denotes the number of thresholds for x_i. Θ splits Ω in open hyperrectangular regions Δ^j, $j = 1, \ldots, s$, $s = \prod_{i=1}^n (p_i + 1)$, called *regulatory domains*. One can show that if $x \in \Delta^j$, then model (1) reduces to $\dot{x} = \mu^j - \nu^j x$,

where $\mu^j = f(x)$ is a constant vector and $\nu^j = g(x)$ is a constant diagonal matrix. In summary, when $x \in \Omega \backslash \Theta$, model (1) is equivalent to the PWA system

$$\dot{x} = h(x) = \mu^j - \nu^j x, \quad \text{if} \quad \lambda(x) = j, \quad j = 1, \ldots, s, \tag{3}$$

where the switching function λ is defined as: $\lambda(x) = j$, if and only if $x \in \Delta^j$. Note that in every domain Δ^j, the map $h(x)$ is affine and in each mode of operation the state variables evolve independently of each other.

3 Hybrid System Identification of Genetic Regulatory Networks

Experimental techniques in biology, like DNA microarrays and gene reporter systems, allow gene expression to be measured at discrete time instants. In what follows, we assume that data are obtained with a uniform sampling period $T > 0$, where T is small with respect to the time constants of gene expression. We denote by $\hat{x}(k)$, $k = 1, \ldots, N+1$, the measured vectors of concentrations $\hat{x}(kT)$. By approximating derivatives through first-order differences, from (3) one obtains the following data model:

$$\hat{x}(k+1) = (I - T\nu^j)\,\hat{x}(k) + T\mu^j + \epsilon(k), \quad \text{if} \quad \lambda(\hat{x}(k)) = j, \tag{4}$$

where $\epsilon(k)$ is an additive noise corrupting the measurements. By focusing on the dynamics of a single protein concentration, say $\hat{x}_i$, model (4) becomes

$$\hat{x}_i(k+1) = \left[\, \hat{x}_i(k)\ 1 \,\right] \phi^j + \epsilon(k), \quad \text{if} \quad \lambda(\hat{x}(k)) = j, \tag{5}$$

where $\phi^j = \left[\, 1 - T(\nu^j)_{ii}\ T(\mu^j)_i \,\right]'$. [1]

Over the last few years, several hybrid system identification algorithms have been proposed for the reconstruction of so-called PieceWise AutoRegressive eXogenous (PWARX) models (see [15] for a review). Without going into details (which can be found in [18]), we just highlight that (5) is a PWARX system with input $u(k) = [\hat{x}_1(k), \ldots, \hat{x}_{l \neq i}(k), \ldots, \hat{x}_n(k)]'$ and output $y(k) = \hat{x}_i(k)$.

The identification of model (5) involves various tasks [15, 18]. In the sequel, we focus on the estimation of the hyperrectangular domains Δ^j, which usually requires an intermediate result produced by all of the above algorithms: the reconstruction of the *switching sequence* $\lambda(\hat{x}(k))$, $k = 1, \ldots, N$. More specifically, as illustrated in [18], a domain Δ^j is found by looking for the $s-1$ hyperplanes separating the set $\mathcal{F}_j = \{\hat{x}(k) : \lambda(\hat{x}(k)) = j\}$ from all sets $\mathcal{F}_l = \{\hat{x}(k) : \lambda(\hat{x}(k)) = l\}$, $l \neq j$. These hyperplanes can be obtained through pattern-recognition techniques such as Multicategory Robust Linear Programming (MRLP) [19] or Support Vector Classifiers (SVC) [20].

A problem with this approach is that both MRLP and SVC do not impose any constraints on the hyperplanes to be estimated. As a consequence, even if the

[1] $(\nu^j)_{ii}$ is the element at position (i,i) of ν^j, $(\mu^j)_i$ is the ith element of μ^j.

switching sequence is perfectly known, there is no guarantee that the estimated domains Δ^j will be hyperrectangular. This may result in hybrid models that are meaningless from a biological point of view, since they do not preserve the concept of a switching threshold associated with a concentration variable. Another problem with existing techniques is that they produce a single model. This is not realistic in our case, because only a fraction of the modes are encountered in the experiments. As a consequence, several hybrid models of the network, each characterized by a different combination of thresholds for the variables, may be consistent with the data and need to be considered.

For all of these reasons, we propose a pattern recognition algorithm tailored to the features of PWARX models of genetic regulatory networks in the next three sections.

4 Switching Thresholds and Multicuts

Let $\mathcal{F}_1, \ldots, \mathcal{F}_s$ be disjoint sets collecting finitely-many points in $\mathbb{R}^n$ and $\mathcal{F}^* = \{\mathcal{F}_1, \ldots, \mathcal{F}_s\}$. Hereafter, we focus on the problem of separating the sets in $\mathcal{F}^*$ with hyperplanes parallel to the linear combination of $n-1$ axes. In order to illustrate the main concepts, we will use the collection $\mathcal{F}^*$ depicted in Figure 1(a). Pairs of distinct sets in $\mathcal{F}^*$ will often be indexed by means of pairs in $U = \{(p,q) \in \{1,\ldots,s\}^2 : p < q\}$.

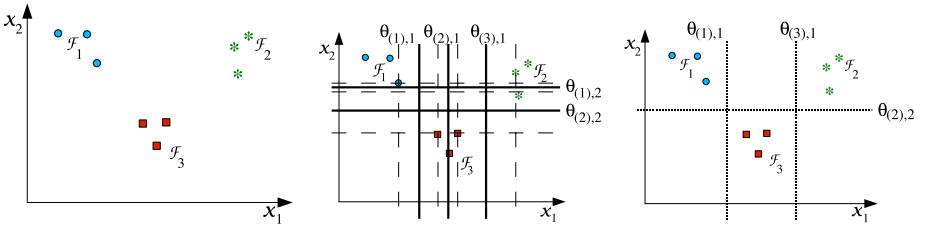


Fig. 1. Simple example of multicuts. (a) Data sets $\mathcal{F}^*$. (b) Multicut $\mathcal{C}^*$: bold lines correspond to cuts and dotted lines are the limits of their equivalence class. (c) Multicut $Max_{\preceq}\mathcal{C}^*$.

Definition 1 (Ap-hyperplane). *An* axis-parallel (ap-) hyperplane *in* $\mathbb{R}^n$ *with direction* $i \in \{1,\ldots,n\}$ *is a hyperplane of equation* $x_i = \alpha$, $\alpha \in \mathbb{R}$, *or equivalently, the zero level set of the function* $\theta(x) = x_i - \alpha$.

By abuse of notation, θ will denote both an ap-hyperplane and its associated function. The function $dir(\theta)$ gives the direction i of the ap-hyperplane θ, while the function $Z(\theta)$ gives the zero-level α. We introduce the following set-valued functions that will turn out to be useful below:

$$\mathcal{I}_-(\theta) = \{j : \forall x \in \mathcal{F}_j, \theta(x) < 0\}, \quad \mathcal{B}_-(\theta) = \cup_{j\in\mathcal{I}_-(\theta)}\mathcal{F}_j,$$
$$\mathcal{I}_+(\theta) = \{j : \forall x \in \mathcal{F}_j, \theta(x) > 0\}, \quad \mathcal{B}_+(\theta) = \cup_{j\in\mathcal{I}_+(\theta)}\mathcal{F}_j.$$

Definition 2 (Separability). *Let $\mathcal{F}_p$ and $\mathcal{F}_q$ be disjoint sets collecting finitely many points in $\mathbb{R}^n$. An ap-hyperplane θ in $\mathbb{R}^n$* separates *$\mathcal{F}_p$ and $\mathcal{F}_q$ if there exists $\delta \in \{+1, -1\}$ such that for all $x \in \mathcal{F}_p \cup \mathcal{F}_q$ one has $\delta\,\theta(x) > 0$, if $x \in \mathcal{F}_p$, and $\delta\,\theta(x) < 0$, if $x \in \mathcal{F}_q$. In this case, we write $\mathcal{F}_p \overset{\theta}{\curlyvee} \mathcal{F}_q$. $\mathcal{F}_p$ and $\mathcal{F}_q$ are* separable *if there exists an ap-hyperplane separating the sets.*

We introduce two additional functions on sets $\mathcal{F}_p$ and $\mathcal{F}_q$, for $i \in \{1, \ldots, n\}$,

$$Inf_i(\mathcal{F}_p, \mathcal{F}_q) = \min(\max_{x \in \mathcal{F}_p} x_i, \max_{x \in \mathcal{F}_q} x_i),$$
$$Sup_i(\mathcal{F}_p, \mathcal{F}_q) = \max(\min_{x \in \mathcal{F}_p} x_i, \min_{x \in \mathcal{F}_q} x_i).$$

In Figure 1, $\mathcal{F}_1$ and $\mathcal{F}_2$ are separable since there exist ap-hyperplanes in the x_1-direction (*e.g.*, $\theta_{(1),1}$ and $\theta_{(2),1}$), such that all points in $\mathcal{F}_1$ lie on one side of the hyperplane $\theta_{(1),1}$ and all points of $\mathcal{F}_2$ on the other side. Notice that the sets $\mathcal{F}_1$ and $\mathcal{F}_2$ are not separable in the x_2-direction. As can be verified in Figure 1, the ap-hyperplane $\theta_{(1),1}$ separates more sets than the ap-hyperplane $\theta_{(2),1}$. The difference in separation power of ap-hyperplanes can be formally defined as follows.

Definition 3 (Separation power). *The* separation power *of an ap-hyperplane θ is the set-valued function $S(\theta) = \{(p, q) \in U : \mathcal{F}_p \overset{\theta}{\curlyvee} \mathcal{F}_q\}$.*

In the remainder of this section, we focus on ap-hyperplanes in the set $\Theta = \{\theta : S(\theta) \neq \emptyset\}$. The comparison of the separation power of ap-hyperplanes in Θ in a given direction motivates the introduction of equivalence classes of ap-hyperplanes.

Definition 4 (Equivalence). *Two ap-hyperplanes $\theta, \theta' \in \Theta$ are* equivalent *if $dir(\theta) = dir(\theta')$ and $S(\theta) = S(\theta')$. Equivalent ap-hyperplanes will be denoted by $\theta \sim \theta'$ and the equivalence class of θ by $[\theta] = \{\theta' : \theta' \sim \theta\}$.*

Following the above definition, the ap-hyperplanes $\theta_{(1),1}$ and $\theta_{(2),1}$ in Figure 1 are not equivalent.

We recall that, given an equivalence relation $\sim$ on a set X and a function $f : X \to Y$, f is *invariant* under $\sim$ if $x \sim y$ implies $f(x) = f(y)$. It is not difficult to show that the functions dir, S, $\mathcal{I}_+$, $\mathcal{I}_-$, $\mathcal{B}_+$ and $\mathcal{B}_-$ are invariant under the equivalence relation $\sim$ defined in Definition 4. This implies that we can generalize these functions to the *quotient set* $\mathcal{E}^* = \Theta/\sim$. Note also that the cardinality of $\mathcal{E}^*$ is finite [21].

Although all ap-hyperplanes in an equivalence class $\mathcal{E} \in \mathcal{E}^*$ have the same separation power, only one is optimal in a statistical sense [20]. This ap-hyperplane will be called a *cut*.

Definition 5 (Cut). *Let $\mathcal{E} \in \mathcal{E}^*$ and $i = dir(\mathcal{E})$. The* cut *associated to $\mathcal{E}$ is the ap-hyperplane $\theta \in \Theta$ such that*

$$Z(\theta) = Inf_i(\mathcal{B}_+(\mathcal{E}), \mathcal{B}_-(\mathcal{E})) + \frac{Sup_i(\mathcal{B}_+(\mathcal{E}), \mathcal{B}_-(\mathcal{E})) - Inf_i(\mathcal{B}_+(\mathcal{E}), \mathcal{B}_-(\mathcal{E}))}{2}. \quad (6)$$

In what follows the set of all cuts is denoted by $\mathcal{C}^*$. Since $\mathcal{E}^*$ and $\mathcal{C}^*$ are isomorphic, the cardinality of $\mathcal{C}^*$ is also finite. In the example with three data sets in Figure 1(a), $\mathcal{C}^*$ is composed of five cuts ($\theta_{(1),1}$, $\theta_{(2),1}$, $\theta_{(3),1}$, $\theta_{(1),2}$, and $\theta_{(2),2}$), which are represented in Figure 1(b) by means of bold lines.

Intuitively, we would be inclined to say that the cut $\theta_{(1),1}$ is more powerful than $\theta_{(2),1}$, in the sense that the former separates $\mathcal{F}_1$ and $\mathcal{F}_2$ as well as $\mathcal{F}_1$ and $\mathcal{F}_3$, whereas the latter separates only $\mathcal{F}_1$ and $\mathcal{F}_2$ (that is, $S(\theta_{(1),1}) = \{(1,2),(1,3)\}$ and $S(\theta_{(2),1}) = \{(1,2)\}$). This motivates the introduction of the following relation on $\mathcal{C}^*$, denoted by $\preceq$:

$$\theta \preceq \theta' \text{ if } S(\theta) \subseteq S(\theta') \text{ and } dir(\theta) = dir(\theta'). \tag{7}$$

It is straightforward to show that $\preceq$ is reflexive, antisymmetric, and transitive, and hence that $\preceq$ is a partial order on $\mathcal{C}^*$. That is, $\mathcal{C}^*$ is a poset (partially ordered set).

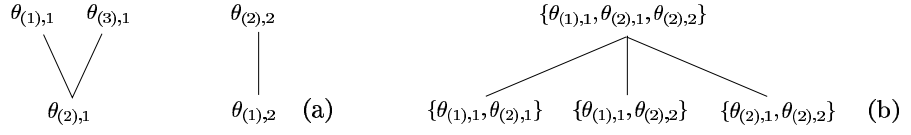


Fig. 2. (a) Poset diagram for the set of cuts $\mathcal{C}^*$ in Figure 1. The diagram shows, *e.g.*, $\theta_{(2),1} \preceq \theta_{(1),1}$. (b) Poset diagram for the down-set of $\mathcal{M} = \{\theta_{(1),1}, \theta_{(3),1}, \theta_{(2),2}\}$, which is a multicut for Figure 1. In fact, $\mathcal{M}$ equals $\text{Max}_{\preceq}\, \mathcal{C}^*$.

The poset diagram corresponding to the example in Figure 1 is shown in Figure 2(a). As for any poset, $\mathcal{C}^*$ admits maximal and minimal elements. The sets of maximal and minimal elements of $\mathcal{C}^*$ are denoted by $\text{Max}_{\preceq}\, \mathcal{C}^*$ and $\text{Min}_{\preceq}\, \mathcal{C}^*$, respectively. For instance, in Figure 2(a) $\text{Max}_{\preceq}\, \mathcal{C}^* = \{\theta_{(1),1}, \theta_{(3),1}, \theta_{(2),2}\}$.

In general, several cuts will be required to separate all sets in $\mathcal{F}^*$. This motivates the introduction of multicuts.

Definition 6 (Multicut). *A* multicut $\mathcal{M}$ *of* $\mathcal{F}^*$ *is a finite set of cuts such that for all* $(p,q) \in U$ *there exists a* $\theta \in \mathcal{M}$*, such that* $\mathcal{F}_p \overset{\theta}{\curlyvee} \mathcal{F}_q$*. A collection* $\mathcal{F}^*$ *is said to be* m-separable *if there exists a multicut of* $\mathcal{F}^*$ *or, equivalently, if* $U = \cup_{\theta \in \mathcal{M}} S(\theta)$*.*

We call $\mathcal{M}^*$ the set of multicuts. Due to the fact that $\mathcal{C}^*$ is finite, $\mathcal{M}^*$ is finite as well. Notice that $\mathcal{M}^*$ may be empty, that is, $\mathcal{F}^*$ may not be m-separable. In the example of Figure 1, $\mathcal{M} = \{\theta_{(3),1}, \theta_{(2),2}\}$ is a multicut since we have $S(\theta_{(3),1}) = \{(1,2),(2,3)\}$ and $S(\theta_{(2),2}) = \{(1,3)\}$.

The following proposition, proven in [21], states a relevant property of $\mathcal{C}^*$.

Proposition 1. *$\mathcal{F}^*$ is m-separable if and only if $\mathcal{C}^*$ is a multicut.*

We define an obvious partial order relation on the set of multicuts $\mathcal{M}^*$, the set inclusion $\subseteq$. The poset $\mathcal{M}^*$ for the example in Figure 1 consists of 20 multicuts (figure not shown).

To every subset $\mathcal{B}$ of $\mathcal{M}^*$ we can associate a down-set, which consists of the multicuts in $\mathcal{M}^*$ upper bounded (according to $\subseteq$) by some multicut in $\mathcal{B}$. For reasons that will become clear below, we focus here on the down-set of singletons $\mathcal{B} = \{\mathcal{M}\}$, for some $\mathcal{M} \in \mathcal{M}^*$.

Definition 7 (Down-set of multicut set). *The* down-set *of* $\{\mathcal{M}\}$, $\mathcal{M} \in \mathcal{M}^*$, *denoted by* $\downarrow \{\mathcal{M}\}$, *is defined by* $\downarrow \{\mathcal{M}\} = \{\mathcal{M}' \in \mathcal{M}^* : \mathcal{M}' \subseteq \mathcal{M}\}$.

Consider the multicut $\mathrm{Max}_{\preceq} \mathcal{C}^*$ in the example (Figure 2(a)). The down-set of $\{\mathrm{Max}_{\preceq} \mathcal{C}^*\}$ is the union of all sets appearing in Figure 2(b). We note that $\downarrow \{\mathcal{M}\}$ is also a poset with respect to set inclusion.

5 Formulation of Switching Threshold Reconstruction Problem

The introduction of the concepts of cut and multicut, and the partial orders defined on them, allows us to formulate the problem of reconstructing switching thresholds in a more precise way. In general, the available data are consistent with a large number of multicuts, and thus with a large number of PWA models of the genetic regulatory network. *A priori* there is no reason to prefer one of these models above the others. However, in practice we are most interested in the minimal models that account for the available data, that is, those models that contain a minimal number of thresholds and separate all pairs of sets in $\mathcal{F}^*$. Assuming that the set of data points is m-separable, so that $\mathcal{C}^*$ is a multicut, it seems reasonable to accept as solutions all multicuts in $\mathrm{Min}_{\subseteq} \downarrow \{\mathcal{C}^*\}$.

Notice though that $\mathcal{C}^*$ may contain many cuts with a weak separation power that could be eliminated beforehand if we are only interested in finding minimal multicuts. That is, we can remove cuts $\theta \in \mathcal{C}^*$ if there exists another $\theta' \in \mathcal{C}^*$, $\theta' \neq \theta$, such that $\theta \preceq \theta'$. Eliminating these cuts does not affect the m-separability of the sets of data points, as indicated by the following proposition (proven in [21]), which should be compared with Proposition 1.

Proposition 2. *$Max_{\preceq} \mathcal{C}^*$ is a multicut if and only if $\mathcal{F}^*$ is m-separable.*

Once $\mathcal{C}^*$ has been reduced to $\mathrm{Max}_{\preceq} \mathcal{C}^*$, our switching threshold reconstruction problem can be recast into the problem of computing the set

$$\mathrm{Min}_{\subseteq} \downarrow \{\mathrm{Max}_{\preceq} \mathcal{C}^*\}. \tag{8}$$

Notice that $\mathrm{Max}_{\subseteq} \downarrow \{\mathrm{Max}_{\preceq} \mathcal{C}^*\}$ is $\{\mathrm{Max}_{\preceq} \mathcal{C}^*\}$ itself, so that we will call $\mathrm{Max}_{\preceq} \mathcal{C}^*$ the *maximal multicut*. In the example of Figure 1, $\mathrm{Max}_{\preceq} \mathcal{C}^*$ consists of three cuts, as shown in Figure 2(a). That is, two cuts with obvious weaker separation power have been eliminated ($\theta_{(2),1}$ and $\theta_{(1),2}$). The down-set of $\{\mathrm{Max}_{\preceq} \mathcal{C}^*\}$ is shown in Figure 2(b). It has three minimal multicuts: $\{\theta_{(1),1}, \theta_{(3),1}\}$, $\{\theta_{(1),1}, \theta_{(2),2}\}$, and $\{\theta_{(3),1}, \theta_{(2),2}\}$. As illustrated by the example, there will generally be several minimal multicuts. We can distinguish between *locally* and *globally* minimal multicuts.

Definition 8. *Let $\mathcal{M}$ be a multicut of $\mathcal{F}^*$. $\mathcal{M}$ is* locally minimal *if for all $\theta \in \mathcal{M}$, the set $\mathcal{M}\backslash\{\theta\}$ is not a multicut of $\mathcal{F}^*$. $\mathcal{M}$ is* globally minimal *if*

$$|\mathcal{M}| = \min_{\tilde{\mathcal{M}} \in \mathcal{M}_{min}} |\tilde{\mathcal{M}}|, \tag{9}$$

where $\mathcal{M}_{min}$ is the set of all locally minimal multicuts of $\mathcal{F}^*$.

It can be shown (see [21]) that the elements of $\mathrm{Min}_{\subseteq} \downarrow \{\mathrm{Max}_{\preceq} \mathcal{C}^*\}$ are locally minimal multicuts, but they are not necessarily globally minimal.

The above remarks lead us to a final refinement of the problem statement:

$$\text{find } \textit{all} \text{ globally minimal multicuts in } \mathrm{Min}_{\subseteq} \downarrow \{\mathrm{Max}_{\preceq} \mathcal{C}^*\}. \tag{10}$$

6 Algorithms for Computing Switching Thresholds

In this section we present an approach to compute the multicuts satisfying criterion (10), and thus infer the minimal set of switching thresholds for a PWA model of a genetic regulatory network from a classified data set $\mathcal{F}^*$.

The computation of the set of all cuts ($\mathcal{C}^*$) is rather straightforward, based on the definition of a cut (Definition 5). For sake of brevity, we omit the algorithm which can be found in [21]. Similarly, the set of maximal cuts ($\mathrm{Max}_{\preceq} \mathcal{C}^*$) can be computed by applying directly the definition of maximal element of $\mathcal{C}^*$ with respect to the partial order (7) (see [21] for further details).

A more challenging task is the computation of all globally minimal multicuts. In order to find them, we could in principle enumerate all subsets of $Max_{\preceq} \mathcal{C}^*$ and verify minimality by means of Definitions 6 and 8. However, this procedure is computationally prohibitive even for simple examples. Therefore, in the sequel, we present an additional result on multicuts that will allow us to reduce the dimension of the search space.

Definition 9 (Redundancy). *Let $\mathcal{M}$ be a multicut of $\mathcal{F}^*$. A cut $\theta \in \mathcal{M}$ is* redundant *in $\mathcal{M}$, if $S(\theta) \subseteq \cup_{\theta' \in \mathcal{M}\backslash\{\theta\}} S(\theta')$.*

In the example of Figure 1, each of the three cuts in the multicut $\{\theta_{(1),1}, \theta_{(3),1}, \theta_{(2),2}\}$ is redundant. The following proposition (proven in [21]), shows that redundant cuts can be safely ignored.

Proposition 3. *A multicut $\mathcal{M}$ of $\mathcal{F}^*$ is locally minimal if and only if no $\theta \in \mathcal{M}$ is redundant in $\mathcal{M}$.*

Definition 10 (Kernel). *Let $\mathcal{M}$ be a multicut of $\mathcal{F}^*$. The* kernel *of $\mathcal{M}$ is defined as* $\ker(\mathcal{M}) = \{\theta \in \mathcal{M} : \exists u \in S(\theta), \nexists \theta' \in \mathcal{M} \setminus \{\theta\}, u \in S(\theta')\}$.

From Definition 10, it is apparent that $\ker(Max_{\preceq} \mathcal{C}^*)$ collects the cuts in $\mathcal{M}$ that must belong to every minimal multicut, otherwise at least one pair of sets in $\mathcal{F}^*$ will not be separated. In the case of $\mathcal{M} = \{\theta_{(1),1}, \theta_{(3),1}, \theta_{(2),2}\}$ in the example of Figure 1, the kernel is empty: none of the cuts is indispensable.

Algorithm 1. Create the set $\mathcal{M}^*_{min}$ of all globally minimal multicuts

1: Initialize the global variables $\mathcal{M}^*_{min} = \emptyset$ and $best = |Max_{\preceq}\mathcal{C}^*|$. Initialize $\mathcal{M}_{in} = \ker(Max_{\preceq}\mathcal{C}^*)$
2: **if** $U = \cup_{\theta\in\mathcal{M}_{in}} S(\theta)$ **then**
3: Append $\ker(Max_{\preceq}\mathcal{C}^*)$ to $\mathcal{M}^*_{min}$ and exit
4: **else**
5: $Branch(\mathcal{M}_{in})$
6: **end if**

function $Branch(\mathcal{M}_{in})$

1: **for all** $\theta \in Max_{\preceq}\mathcal{C}^* \backslash \mathcal{M}_{in}$ **do**
2: **if** $S(\theta) \not\subseteq \cup_{\theta'\in\mathcal{M}_{in}} S(\theta')$ **then** //θ is not redundant in $\mathcal{M}_{in} \cup \{\theta\}$.
3: Set $\mathcal{M}_{out} = \mathcal{M}_{in} \cup \{\theta\}$
4: **if** $U = \cup_{\theta'\in\mathcal{M}_{out}} S(\theta')$ **then** //$\mathcal{M}_{out}$ is a multicut.
5: **if** $|\mathcal{M}_{out}| = best$ and $\mathcal{M}_{out} \notin \mathcal{M}^*_{min}$ **then**
6: Append $\mathcal{M}_{out}$ to $\mathcal{M}^*_{min}$
7: **else if** $|\mathcal{M}_{out}| < best$ **then**
8: Set $\mathcal{M}^*_{min} = \{\mathcal{M}_{out}\}$ and $best = |\mathcal{M}_{out}|$ //Reset $\mathcal{M}^*_{min}$ and update *best*.
9: **end if**
10: **else if** $|\mathcal{M}_{out}| < best$ **then**
11: $Branch(\mathcal{M}_{out})$
12: **end if**
13: **end if**
14: **end for**

The notions of redundancy and kernel are used to speed up the branch-and-bound strategy of Algorithm 1 below, computing the set $\mathcal{M}^*_{min} \subseteq \mathcal{M}^*$ of globally minimal multicuts. The basic idea is to start with a small subset of $Max_{\preceq}\mathcal{C}^*$, given by $\ker(Max_{\preceq}\mathcal{C}^*)$, and add new cuts iteratively.

During the execution of Algorithm 1, the global variable *best* stores the size of the smaller multicut found so far. If $\ker(Max_{\preceq}\mathcal{C}^*)$ is a multicut, it is also the only globally minimal multicut in $Max_{\preceq}\mathcal{C}^*$ and the algorithm terminates (lines 1 and 1 of the main procedure). Otherwise, the function *Branch* is called in order to add suitable cuts to $\ker(Max_{\preceq}\mathcal{C}^*)$. At line 1 of the function *Branch*, the addition of a new cut θ to $\mathcal{M}_{in}$ is considered only if θ is not redundant in $\mathcal{M}_{out} = \mathcal{M}_{in} \cup \{\theta\}$ (following Proposition 3). Lines 1-1 process sets $\mathcal{M}_{out}$ that are multicuts and modify the set $\mathcal{M}^*_{min}$ accordingly. More specifically, a multicut of size *best* is added to $\mathcal{M}^*_{min}$ (line 1), while a multicut of size less than *best* causes the reset of the set $\mathcal{M}^*_{min}$ (line 1) and the update of *best*. These operations guarantee that only globally minimal multicuts will be stored in $\mathcal{M}^*_{min}$.

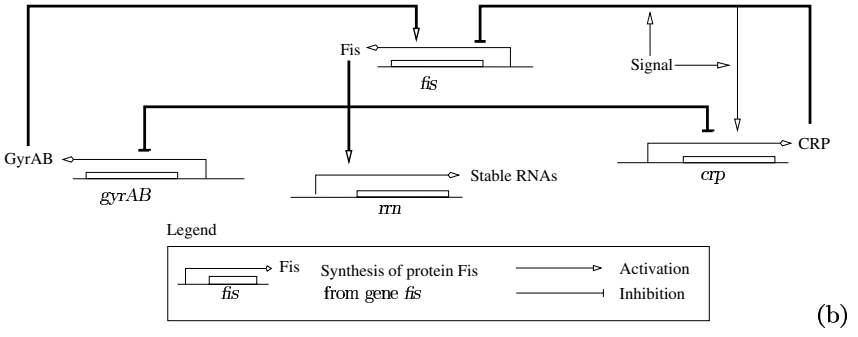
7 Reconstruction of Switching Thresholds in PWA Model of Carbon Starvation Response of *E. coli*

In order to test the applicability of the multicut approach, we have used it for the reconstruction of switching thresholds in a PWA model of the carbon starvation

response in the bacterium *Escherichia coli*. In the absence of essential carbon sources, an *E. coli* population abandons exponential growth and enters a non-growth state called *stationary phase*. On the molecular level, the transition from exponential phase to stationary phase in response to a carbon stress is controlled by a complex genetic regulatory network.

A PWA model of the carbon starvation response has been developed in *E. coli* [14]. The model describes how a carbon stress signal is propagated through a network of interactions between global transcriptional regulators of the bacterium, so as to influence the synthesis of stable RNAs and thereby adapt the growth of the cell. For this study, we have used a simplified version of this model (Figure 3), which preserves essential properties of the qualitative dynamics

$$
\begin{aligned}
\dot{x}_{CRP} &= \kappa^0_{CRP} + \kappa^1_{CRP}\, s^-(x_{Fis}, \theta^1_{Fis})\, s^+(x_{CRP}, \theta^1_{CRP})\, s^+(x_S, \theta_S) - \gamma_{CRP}\, x_{CRP} \\
\dot{x}_{Fis} &= \kappa^1_{Fis}\,(1 - s^+(x_{CRP}, \theta^1_{CRP})\, s^+(x_S, \theta_S)) \\
&\quad + \kappa^2_{Fis}\, s^+(x_{GyrAB}, \theta_{GyrAB})\,(1 - s^+(x_{CRP}, \theta^1_{CRP})\, s^+(x_S, \theta_S)) - \gamma_{Fis}\, x_{Fis} \\
\dot{x}_{GyrAB} &= \kappa_{GyrAB}\, s^-(x_{Fis}, \theta^3_{Fis}) - \gamma_{GyrAB}\, x_{GyrAB} \\
\dot{x}_{rrn} &= \kappa_{rrn}\, s^+(x_{Fis}, \theta^2_{Fis}) - \gamma_{rrn}\, x_{rrn} \\
\dot{x}_S &= 0
\end{aligned}
\qquad \text{(a)}
$$



(b)

Fig. 3. (a) Simplified PWA model of the carbon starvation network in *E. coli* [14]. The variables x_{CRP}, x_{Fis}, x_{GyrAB}, and x_{rrn} denote the concentrations of CRP, Fis, GyrAB, and stable RNAs, while x_S represents the carbon starvation signal ($s^+(x_S, \theta_S) = 1$ means that the carbon starvation signal is present). The variables have been rescaled to the interval $[0, 1]$, and the following parameter values have been used for the simulations: $\theta^1_{CRP} = 0.33$, $\theta^2_{CRP} = 0.67$, $\theta^1_{Fis} = 0.1$, $\theta^2_{Fis} = 0.5$, $\theta^3_{Fis} = 0.75$, $\theta_{GyrAB} = 0.5$, $\theta_{rrn} = 0.5$, $\theta_S = 0.5$, $\gamma_{CRP} = 0.5$;, $\gamma_{Fis} = 2$, $\gamma_{GyrAB} = 1$, $\gamma_{rrn} = 1.5$, $\gamma_S = 0.5$, $\kappa^0_{CRP} = 0.25$, $\kappa^1_{CRP} = 0.4$, $\kappa^1_{Fis} = 0.6$, $\kappa^2_{Fis} = 1.15$, $\kappa_{GyrAB} = 0.75$, $\kappa_{rrn} = 1.12$, (b) Graphical representation of the PWA model, indicating genes and their regulatory interactions. The interactions in bold have been correctly identified by the best globally minimal multicuts obtained from the data for the reentry into exponential phase after a carbon upshift (MC_2 in Figure 5(c)) and for the entry into stationary phase (results not shown).

predicted by the original model, as verified by means of the approach described in [8]. In response to a carbon starvation signal, the system switches from an equilibrium point characteristic for exponential growth to another equilibrium point, corresponding to stationary phase. Reentry into exponential phase after a carbon upshift gives rise to a damped oscillation towards the exponential-phase equilibrium point.

The use of reporter genes encoding fluorescent and luminescent proteins makes it possible to obtain precise and densely-spaced measurements of the expression of the genes in the carbon starvation response network. This kind of data is well-suited for system identification purposes, as shown previously in [2, 6]. In this paper, we use simulated data to test the multicut approach, staying close to the expected noise and sample density of the real measurements.

Figure 4 gives an indication of the data obtained from simulating the reentry into exponential phase after a carbon upshift. In order to separate the threshold reconstruction problem from the classification problem for the purpose of this paper, we have generated the correct classification by detecting mode switches during simulation.

The resulting datasets have been analyzed by means of a Matlab implementation of the algorithms presented in Section 6. The results for the transition from stationary to exponential phase after a carbon upshift are summarized in Figure 5. The algorithm finds the maximal multicut $\mathcal{C}^*$, consisting of six cuts $(\theta_1, \ldots, \theta_6)$. In order to get an idea of the separation power of the cuts, Figure 5(b) pictures the projection of the data points on the (x_{Fis}, x_{GyrAB})-subspace. As can be seen, the cuts θ_2, θ_5, and θ_6 nicely separate the classes generated from the damped oscillation (Figure 4).

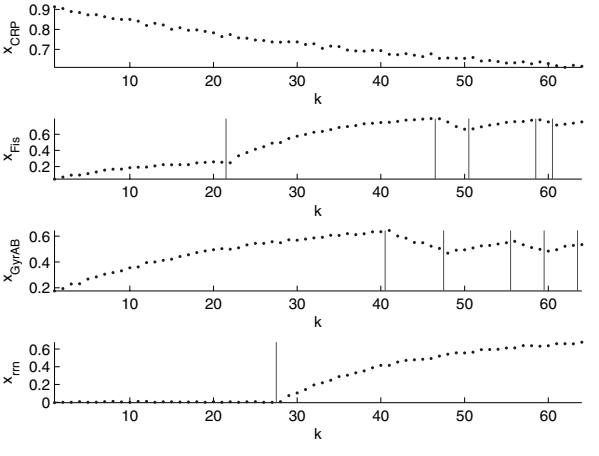
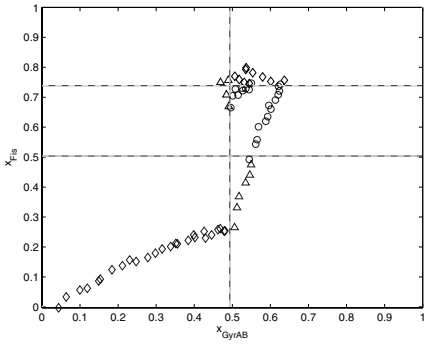


Fig. 4. Simulation of the reentry into exponential phase following a carbon upshift, using the PWA model in Figure 3(a). In order to mimic the absence of a carbon stress, $x_S(0)$ has been set to 0. For each protein concentration variable, the mode switches are indicated by means of vertical bars.

Cut	Variable	Threshold value	Interaction	Correct? (Y/N)
θ_1	x_{Fis}	0.26	Fis activates *fis*	N
θ_2	x_{GyrAB}	0.49	GyrAB activates *fis*	Y
θ_3	x_{rrn}	0.03	Stable RNAs activate *rrn*	N
θ_4	x_{CRP}	0.65	CRP inhibits *fis*	Y
θ_5	x_{Fis}	0.5	Fis activates *rrn*	Y
θ_6	x_{Fis}	0.74	Fis inhibits *gyrAB*	Y

(a)



(b)

Multicut	Composing cuts	Correct?
MC_1	$\{\theta_2, \theta_3, \theta_6\}$	$\{Y, N, Y\}$
MC_2	$\{\theta_2, \theta_4, \theta_6\}$	$\{Y, Y, Y\}$
MC_3	$\{\theta_2, \theta_5, \theta_6\}$	$\{Y, Y, Y\}$

(c)

Fig. 5. (a) Maximal multicut for the data in Figure 4. (b) Illustration of the separation power of the cuts θ_2, θ_5, and θ_6, included in the globally minimal multicut MC_3 in (c). The data have been projected on the (x_{Fis}, x_{GyrAB})-subspace. (c) Globally minimal multicuts generated by Algorithm 1 from the maximal multicut in (a).

To each of the cuts corresponds a switching threshold, associated with a regulatory interaction in the network. For instance, one can verify in Figure 4 that when x_{Fis} crosses the threshold value 0.5 from below, the concentration x_{rrn} of stable RNAs starts to increase as well. This motivates the conclusion that the threshold where x_{Fis} equals 0.5 corresponds to the activation of the *rrn* operon by Fis, an interaction that is correctly inferred from the simulation data (Figure 4). Four of the cuts in the maximal multicut correspond to real switching thresholds of the system.

Applying Algorithm 1 to the maximal multicut yields three globally minimal multicuts, shown in Figure 5(c). Each of the multicuts consists of three cuts, two of which occur in every solution. The cut θ_6 corresponds to the switching threshold above which Fis starts to inhibit the expression of the gene *gyrAB*, while θ_2 represents the switching threshold associated with the activation of *fis* by GyrAB. Notice that the globally minimal multicuts MC_2 and MC_3 contain only cuts corresponding to correct switching thresholds, whereas for MC_1 two out of three thresholds are correct.

Repeating the above procedure for the second set of simulation data, corresponding to the entry into stationary phase, yields a maximal multicut consisting of four cuts, three of which correspond to a real switching threshold of the system (results not shown). From this information, Algorithm 1 generates four globally

minimal multicuts, each composed of two cuts. Two of the globally minimal multicuts entirely consist of cuts corresponding to correct switching thresholds, whereas in the other two cases one of the cuts corresponds to a non-existing threshold.

Summarizing the results of the switching threshold reconstruction process, the best globally minimal multicuts for the first and second data series have been projected on the graphical representation of the carbon starvation network in Figure 3. As can be seen, the multicut approach has inferred five out of six interactions from the data (only the autoactivation of CRP is missing). As for the worst globally minimal multicuts found by the algorithm, they nevertheless achieve the correct identification of three of the switching thresholds in the model. These results confirm the in-principle applicability of our approach.

8 Conclusions

In this paper we have proposed a pattern recognition technique for reconstructing all combinations of switching thresholds that are consistent with measured data in PWA models of genetic regulatory networks. We have shown how to recast this problem into finding all globally minimal multicuts of maximal cuts that separate different sets of points within a given collection. This algorithm is intended to be used in combination with hybrid identification procedures for classifying the data (*i.e.*, partitioning temporal gene expression data into subsets associated with different regulatory modes) and for reconstructing the values of synthesis/degradation parameters characterizing the dynamics of the network in different regulatory domains.

A potential pitfall of the multicut approach is that the algorithms presented in Section 6 have been derived under the assumption that the sets of points considered are m-separable. Although this assumption is satisfied in the example of Section 7, it may be violated in other situations for two main reasons. The first one is that noisy data may affect the quality of the results obtained through hybrid systems identification, and lead to a misclassification of some data points [15]. The second reason is that genetic regulatory networks may exhibit the same dynamics on different regulatory domains, a fact that may result in a structural loss of m-separability. However, we stress that even if some pairs of sets are not separable, this does not prevent the multicut algorithm from finding *some* of the thresholds. Most importantly, the m-separability assumption can be verified once $\mathcal{C}^*$ has been found. We also believe that even if the mathematical framework for multicuts developed in Sections 4 to 6 is tailored to an idealized case, it provides a sound background for developing new methods capable of dealing with m-inseparable collections of sets.

Acknowledgments. This research has been supported by the European Commission under project HYGEIA (NEST-4995).

References

1. D'haeseleer, P., Liang, S., Somogyi, R.: Genetic network inference: From co-expression clustering to reverse engineering. Bioinformatics **16** (2000) 707–726
2. Gardner, T., di Bernardo, D., Lorenz, D., Collins, J.: Inferring genetic networks and identifying compound mode of action via expression profiling. Science **301** (2003) 102–105
3. van Someren, E., Wessels, L., Reinders, M.: Linear modeling of genetic networks from experimental data. In Altman, R., et *al.*, eds.: Proc. Eight Int. Conf. Intell. Syst. Mol. Biol., ISMB 2000, Menlo Park, CA, AAAI Press (2000) 355–366
4. Lemeille, S., Latifi, A., Geiselmann, J.: Inferring the connectivity of a regulatory network from mRNA quantification in *Synechocystis* PCC6803. Nucleic Acids Res. **33** (2005) 3381–3389
5. Jaeger, J., Surkova, S., Blagov, M., Janssens, H., Kosman, D., Kozlov, K., Manu, Myasnikova, E., Vanario-Alonso, C., Samsonova, M., Sharp, D., Reinitz, J.: Dynamic control of positional information in the early *Drosophila* embryo. Nature **430** (2004) 368–371
6. Ronen, M., Rosenberg, R., Shraiman, B., Alon, U.: Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics. Proc. Natl. Acad. Sci. USA **99** (2002) 10555–10560
7. Glass, L., Kauffman, S.: The logical analysis of continuous non-linear biochemical control networks. J. Theor. Biol. **39** (1973) 103–129
8. Batt, G., Ropers, D., de Jong, H., Geiselmann, J., Page, M., Schneider, D.: Qualitative analysis and verification of hybrid models of genetic regulatory networks: Nutritional stress response in *Escherichia coli*. In Morari, M., Thiele, L., eds.: Proc. Hybrid Systems: Computation and Control (HSCC 2005). Volume 3414 of LNCS. Springer-Verlag, Berlin (2005) 134–150
9. Belta, C., Finin, P., Habets, L., Halász, A., Imiliǹski, M., Kumar, R., Rubin, H.: Understanding the bacterial stringent response using reachability analysis of hybrid systems. In Alur, R., Pappas, G., eds.: Proc. Hybrid Systems: Computation and Control (HSCC 2004). Volume 2993 of LNCS. Springer-Verlag, Berlin (2004) 111–125
10. de Jong, H., Gouzé, J.L., Hernandez, C., Page, M., Sari, T., Geiselmann, J.: Qualitative simulation of genetic regulatory networks using piecewise-linear models. Bull. Math. Biol. **66** (2004) 301–340
11. Edwards, R., Siegelmann, H., Aziza, K., Glass, L.: Symbolic dynamics and computation in model gene networks. Chaos **11** (2001) 160–169
12. Ghosh, R., Tomlin, C.: Symbolic reachable set computation of piecewise affine hybrid automata and its application to biological modelling: Delta-Notch protein signalling. Syst. Biol. **1** (2004) 170–183
13. Mestl, T., Plahte, E., Omholt, S.: A mathematical framework for describing and analysing gene regulatory networks. J. Theor. Biol. **176** (1995) 291–300
14. Ropers, D., de Jong, H., Page, M., Schneider, D., Geiselmann, J.: Qualitative simulation of the carbon starvation response in *Escherichia coli*. BioSystems (2006) In press.
15. Juloski, A., W.P.M.H. Heemels, W., Ferrari-Trecate, G., Vidal, R., Paoletti, S., Niessen, J.: Comparison of four procedures for the identification of hybrid systems. In Morari, M., Thiele, L., eds.: Proc. Hybrid Systems: Computation and Control (HSCC-05). Volume 3414 of LNCS., Springer-Verlag, Berlin (2005) 354–369

16. Perkins, T., Hallett, M., Glass, L.: Inferring models of gene expression dynamics. J. Theor. Biol. **230** (2004) 289–299
17. de Jong, H.: Modeling and simulation of genetic regulatory systems: A literature review. J. Comput. Biol. **9** (2002) 67–103
18. Ferrari-Trecate, G., Muselli, M., Liberati, D., Morari, M.: A clustering technique for the identification of piecewise affine and hybrid systems. Automatica **39** (2003) 205–217
19. Bennett, K., Mangasarian, O.: Multicategory discrimination via linear programming. Optimization Methods and Software **3** (1993) 27–39
20. Vapnik, V.: Statistical Learning Theory. John Wiley, NY (1998)
21. Drulhe, S., Ferrari-Trecate, G., de Jong, H., Viari, A.: Reconstruction of switching thresholds in piecewise-affine models of genetic regulatory networks. Technical report, INRIA (2005) `http://www.inria.fr/rrrt/index.en.html`.