Proceedings of the 17th World Congress
The International Federation of Automatic Control
Seoul, Korea, July 6-11, 2008

Identification of PieceWise Affine Models of Genetic Regulatory Networks: the Data Classification Problem ⋆

Riccardo Porreca * **Giancarlo Ferrari-Trecate** *

* *Dipartimento di Informatica e Sistemistica,*
Università degli Studi di Pavia, Via Ferrata 1, 27100 Pavia, Italy
(e-mail: {riccardo.porreca, giancarlo.ferrari}@unipv.it)

Abstract: In this paper we consider the identification of PieceWise Affine (PWA) models of Genetic Regulatory Networks (GRNs) and focus on data classification that is a task of the whole identification process. By assuming that gene expression profiles have been split into segments generated by a single affine mode, data classification amounts to group together segments that have been produced by the same mode. In particular, this operation must be performed in a noisy setting and without using any knowledge on the number of modes excited in the experiment. At a mathematical level, classification amounts to find all partitions of the set of segments that verify a statistical criterion and as such it has a combinatorial nature. In order to minimize the computational complexity we propose a pruning strategy for reducing the dimension of the search space. In particular, our approach hinges on a new algorithm for generating in an efficient way all partitions of a finite set that verify a bound on a monotone cost function.

Keywords: genetic regulatory networks; piecewise affine models; data classification; system identification; hybrid systems.

1. INTRODUCTION

Research on methods for inferring structure and parameters of Genetic Regulatory Networks (GRNs) from experimental data has gained momentum in recent years due to the availability of experimental techniques, such as RT-PCR and gene reporter systems (Ronen et al. [2002]), for measuring gene expression with a sampling time that is sufficiently small with respect to the time constants of the network.

Among the modeling formalisms for GRNs proposed in the literature, in this paper we consider PieceWise Affine (PWA) systems (Glass and Kauffman [1973]). The main reason for this choice is that, differently from linear models of GRNs, PWA models are capable to describe the strong nonlinear behavior of the network. Moreover, compared to smooth nonlinear models, they capture the switching nature of GRNs using a reduced number of parameters. While there are methods for the qualitative simulation of PWA models of GRNs (see e.g. Batt et al. [2005], Casey et al. [2005] and de Jong et al. [2004]), identification of these systems has received so far little attention.

In the hybrid systems literature, many different algorithms have been proposed for the identification of PWA input-output models (Paoletti et al. [2007]), and in principle they could be used for the data-based reconstruction of GRNs. However, PWA systems describing GRNs possess a specific structure that must be preserved in order to guarantee the biological interpretability of the identified model and all existing identification methods have a limited capability of incorporating such constraints.

In this paper we focus on a subtask of the identification process: the data classification problem. This problem consists in the attribution of the data to distinct modes of operation of the network. Classification represents the basis for another important identification task, i.e. the reconstruction of the thresholds characterizing the model and defining regulatory interactions among genes (see Drulhe et al. [2006]). Classification is performed assuming that gene expression profiles have been processed by switch detection algorithms such as those proposed by Porreca et al. [2007]. These algorithms produce segments of data generated by a single affine mode of the network. Therefore, data classification amounts to find which segments are generated by the same mode. This operation must be carried out in a statistical setting, because of noise affecting the data, and without any knowledge on the number of modes. From a mathematical viewpoint, the problem can be formulated in terms of finding all partitions of the set of segments that verify an upper bound on a monotone function. In order to minimize the computational complexity stemming from the combinatorial nature of the problem, we propose a pruning strategy inspired by the *Apriori* algorithm (Agrawal and Srikant [1994]). The resulting procedure aims at generating, in an efficient way, all aggregations of segments verifying a statistical criterion.

The paper is structured as follows. In Section 2 we introduce PWA models of GRNs, whose identification is addressed in Section 3. The data classification problem is

⋆ This work was supported by the European Commission under project HYGEIA (NEST-4995).

978-3-902661-00-5/08/$20.00 © 2008 IFAC

10.3182/20080706-5-KR-1001.2527

formally introduced in Section 4. A new algorithm for the efficient generation of partitions is described in Section 5 and specialized to segment aggregation in Section 6. The performance of the algorithm is analyzed in Section 7 by means of multiple experiments on synthetic data.

2. PWA MODELS OF GENETIC REGULATORY NETWORKS

We consider a GRN composed by n genes, each one coding for a molecule (e.g. a protein) whose concentration at time t is denoted with $x_i(t)$, for $i = 1, \dots, n$. Genes are represented as dynamical systems regulated by the concentration of molecules involved in the network. PWA models of GRNs have been introduced by Glass and Kauffman [1973] by approximating sigmoidal functions, commonly used for describing regulatory interactions, with step functions, hence modeling genes as switching units. Due to lack of space, in this section we summarize the main features of the resulting class of PWA models, deferring the reader to Glass and Kauffman [1973] and de Jong et al. [2004] for further details and examples.

The concentration vector $\boldsymbol{x} = [x_1, \dots, x_n]'$ represents the continuous state of the network and lies within a bounded hyperrectangle $\Omega \subseteq \mathbb{R}^n_+$ including the origin. To the i-th concentration variable it is associated a (possibly empty) set of positive thresholds $\{\theta_i^{\ell_i}\}_{\ell_i=1}^{p_i}$. All thresholds define the grid

$$G = \bigcup_{i\in\{1,\dots,n\},\ell_i\in\{1,\dots,p_i\}} \{\boldsymbol{x} \in \Omega : x_i = \theta_i^{\ell_i}\}$$

that splits Ω into open hyperrectangular regions Δ^j, $j = 1, \dots, s$, called *regulatory domains*. The dynamics of the GRN is then captured by the autonomous PWA system

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

where $\boldsymbol{\mu}^j = [\mu_1^j \cdots \mu_n^j]' \geq 0$, $\boldsymbol{\nu}^j = \mathrm{diag}(\nu_1^j, \dots, \nu_n^j) > 0$ are suitable coefficient matrices and $\lambda(\boldsymbol{x}) = j \Leftrightarrow \boldsymbol{x} \in \Delta^j$ is the *switching function*. Note that the r.h.s. of (1) is the difference of synthesis rates $\boldsymbol{\mu}^j$ and degradation rates $\boldsymbol{\nu}^j\boldsymbol{x}$. In particular, spontaneous degradation is always present, and the i-th gene is off when $\mu_i^j = 0$. Each tuple $(\boldsymbol{\mu}^j, \boldsymbol{\nu}^j, \Delta^j)$ defines a *mode of operation* of the network. Note that the dynamics (1) can be defined also on G by using the notion of Filippov solutions (de Jong et al. [2004]).

We introduce now *molecule domains* as the regions where a single molecule concentration evolves according to a single affine dynamics. We associate to the i-th molecule the set

$$R_i = \left\{(\mu_i^j, \nu_i^j), j = 1, \dots, s\right\} \tag{2}$$

collecting all the distinct pairs of rate coefficients and having cardinality s_i. Molecule domains M_i^q, $q = 1, \dots, s_i$, are defined as

$$M_i^q = \bigcup_{j=1}^{s} \left(\Delta^j : (\mu_i^j, \nu_i^j) = R_{iq}\right) \ , \tag{3}$$

where $R_{iq} = (\kappa_i^q, \gamma_i^q)$ is the q-th pair in R_i. Apparently, $\{M_i^q\}_{q=1}^{s_i}$ is a partition of $\Omega \setminus G$. The dynamics of x_i is then given by the PWA system

$$\dot{x}_i = \kappa_i^q - \gamma_i^q x_i \ , \quad \text{if } \lambda_i(\boldsymbol{x}) = q \ , \tag{4}$$

where $\lambda_i(\boldsymbol{x}) = q \Leftrightarrow \boldsymbol{x} \in M_i^q$ is the *molecular switching function*. In (4) the variables x_ℓ, with $\ell \neq i$, play the role of inputs affecting the selection of the active dynamics. The tuples $(\kappa_i^q, \gamma_i^q, M_i^q)$, $q = 1, \dots, s_i$, will be called *molecular modes of operation* (of the i-th molecule).

Experimental data that can be obtained with gene reporter systems are measurements of molecular concentrations collected at sampling instants t_k, $k \in \mathbb{N}$, sufficiently close to each other with respect to the time constants of the network dynamics. We assume that measurements $y_i(t_k)$ are generated by the output-error model

$$y_i(t_k) = x_i(t_k) + \xi_i(k), \quad \xi_i(\cdot) \sim WGN(0, \sigma_i^2) \ , \tag{5}$$

where $\xi_i(\cdot)$ is a white gaussian noise with zero mean and variance σ_i^2.

3. IDENTIFICATION OF PWA MODELS OF GENETIC REGULATORY NETWORKS

The data-based reconstruction of model (4)–(5) could be thought of as a classic hybrid identification problem, for which identification methods are available in the literature. In particular, several algorithms have been proposed for identifying PieceWise AutoRegressive eXogenous (PWARX) and PWA Output-Error (PWA-OE) models (see Paoletti et al. [2007] and the references therein). Techniques for the identification of PWA-OE models assume that the number of modes of operation composing the system is known in advance, that is seldom the case in the context of GRNs. Moreover, as pointed out in Drulhe et al. [2006], all existing procedures for the identification of hybrid models are general-purpose and hence do not account for features and constraints specific to GRNs models. This fact has some important consequences. First, neglecting the constraints on the model structure can result in models that have no biological meaning. Second, existing hybrid identification techniques typically produce a single model, while the scarcity of expression data often does not allow one to uniquely reconstruct the switching mechanisms characterizing the GRN. Therefore it makes sense to generate multiple results in order to provide biologists with multiple and plausible hypotheses on the network functioning.

In view of the previous remarks, we are developing a gray-box procedure for the identification of PWA models of GRNs that is conceptually split in the following tasks:

(1) detection of the switches in time series of gene expression data;
(2) attribution of the data to distinct modes of operation of the whole GRN (classification problem);
(3) reconstruction of thresholds on concentration variables and of all combinations of thresholds consistent with the data;
(4) estimation of the kinetic parameters in each mode of operation for all models generated in point 3.

In the sequel, we focus on the classification problem (task 2), assuming task 1 has been carried out. Algorithms for detecting switches in gene expression profiles can be found in Porreca et al. [2007]. Task 3 can be performed, under suitable assumptions, using the multicut algorithm proposed by Drulhe et al. [2006]. As pointed out in Paoletti et al. [2007], task 4 can be easily carried out relying on the data classification produced in step 2.

4. CLASSIFICATION PROBLEM

In this section we discuss the classification problem and describe our approach for solving it. As mentioned above, classification is based on the results produced by switch detection algorithms that split time series of molecular concentrations into segments of consecutive data generated by a single affine mode. In particular, we focus on the concentration profile of a single molecule and propose an algorithm for detecting and aggregating segments generated by the same molecular mode of operation. Classification with respect to modes of operation of the whole network is then easily obtained by merging the aggregation results obtained for all molecules, as described in Porreca and Ferrari-Trecate [2007].

For simplifying the notation, we will not specify the considered molecule and use the index i with a different meaning with respect to Section 2. First of all, we assume that a set of m segments

$$\mathcal{S} = \left\{S_i = \{(t_{i1}, y_{i1}), \dots, (t_{iN_i}, y_{iN_i})\}\right\}_{i=1}^{m} \qquad (6)$$

is available for the molecule. Each segment S_i is a collection of N_i consecutive data points (t_{ij}, y_{ij}), where $y_{ij} = y(t_{ij})$ is the measurement obtained at the sampling instant t_{ij}. The total number of data is $N = \sum_{i=1}^{m} N_i$. We also assume that segments are disjoint, i.e. $S_i \cap S_j = \emptyset$ for $i \neq j$, and that $N_i > 3$ [1]. Since data points belonging to the segment S_i have been attributed to the same dynamics, they can be described by the exponential model

$$y_{ij} = \phi(\kappa_i, \gamma_i, x_{0i}, t_{ij} - t_{i1}) + \xi(j), \xi(\cdot) \sim WGN(0, \sigma^2), \quad (7)$$

$$\text{with } \phi(\kappa, \gamma, x_0, \Delta t) = \frac{\kappa}{\gamma} - \left(\frac{\kappa}{\gamma} - x_0\right) e^{-\gamma \Delta t},$$

that is obtained by exact integration of the affine dynamics (4). In (7), $\kappa_i \geq 0$ and $\gamma_i > 0$ are the true and unknown rate parameters and x_{0i} is the (true and unknown) concentration at the beginning of the i-th segment. The problem of aggregating segments generated by the same affine mode amounts to find *equivalence relations* $\sim$ on $\mathcal{S}$ such that $S_i \sim S_j$ implies $(\kappa_i, \gamma_i) = (\kappa_j, \gamma_j)$. Note that equivalence classes form a *partition* of the set $\mathcal{S}$. Moreover, there is a bijection between the set of all possible equivalence relations on $\mathcal{S}$ and the set of all possible partitions of $\mathcal{S}$ (Cameron [1994]). This allows to state our problem in terms of looking for partitions of $\mathcal{S}$ corresponding to equivalence relations $\sim$ consistent with the data.

Recalling the statistical setting we have assumed by modeling measurements as in (5), the problem of aggregating segments is characterized by the following features:

- every possible aggregation is a partition of the set of segments $\mathcal{S}$;
- due to the noise affecting the data, aggregation must be based on a statistical criterion;
- the solution might be not unique, since several partitions can be statistically consistent with the data.

In principle, the problem could be solved by considering all possible partitions of $\mathcal{S}$ and discarding the ones that do not fulfill the statistical criterion. The main drawback of this approach is the combinatorial complexity of exhaustive search. Therefore we aim at finding an efficient strategy for generating all partitions that are statistically consistent with the data. To this purpose we first introduce, in the next section, an abstract algorithm for generating partitions verifying a bound on a monotone function. Then, in Section 6 we will show how to exploit it for performing classification.

Rank	Partition	RGF	Rank	Partition	RGF
1	1 2 3 4	1,1,1,1	9	1 4/2 3	1,2,2,1
2	1 2 3/4	1,1,1,2	10	1/2 3 4	1,2,2,2
3	1 2 4/3	1,1,2,1	11	1/2 3/4	1,2,2,3
4	1 2/3 4	1,1,2,2	12	1 4/2/3	1,2,3,1
5	1 2/3/4	1,1,2,3	13	1/2 4/3	1,2,3,2
6	1 3 4/2	1,2,1,1	14	1/2/3 4	1,2,3,3
7	1 3/2 4	1,2,1,2	15	1/2/3/4	1,2,3,4
8	1 3/2/4	1,2,1,3			

Table 1. Partitions of the set $\{1, 2, 3, 4\}$ and corresponding RGFs; ranking is with respect to lexicographic order of RGFs.

5. GENERATION OF PARTITIONS VERIFYING A BOUND ON A MONOTONE FUNCTION

A partition $P = \{X_1, X_2, \dots, X_k\}$ of the finite set $X = \{1, 2, \dots, m\}$ is any set of nonempty and mutually disjoint subsets of X, called *blocks*, such that their union is equal to X. As an example, the 15 partitions of $\{1, 2, 3, 4\}$ are reported in Table 1 where, in order to simplify the notation, a slash character is used to separate blocks. Let $\mathcal{P}$ be the set of all possible partitions of X and $\mathcal{P}_k \subseteq \mathcal{P}$ be the set of partitions having k blocks. As an example, with reference to Table 1, one has $\mathcal{P}_3 = \{1\,2/3/4, 1\,3/2/4, 1/2\,3/4, 1\,4/2/3, 1/2\,4/3, 1/2/3\,4\}$. Note that $\{\mathcal{P}_k\}_{k=1}^{m}$ is a partition of $\mathcal{P}$. It is possible to consider a partial order relation $\leq$ on $\mathcal{P}$, corresponding to the concept of "being finer than" (Stanley [1997]).

Definition 1. Given two partitions $P, Q \in \mathcal{P}$, $P \leq Q$ (P is finer than Q, Q is coarser than P) if for each block $X_i \in P$ there exists a block $X_j \in Q$ such that $X_i \subseteq X_j$.

Moreover, denote with $\prec$ and $\succ$ the usual covering relations associated with $\leq$. In particular, $Q \succ P$ (Q covers P) is obtained by replacing exactly two blocks of P by their union. It is easy to verify that the partially ordered set [2] $(\mathcal{P}, \leq)$ is a complete lattice.

Set partitions are often represented using a sequence known as restricted growth function (RGF) (Knuth [2005]), i.e. a string $\boldsymbol{p} = (p_1, p_2, \dots, p_m)$, with $p_i \in \mathbb{Z}$, satisfying the restricted growth condition

$$p_i \leq 1 + \max\{p_1, \dots, p_{i-1}\}, \forall i > 1 \ . \qquad (8)$$

Without loss of generality, we assume $p_1 = 1$. The main idea of using RGFs is that, for any $i, j \in \{1, \dots, m\}$, $p_i = p_j$ means that i and j belong to the same block of the partition. This yields a one-to-one correspondence between partitions of $\{1, 2, \dots, m\}$ and RGFs of length m (see Table 1 for an example). For the sake of clarity, the same letter will be used to denote a partition (capital letter, e.g. P) and the corresponding RGF (bold small letter, e.g. $\boldsymbol{p}$). The set of all RGFs (that corresponds to $\mathcal{P}$) will be denoted with $\mathcal{L}$. Note that the number of blocks, i.e. the

[1] This assumption is needed in order to estimate three parameters for each segment.

[2] In the sequel, partially ordered sets will be termed "posets".

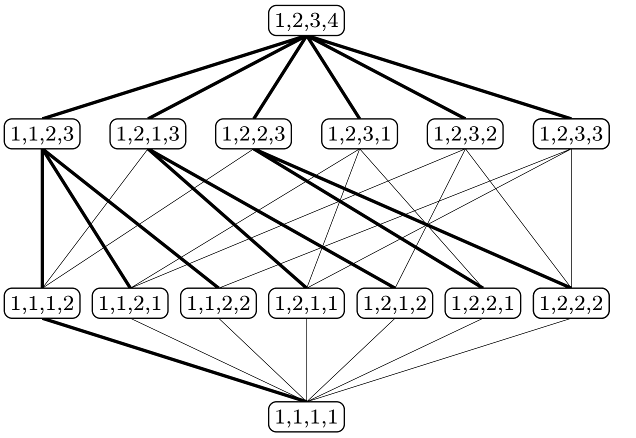

Fig. 1. Graphical representation, using RGFs, of the poset $(\mathcal{P}, \leq)$ for $m = 4$. Partitions in $\mathcal{P}_k$ appear (in lexicographic order) on the k-th level from the bottom, and edges represent covering relations. Thick edges represent the generation tree used in Algorithm 1.

cardinality $|P|$ of the partition P, is given by $\max \boldsymbol{p}$. RGFs can be sorted according to the lexicographic ascending and descending total orders. RGFs in Table 1 appear in ascending lexicographic order. Figure 1 represents the poset $(\mathcal{P}, \leq)$ for $m = 4$ using the RGF notation.

Definition 2. For any $\boldsymbol{p} \in \mathcal{L}$, $C(\boldsymbol{p})$ is a full-rank $(m-k) \times m$ *constraint matrix* such that:

- $k = \max \boldsymbol{p}$;
- the elements of $C(\boldsymbol{p})$ belong to $\{-1, 0, +1\}$;
- in each row there are exactly two elements whose values are $+1$ and -1; if i and j are the indexes of such elements, then one has $p_i = p_j$.

As an example, for $\boldsymbol{p} = (1, 1, 2, 1, 3, 2)$, a possible constraint matrix is

$$C(\boldsymbol{p}) = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \end{pmatrix} .$$

Note that the constraint matrix for a given partition is not unique. Based on Definition 2, it is possible to provide an algebraic characterization of the partial order relation. Indeed, considering RGFs as column vectors one has

$$P \leq Q \Leftrightarrow C(\boldsymbol{p}) \cdot \boldsymbol{q} = \boldsymbol{0} \ . \tag{9}$$

Let $J : \mathcal{P} \to \mathbb{R}$ be a monotone function, i.e. $\forall P, Q \in \mathcal{P}, P \leq Q \Rightarrow J(P) \leq J(Q)$. We consider the problem of generating every partition P satisfying the condition

$$J(P) < \bar{J} \ , \tag{10}$$

where $\bar{J} \in \mathbb{R}$ is a given upper bound. Define as *accepted* the partitions satisfying (10), and as *rejected* those that are not accepted. Therefore we aim at generating the set $\mathcal{P}^{\mathrm{a}}$ of all the accepted partitions:

$$\mathcal{P}^{\mathrm{a}} = \left\{ P \in \mathcal{P} : J(P) < \bar{J} \right\} \ . \tag{11}$$

In principle, one could generate all the partitions in $\mathcal{P}$ and test if they are accepted or not. However, it is possible to exploit the monotonicity property of function J in order to avoid the complete generation hence reducing the search space. The next proposition, that is a direct consequence of the monotonicity of J, will play a key role in our generation procedure.

Algorithm 1. *Apriori*-like strategy for generating $\mathcal{P}^{\mathrm{a}}$

```
$k = m$
$\mathcal{P}^{\mathrm{a}}_m = \left\{ P \in \mathcal{P}_m = \{1/2/\ldots/m\} : J(P) < \bar{J} \right\}$
while ($\mathcal{P}^{\mathrm{a}}_k \neq \emptyset$) and ($k > 1$) do
    $k = k - 1$
    $\mathcal{P}^{\mathrm{c}}_k = \mathrm{candidateGeneration}(\mathcal{P}^{\mathrm{a}}_{k+1})$
    $\mathcal{P}^{\mathrm{a}}_k = \left\{ P \in \mathcal{P}^{\mathrm{c}}_k : J(P) < \bar{J} \right\}$
end while
```

Proposition 3. $P \in \mathcal{P}^{\mathrm{a}} \Rightarrow Q \in \mathcal{P}^{\mathrm{a}}, \forall Q \leq P$

The efficient construction of $\mathcal{P}^{\mathrm{a}}$ based on Proposition 3 is inspired by *Apriori* (Agrawal and Srikant [1994]), a classic data mining algorithm for learning association rules (Tan et al. [2005]). This algorithm exploits a particular property, called *Apriori* principle, for efficiently generating subsets. Proposition 3 actually corresponds to an *Apriori*-like principle for partitions: if a partition is accepted then all finer partitions must be accepted. As for the *Apriori* algorithm, this principle allows one to reduce the search space by pruning partitions that are coarser than a rejected partition. Therefore, the *Apriori*-like algorithm for constructing $\mathcal{P}^{\mathrm{a}}$ is based on a "fine-to-coarse" search strategy and considers at each step, for testing condition (10), only partitions in the candidate set

$$\mathcal{P}^{\mathrm{c}}_k = \left\{ Q \in \mathcal{P}_k : \forall P \prec Q, P \in \mathcal{P}^{\mathrm{a}}_{k+1} \right\} \ , \tag{12}$$

constructed starting from $\mathcal{P}^{\mathrm{a}}_{k+1} = \mathcal{P}^{\mathrm{a}} \cap \mathcal{P}_{k+1}$. Proposition 3 ensures that $\mathcal{P}^{\mathrm{a}}_k \subseteq \mathcal{P}^{\mathrm{c}}_k$. The resulting iterative procedure is shown in Algorithm 1.

The core of the algorithm is the function candidateGeneration, which needs to be expressly designed for dealing with partitions. A detailed description of the procedure for constructing $\mathcal{P}^{\mathrm{c}}_k$ from $\mathcal{P}^{\mathrm{a}}_{k+1}$ is out of the scope of this paper and can be found in Porreca and Ferrari-Trecate [2007]. The idea is to define a partial order $\leq_{\mathrm{T}}$ such that $(\mathcal{P}, \leq_{\mathrm{T}})$ is a weak subposet of $(\mathcal{P}, \leq)$ and has a tree structure, as the one represented in Fig. 1 by thick edges. Then, the two main steps for generating candidates are:

(1) generation step: build the superset $\widetilde{\mathcal{P}}^{\mathrm{c}}_k$ of $\mathcal{P}^{\mathrm{c}}_k$ given by
$$\widetilde{\mathcal{P}}^{\mathrm{c}}_k = \{ Q \in \mathcal{P}_k : \exists P \in \mathcal{P}^{\mathrm{a}}_{k+1} \text{ verifying } P \leq_{\mathrm{T}} Q \} \ ;$$
(2) pruning step (like in *Apriori* candidate generation): remove from $\widetilde{\mathcal{P}}^{\mathrm{c}}_k$ partitions that cover some $P \notin \mathcal{P}^{\mathrm{a}}_{k+1}$.

We highlight that, as detailed in Porreca and Ferrari-Trecate [2007], an efficient procedure for building $\widetilde{\mathcal{P}}^{\mathrm{c}}_k$ is based on the generation of RGFs in lexicographic order.

6. SEGMENT AGGREGATION

As discussed in Section 4, the problem of aggregating segments amounts to find partitions of $\mathcal{S}$ that are statistically consistent with the data. In a statistical setting, each partition can be considered as the hypothesis that $S_i \sim S_j$ for all S_i, S_j belonging to the same block, i.e. that equality constraints between rate parameters hold. Recalling the meaning of representing a partition P with the RGF $\boldsymbol{p}$, one has

$$p_i = p_j \Leftrightarrow S_i \sim S_j, \ S_i \sim S_j \Rightarrow (\kappa_i, \gamma_i) = (\kappa_j, \gamma_j) \ . \tag{13}$$

Let $\Theta = (\kappa_1, \gamma_1, x_{01}, \ldots, \kappa_m, \gamma_m, x_{0m})'$ be the vector collecting parameters of model (7) that characterize the segments $S_1, \ldots, S_m$. Then, the constraints on rate parameters associated with a partition P can be expressed as

$$(C(\boldsymbol{p}) \otimes I_{2,3}) \cdot \Theta = \mathbf{0} \ , \quad \text{with } I_{2,3} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \ , \qquad (14)$$

where symbol $\otimes$ denotes the Kronecker product. For each partition P, least squares estimation of Θ can be performed under the hypothesis that constraint (14) holds, thus obtaining the estimate $\hat{\Theta}(P)$ as

$$\hat{\Theta}(P) = \arg\min_{\Theta} J(\Theta) \qquad (15)$$
$$\text{subject to } \ (C(\boldsymbol{p}) \otimes I_{2,3}) \cdot \Theta = \mathbf{0}$$

where

$$J(\Theta) = \sum_{i=1}^{m} \sum_{j=1}^{N_i} \left[(y_{ij} - \phi(\kappa_i, \gamma_i, x_{0i}, t_{ij} - t_{i1}) \right]^2 \ . \qquad (16)$$

In order to asses if P is consistent with the data one can use a statistical test for the hypothesis $\mathrm{H}_0 : (C(\boldsymbol{p}) \otimes I_{2,3}) \cdot \Theta = \mathbf{0}$ against the hypothesis H_1 that no constraints hold. Note that H_1 corresponds to the partition $P_1 = 1/2/\ldots/m$. Let $\hat{J}(P) = J(\hat{\Theta}(P))$ be the Sum of Squared Residuals (SSR) under H_0 and $\hat{J}_1 = J(\hat{\Theta}(P_1))$ the SSR under H_1. According to Rohatgi and Saleh [2000], the Generalized Likelihood Ratio (GLR) $(1-\alpha)$-level test [3] is to reject H_0, and the corresponding partition, if

$$\hat{J}(P) \geq (1 + \Delta_\alpha(k))\hat{J}_1 \ , \qquad (17)$$

$$\text{with } \Delta_\alpha(k) = \frac{2(m-k)}{N-3m} F_\alpha(2(m-k), N-3m) \ . \qquad (18)$$

In (18), $F_\alpha(v_1, v_2)$ represents the $(1-\alpha)$-th quantile of the F distribution with (v_1, v_2) degrees of freedom, and k is the number of blocks of P. Therefore we aim at generating the set of partitions

$$\widetilde{\mathcal{P}^{\mathrm{a}}} = \left\{ P \in \mathcal{P} : \hat{J}(P) < (1 + \Delta_\alpha(|P|))\hat{J}_1 \right\} \ . \qquad (19)$$

Moreover, among the partitions in $\widetilde{\mathcal{P}^{\mathrm{a}}}$, we are particularly interested in those that are *maximal*, i.e. partitions $P \in \widetilde{\mathcal{P}^{\mathrm{a}}}$ such that $\forall Q > P, Q \notin \widetilde{\mathcal{P}^{\mathrm{a}}}$. In fact, maximal partitions are the most informative ones, since they fulfill the constraints imposed by any finer partition.

Proposition 4. (Monotonicity of $\hat{J}$).

$$Q \geq P \ \Rightarrow \ \hat{J}(Q) \geq \hat{J}(P)$$

Proof. By (9), one has that the constraint $C(\boldsymbol{p}) \cdot \boldsymbol{q} = \mathbf{0}$ on the RGF $\boldsymbol{q}$ holds for all partitions $Q \geq P$. According to (13), this implies that the constraint on the parameters $(C(\boldsymbol{p}) \otimes I_{2,3}) \cdot \Theta = \mathbf{0}$ holds for Q. Under this constraint $J(\Theta)$ is minimized by $\hat{\Theta}(P)$, thus implying $\hat{J}(Q) \geq \hat{J}(P)$ for any $Q \geq P$.

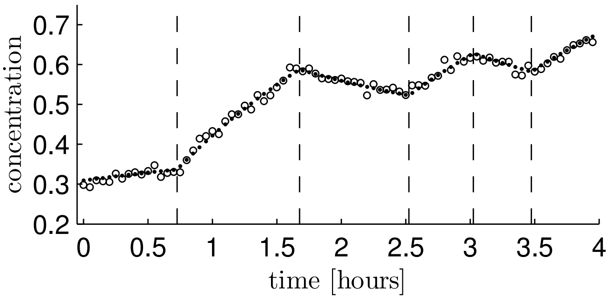

Fig. 2. Time series used for validation; dots and circle represent the noiseless and noisy data ($\sigma = 0.01$); vertical lines denote switching times defining segments.

Proposition 4 is the basis for applying the algorithm described in Section 5. However, the rejection condition (17) is based on a bound that depends on the number of blocks, since $\Delta_\alpha(k)$ increases as k decreases. Therefore it is not true, in general, that if P is rejected by test (17) then each $Q \geq P$ will be rejected. This prevents the application of Algorithm 1 for generating $\widetilde{\mathcal{P}^{\mathrm{a}}}$. The idea to overcome this problem is to use as bound the largest values of the r.h.s. of (17), which is obtained for $k = 1$, thus considering the set

$$\mathcal{P}^{\mathrm{a}} = \left\{ P \in \mathcal{P} : \hat{J}(P) < (1 + \Delta_\alpha(1))\hat{J}_1 \right\} \ . \qquad (20)$$

Partitions in $\mathcal{P}^{\mathrm{a}}$ can be actually generated by Algorithm 1. Apparently, $\mathcal{P}^{\mathrm{a}}$ is a superset of $\widetilde{\mathcal{P}^{\mathrm{a}}}$, and typically it is not much larger than $\widetilde{\mathcal{P}^{\mathrm{a}}}$. Therefore, once $\mathcal{P}^{\mathrm{a}}$ has been generated, $\widetilde{\mathcal{P}^{\mathrm{a}}}$ is easily obtained by removing partitions rejected by the test (17). Note that such a test is not valid for P_1, which however has always to be considered as accepted, thus implying $\mathcal{P}^{\mathrm{a}}_m = \mathcal{P}_m$. The algorithm can produce multiple maximal partitions, each one corresponding to an aggregation of segments consistent with the data. This agrees with the goal in GRN identification of generating all the models that are consistent with the data.

To get an idea of the search space reduction, consider the ideal case where test (17) only accepts the "true" partition P^* with RGF $\boldsymbol{p}^* = (1, 2, 1, 2, \ldots)$ and all its refinements. Then, the number of partitions explored and tested by Algorithm 1 is 8 over a total of 15 for $m = 4$, 15 over 52 for $m = 5$, 34 over 203 for $m = 6$.

7. EXPERIMENTAL RESULTS

The effectiveness of the proposed method was tested by running multiple experiments on synthetic data. We considered the noiseless time series of normalized concentration values depicted in Fig. 2. In particular, the six segments visible in Fig. 2 arise from switches between two molecular modes of operation, with rate parameters (κ, γ) equal to $(0.2, 0.5)$ and $(0.5, 0.5)$ for the first and second mode, respectively. Noisy data points in each mode were generated according to model (7). The time series mimic behaviors and features of data obtained from PWA models of real GRNs, such as those used in Drulhe et al. [2006] and Porreca et al. [2007]. We added to the measurements noise with standard deviations $\sigma \in \{0.001, 0.005, 0.01\}$, i.e. spanning one decade around the expected noise level of data produced by gene reporter systems (Drulhe et al. [2006]). Additional simulations for lower values of σ displayed performances very similar to the case $\sigma = 0.001$, whereas for $\sigma > 0.01$ the noise level was so high that almost no dynamics was visible in the data. A set of 200 noisy trajectories was generated for each value of σ. For each trajectory the algorithms described in Sections 5–6 were applied using the correct segmentation, i.e. assuming

[3] The considered GLR test is designed in Rohatgi and Saleh [2000] for linear models. When applied to nonlinear models, the level of the test is $(1-\alpha)$ under some approximation (Gallant [1986]).

		$N_{\max}$					
σ	$p_{\rm s}$	1	2	3	4	5	$N_{\rm e}$
0.001	98.0%	98.5%	0.0%	0.0%	1.0%	0.5%	33.92
0.005	98.5%	99.0%	0.5%	0.0%	0.0%	0.5%	34.01
0.01	99.5%	69.0%	25.0%	5.0%	1.0%	0.0%	36.57

Table 2. Performance indexes for different noise levels.

the true switching instants to be known. For test (17) the level of 0.95 ($\alpha = 0.05$) was used.

The optimal situation is when the "true" partition $P^* = 1\,3\,5/2\,4\,6$ is the only maximal partition in $\widetilde{\mathcal{P}^{\rm a}}$. Moreover, it is important to evaluate the reduction of the search space. Therefore we consider the following performance indexes:

- the percentage $p_{\rm s}$ of successful cases in which P^* appears among the maximal partitions;
- the distribution of the number $N_{\max}$ of maximal partitions generated in each experiment;
- the average number $N_{\rm e}$ of explored partitions.

Performances obtained for the considered noise levels are reported in Table 2. First of all, the effectiveness in reducing the search space is demonstrated by values of $N_{\rm e}$ that are less than 20% of the total number of partitions (203). Results also show an excellent capability of including P^* among the retained partitions, with values of $p_{\rm s}$ bigger than 98%. Moreover, for $\sigma = 0.001$ and $\sigma = 0.005$ at most 3 among 200 experiments produced multiple maximal partitions. The situation is different for $\sigma = 0.01$. Although the value of $p_{\rm s}$ is even higher than the ones obtained for lower noise levels, multiple maximal partitions are generated in the 31% of the experiments, producing only one spurious partition in the majority of such cases. As a final remark, the high values of $p_{\rm s}$ with respect to $(1 - \alpha)$ suggests that the level of test (17) is underestimated due to the model nonlinearity.

8. CONCLUSION

In this paper we have considered the problem of classifying gene expression data, that is a crucial step in the process of identifying PWA models of GRNs. We have assumed that, for each molecule involved in the network, segments of data generated by a single affine mode are available. Then, data classification aims at generating partitions of the set of segments whose blocks correspond to data generated by the same mode of operation. Given the combinatorial nature of the problem, we have proposed a statistical method for generating partitions consistent with the data that ultimately relies on an algorithm for the efficient generation of partitions verifying a bound on a monotone function. Experimental results, obtained assuming the knowledge of the true segments, have shown the effectiveness of our method in producing the correct partition and in reducing the dimension of the search space. Future research will consider the effect of coupling the proposed method with switch detection algorithms that reconstruct segments in concentration time series. In particular, it will be important to evaluate the joint performance of the algorithms, considering how possible errors in detecting switches propagate to the results of classification. The final goal of this research is to integrate these procedures with the algorithm proposed in Drulhe et al. [2006], in order to have a complete procedure for the data-based identification of PWA models of GRNs.

REFERENCES

R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In J.B. Bocca, M. Jarke, and C. Zaniolo, editors, *Proc. 20th Int. Conf. Very Large Data Bases (VLDB)*, pages 487–499. Morgan Kaufmann, 1994.

G. Batt, D. Ropers, H. de Jong, J. Geiselmann, M. Page, and D. Schneider. Qualitative analysis and verification of hybrid models of genetic regulatory networks: nutritional stress response in *Escherichia coli*. In M. Morari and L. Thiele, editors, *Proc. Hybrid Systems: Computation and Control (HSCC 2005)*, volume 3414 of *LNCS*, pages 134–150. Springer-Verlag, 2005.

P.J. Cameron. *Combinatorics: Topics, Techniques, Algorithms*. Cambridge University Press, 1994.

R. Casey, H. de Jong, and J.-L. Gouzé. Piecewise-linear models of genetic regulatory networks: equilibria and their stability. *J. Math. Biol.*, 52(1):27–56, 2005.

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

S. Drulhe, G. Ferrari-Trecate, H. de Jong, and A. Viari. Reconstruction of switching thresholds in piecewise-affine models of genetic regulatory networks. In J. Hespanha and A. Tiwari, editors, *Proc. Hybrid Systems: Computation and Control (HSCC 2006)*, volume 3927 of *LNCS*, pages 184–199. Sringer-Verlag, 2006.

A.R. Gallant. *Nonlinear Statistical Models*. Wiley, 1986.

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

D.E. Knuth. *The Art of Computer Programming, Volume 4, Fascicle 3: Generating All Combinations and Partitions*. Addison-Wesley, 2005.

S. Paoletti, A.L. Juloski, G. Ferrari-Trecate, and R. Vidal. Identification of hybrid systems: a tutorial. *Eur. J. Contr.*, 13(2-3):242–260, 2007.

R. Porreca, G. Ferrari-Trecate, D. Chieppi, L. Magni, and O. Bernard. Switch detection in genetic regulatory networks. In A. Bemporad, A. Bicchi, and G. Buttazzo, editors, *Proc. Hybrid Systems: Computation and Control (HSCC 2007)*, volume 4416 of *LNCS*, pages 754–757. Sringer-Verlag, 2007.

R. Porreca and G. Ferrari-Trecate. Identification of piecewise affine models of genetic regulatory networks: the data classification problem. Technical Report 143/07, Università degli Studi di Pavia, 2007. `http://sisdin.unipv.it/lab/personale/pers_hp/ferrari/publications.html`.

V.K. Rohatgi and A.K.M. Saleh. *An Introduction to Probability and Statistics*. Wiley, 2nd edition, 2000.

M. Ronen, R. Rosenberg, B.I. Shraiman, and U. Alon. Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc. Natl. Acad. Sci. USA*, 99(16):10555–10560, 2002.

R.P. Stanley. *Enumerative Combinatorics, Volume 1*. Cambridge University Press, 1997.

P.-N. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison-Wesley, 2005.