Switch Detection in Genetic Regulatory Networks

Riccardo Porreca[1], Giancarlo Ferrari-Trecate[1], Daniela Chieppi[1], Lalo Magni[1], and Olivier Bernard[2]

[1] Dipartimento di Informatica e Sistemistica, Università degli Studi di Pavia, via Ferrata 1, 27100 Pavia, Italy
e-mail: {giancarlo.ferrari, daniela.chieppi, lalo.magni}@unipv.it
[2] INRIA, Sophia Antipolis, 2004 route des Lucioles, B.P. 93, 06902 Sophia Antipolis, France
e-mail: olivier.bernard@inria.fr

Technical Report 142/06

Abstract. This paper considers piecewise affine models of genetic regulatory networks and focuses on the problem of detecting switches among different modes of operation in gene expression data. This task constitutes the first step of a procedure for the complete identification of the network and complements the algorithms proposed in [1]. We propose two methods for switch detection. The first one is based on the computation of suitable indexes that emphasize the occurrence of switches in the data. The second one exploits nonlinear identification techniques in order to recast switch detection into an hypothesis testing problem. In both cases we assume that the expression of individual genes obeys to an output-error piecewise affine dynamics and we study the performance of the proposed algorithms for different noise levels. We also illustrate the application of our methods to the reconstruction of switching times in data produced by a piecewise affine model of the network regulating the carbon starvation response in *Escherichia coli.*

1 Introduction

The reconstruction of biochemical networks from experimental data is nowadays recognized as one of the most important goals of Systems Biology. Research in this field has been promoted by the availability of experimental techniques for measuring the concentration of various molecules regulating the functioning of cells. As far as Genetic Regulatory Networks (GRN) are concerned, several measurements techniques have been developed for sampling gene expression, ranging from DNA microarrays [2] to RT-PCR and gene reporter systems [3]. These methods differ under many aspects including the number of genes that can be measured simultaneously and the allowed sampling time, thus highlighting different features of the network at different timescales.

We consider the problem of identifying the dynamics of GRNs using gene expression data collected with a sampling time that is sufficiently short with respect to the time constants of the network. As an example, data produced by gene reporter systems possess this feature and can capture transitory phenomena [3]. Moreover, we restrict our attention to PieceWise Affine (PWA) models of GRNs [4,5] since they are attractive under many respects. First, they preserve the nonlinear character of the underlying biological process thus capturing behaviors that cannot be represented by means of linear models. Second, they usually involve a reduced number of parameters with respect to general nonlinear models of GRNs, a feature that is appealing from the identification viewpoint. Third, powerful techniques exist for analysis and qualitative simulation of PWA models of GRNs [6,7].

Recently, many different algorithms have been proposed for the identification of PWA input-output models [8,9,10,11], 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 basic task in the whole identification procedure: the detection of switches in data generated by PWA input-output models of GRNs. In particular, our aim is to find switches between different modes of operation without assuming any knowledge of the model parameters. For general piecewise affine autoregressive exogenous systems, this problem has been addressed in [12], where the links with fault detection techniques are also discussed. However, our methods differ from those proposed in [12] under two respects: first, we focus on Output-Error (OE) models of GRNs and second, we exploit the structure of these models for improving the switch detection capabilities of our algorithms.

Two different techniques for switch detection are proposed. The first one, described in Section 5, is based on the construction of switching indexes that have a constant value in absence of switches and a non constant profile at switching times. We provide a statistical characterization of these indexes and compute their confidence sets that are used in an iterative algorithm for aggregating consecutive data points generated by the same mode of operation. The second method, presented in Section 6, is based on the iterative identification of an OE nonlinear model that allows switches to be detected by means of hypothesis testing. In Section 7, we provide an extensive experimental comparison of the two algorithms and highlight the pros and cons of each method. In particular, our analysis reveals that the method based on switching indexes produces better results when the noise is sufficiently small. On the contrary, the algorithm based on nonlinear identification is preferable for higher noise levels. Finally, Section 8 presents the detection of switching times in gene expression data generated by a PW-OE model of the GRN governing the carbon starvation response in *Escherichia coli.*

2 PWA Models of Genetic Regulatory Networks

PWA models of GRNs have been introduced by Glass and Kauffmann [4] by approximating sigmoidal functions, commonly used for describing gene activation, with step functions, hence modeling genes as switching units that can be turned on and off. In this section we summarize the main features of the resulting class of PWA models, deferring the reader to [4], [5] and [1] for further details.

We assume that the regulation mechanism is described by the interactions of n genes, each one coding for a molecule (e.g. a protein). Molecule concentrations are denoted with x_i, for $i = 1, \dots, n$, and the concentration vector $\boldsymbol{x} = [x_1, \dots, x_n]$ lies within a bounded hyperrectangle $\varOmega \subseteq \mathbb{R}_+^n$ containing 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

$$\varTheta = \bigcup\nolimits_{i \in \{1,\dots,n\}, \ell_i \in \{1,\dots,p_i\}} \{\boldsymbol{x} \in \varOmega : x_i = \theta_i^{\ell_i}\}$$

that splits $\varOmega$ into open hyperrectangular regions $\varDelta^j$, $j = 1, \dots, \prod_{i=1}^n (p_i + 1)$, called regulatory domains. The dynamics of the GRN is then captured by the autonomous PWA system

$$\dot{\boldsymbol{x}} = \boldsymbol{\mu}^j - \boldsymbol{\nu}^j \boldsymbol{x} \ , \quad \text{if } \boldsymbol{x} \in D^j, \ \text{ with } j \in \{1, \dots, s\} \ , \tag{1}$$

where $\boldsymbol{\mu}^j = \mathrm{diag}\{\mu_1^j, \dots, \mu_n^j\} \geq 0$, $\boldsymbol{\nu}^j = \mathrm{diag}\{\nu_1^j, \dots, \nu_n^j\} > 0$ and the s regions D^j are disjoint unions of regulatory domains. Without loss of generality, we also assume that all pairs $(\boldsymbol{\mu}^j, \boldsymbol{\nu}^j)$ are different, meaning that if $i \neq j$ then $\boldsymbol{\mu}^i \neq \boldsymbol{\mu}^j$ or $\boldsymbol{\nu}^i \neq \boldsymbol{\nu}^j$. For subsequent use, we define the set $\mathcal{D} = \bigcup_{j=1}^s D^j = \varOmega \setminus \varTheta$. 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 $\mu_i^j = 0$.

In this paper we consider the problem of detecting switches in the dynamics of a single molecule. Therefore it is important to characterize regions in $\varOmega$ where the i-th molecule concentration obeys to the same dynamics. To this purpose, consider the set

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

collecting all distinct pairs of synthesis and degradation rates for the i-th molecule, and denote with s_i its cardinality. Molecule domains M_i^j, $j = 1, \dots, s_i$, are defined as:

$$M_i^j = \bigcup_{\ell=1}^{s} \left\{ D^\ell \mid (\mu_i^\ell, \nu_i^\ell) = \left(\kappa_i^j, \gamma_i^j\right) \right\} \ , \tag{3}$$

where $\left(\kappa_i^j, \gamma_i^j\right)$ is the j-th element of R_i. Apparently, for a fixed i, $\{M_i^j\}_{j=1}^{s_i}$ is a partition of $\mathcal{D}$. The dynamics of x_i is then given by the PWA system

$$\dot{x}_i = \kappa_i^j - \gamma_i^j x_i \ , \quad \text{if } \boldsymbol{x} \in M_i^j \ , \tag{4}$$

where the variables x_j, $\forall j \neq i$, play the role of inputs.

Experimental data are measurements y_i of the gene expression x_i, collected with a uniform sampling time $T > 0$. In order to account for the measurement noise, we introduce the PW Output Error (OE) model

$$\begin{cases} x_i(k+1) = \tilde{\kappa}_i^j - \tilde{\gamma}_i^j x_i \\ \quad\; y_i(k) = x_i(k) + n_i(k) \end{cases} , \quad \text{if } \boldsymbol{x}(k) \in M_i^j \ , \tag{5}$$

where $\tilde{\kappa}_i^j = \left(\kappa_i^j/\gamma_i^j\right)\left(1 - e^{-\gamma_i^j T}\right)$, $\tilde{\gamma}_i^j = -e^{-\gamma_i^j T}$, $n_i(k)$ is a white gaussian noise with zero mean and variance σ_n^2, and $k \in \mathbb{N}$ denotes the sampling instant $t = kT$. The modes of operation in (5) are the triples $\left(\tilde{\kappa}_i^j, \tilde{\gamma}_i^j, M_i^j\right)$.

Note that the ODE (1) has a discontinuous r.h.s. on Θ. When $\boldsymbol{x} \in \Theta$, solutions to (1) must be understood in the sense of Filippov [5], possibly giving rise to sliding modes on Θ. The results of this paper hinge on the assumption that sliding-mode behaviors are absent in the measured data. For the sake of simplicity we also assume that $\boldsymbol{x}(k) \notin \Omega$, $\forall k \in \mathbb{N}$, thus enforcing the well-posedness of (1) and (5).

3 Identification of PWA Models of Genetic Regulatory Networks

At a first sight, the data-based reconstruction of model (5) seems a classic hybrid identification problem, for which identification methods exist [8,9,10,11]. However, most of such algorithms assume gene expression data generated by an autoregressive exogenous model in each regulatory domain, and therefore they are not tailored to Output-Error models like (5). Identification of PW-OE models has been considered in [13] and [14]. However, in both papers, it is assumed 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 [1], all the above-mentioned procedures do not preserve the particular structure of PWA models of GRNs, and then could generate models that are meaningless from a biological viewpoint. Finally, existing hybrid identification techniques produce a single model while scarcity of expression data does not allow one to uniquely reconstruct the switching mechanisms characterizing the GRNs.

The data-based reconstruction of GRNs can be conceptually split in the following tasks:

1. detection of the switches in single time series of gene expression data (without assuming any knowledge of the mode parameters and regions);
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.

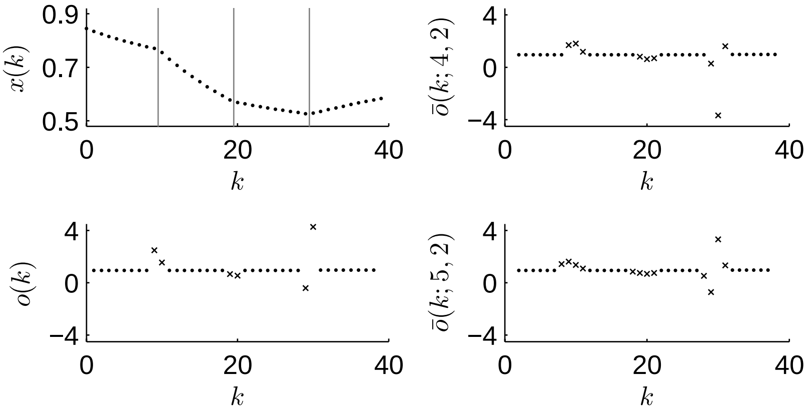

Fig. 1. Behavior of switching indexes $o(k)$, $\bar{o}(k;4,2)$ and $\bar{o}(k;5,2)$. The samples $x(k)$ are depicted in the upper left panel together with switching instants, represented by vertical lines.

In the sequel, we propose two algorithms for solving task 1. *Ad hoc* methods for task 2 are currently under study, while task 3 can be performed, under suitable assumptions, using the multicut algorithm proposed in [1]. As pointed out in [8], task 4 can be easily carried out relying on the data classification produced in step 2.

4 Switching Indexes

This section is devoted to the study of indexes that are at the core of the switch detection algorithm proposed in Section 5.

Without loss of generality, in the following we will denote with $x \in \mathbb{R}^+$ a concentration variable and with y its noisy measurement, omitting the subscript i.

In order to detect switches in a molecule concentration dynamics, we introduce the *switching index* $o(k)$, which performs a nonlinear filtering on experimental data and emphasizes switches. The index is defined as follows:

$$o(k) = \frac{x(k+1) - x(k)}{x(k) - x(k-1)} \quad . \tag{6}$$

From (5) it is easy to verify that $o(k)$ takes the constant value $-\tilde{\gamma}^j = e^{-\gamma^j T}$ for every k such that $x(k-1)$, $x(k)$, $x(k+1)$ are generated by the j-th mode. As we can see in Fig. 1, a switch causes the index to move away from its constant value at times $k-1$ and k. Therefore, switching instants can be detected by looking for instants at which $o(k)$ is not constant.

Since we aim at using noisy experimental data, it is interesting to consider the behavior of switching indexes computed on Moving Average (MA) values $\bar{x}$

of concentrations

$$\bar{x}(k) = \frac{1}{W-2} \sum_{i=1}^{W-2} x(k-w+i) \ , \tag{7}$$

where $W \geq 3$ and $0 \leq w < W$ is an offset. Parameters W and w give rise to the family of switching indexes

$$\bar{o}(k) = \bar{o}(k; W, w) = \frac{\bar{x}(k+1) - \bar{x}(k)}{\bar{x}(k) - \bar{x}(k-1)} = \frac{x(k-w+W-1) - x(k-w+1)}{x(k-w+W-2) - x(k-w)} \ , \tag{8}$$

where W is the number of samples between the first and the last concentration values involved in the index definition and w is the offset between the first sample of x used in the index and the instant k at which the index is defined. Notice that $\bar{o}(k; 3, 1) \equiv o(k)$.

As shown in Fig. 1, $\bar{o}(k)$ behaves like $o(k)$, taking the constant value $-\tilde{\gamma}_i^j = e^{-\gamma^j T}$ when all samples $x(k-w), \ldots, x(k-w+W-1)$ belong to the same mode. However, the MA operation smooths out the behavior of $\bar{o}(k)$ and leads to an increase in the number of instants where $\bar{o}(k)$ is nonconstant because of a switch.

4.1 Statistical Analysis of Switching Indexes

In this section, we study the statistical properties of $\bar{o}(k)$ when it is computed on the basis of the noisy measurements $y(k)$. Let $\tilde{o}(k; W, w)$ be defined as in (8) but replacing $x(k)$ with $y(k)$. Denoting with Y_1 and Y_2 the numerator and denominator of $\tilde{o}(k)$ respectively, one has

$$\begin{aligned} &Y_1 = y(k-w+W-1) - y(k-w+1) \\ &Y_1 \sim N\left(x(k-w+W-1) - x(k-w+1),\, 2\sigma_n^2\right) \end{aligned} \tag{9}$$

$$\begin{aligned} &Y_2 = y(k-w+W-2) - y(k-w) \\ &Y_2 \sim N\left(x(k-w+W-2) - x(k-w),\, 2\sigma_n^2\right) \ . \end{aligned} \tag{10}$$

The Random Variables (RVs) Y_1 and Y_2 are jointly Gaussian and independent for $W > 3$ or characterized by a correlation coefficient $\rho = -\frac{1}{2}$ for $W = 3$. Therefore, the problem of finding the probability density function (pdf) of $\tilde{o}(k)$ amounts to the problem of characterizing the pdf of the ratio of two RVs, which has been investigated since the thirties [15,16,17,18].

Let $f_{\tilde{o}(k)}$ be the pdf of $\tilde{o}(k)$. From [15] and [18], $f_{\tilde{o}(k)}$ can be expressed as the pdf of a modified Cauchy distribution. Possible shapes of $f_{\tilde{o}(k)}$ are plotted in Fig. 2. In particular, the mean of Cauchy distributions is not defined and also the median can be a misleading estimator of the true index value $\bar{o}(k)$ (see Fig. 2), especially for high noise levels. However, Fieller's theorem [19] allows one to construct confidence sets for $\bar{o}(k)$, as shown in the next theorem.

Theorem 1. *Let Y_1 and Y_2 be the jointly Gaussian RVs defined by* (9)–(10) *and characterized by a correlation coefficient $\rho = -\frac{1}{2}\varrho$, with*

$$\varrho = \begin{cases} 1, & \text{if} \quad W = 3 \\ 0, & \text{if} \quad W > 3 \end{cases} \ . \tag{11}$$

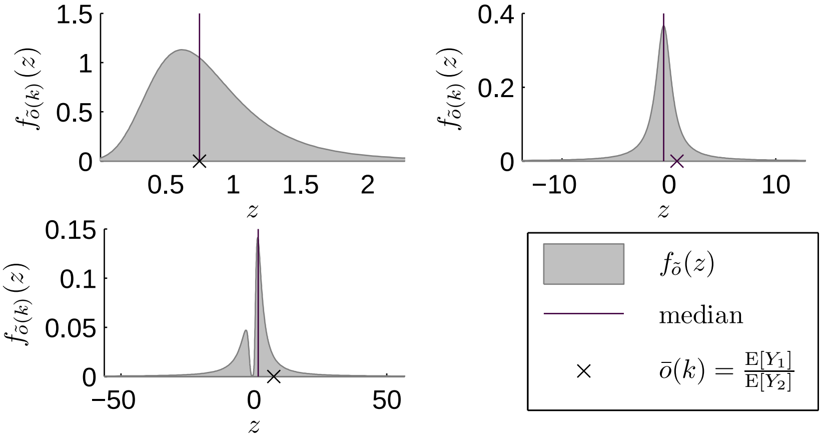

Fig. 2. Possible shapes of the pdf $f_{\tilde{o}(k)}$ of $\tilde{o}(k) = Y_1/Y_2$.

Let $\bar{o}(k)$ *be defined as in* (8)*, i.e.* $\bar{o}(k) = \mathrm{E}[Y_1]/\mathrm{E}[Y_2]$*, where* $\mathrm{E}[\,\cdot\,]$ *denotes the expectation operator.*
Then,

$$T\big([Y_1\ Y_2],\, \bar{o}(k)\big) = \frac{Y_1 - \bar{o}(k) Y_2}{\sigma_n \sqrt{2\left(\bar{o}(k)^2 + \varrho \bar{o}(k) + 1\right)}} \sim N(0,1) \tag{12}$$

is a pivotal quantity [20] for $\bar{o}(k)$ *and* $(1-\alpha)$*-level confidence sets for* $\bar{o}(k)$ *are given by*

$$S_\alpha\big([Y_1\ Y_2]\big) = \begin{cases} [\bar{o}_1^*, \bar{o}_2^*]\,, & \text{if } \ Y_2^2 > 2\bar{z}_\alpha^2 \sigma_n^2 \\ \mathrm{I\!R} \setminus (\bar{o}_1^*, \bar{o}_2^*)\,, & \text{if } \ Y_2^2 < 2\bar{z}_\alpha^2 \sigma_n^2 \text{ and} \\ & \quad 2\left(Y_1^2 + \varrho Y_1 Y_2 + Y_2^2\right) > (4-\varrho)\bar{z}_\alpha^2 \sigma_n^2 \\ \mathrm{I\!R}, & \text{if } \ Y_2^2 < 2\bar{z}_\alpha^2 \sigma_n^2 \text{ and} \\ & \quad 2\left(Y_1^2 + \varrho Y_1 Y_2 + Y_2^2\right) \le (4-\varrho)\bar{z}_\alpha^2 \sigma_n^2 \end{cases} , \tag{13}$$

with

$$\bar{o}_{1,2}^* = \frac{Y_1 Y_2 + \varrho \bar{z}_\alpha^2 \sigma_n^2 \mp \bar{z}_\alpha \sigma_n \sqrt{2\left(Y_1^2 + \varrho Y_1 Y_2 + Y_2^2\right) - (4-\varrho)\bar{z}_\alpha^2 \sigma_n^2}}{Y_2^2 - 2\bar{z}_\alpha^2 \sigma_n^2} \ , \tag{14}$$

where $\bar{z}_\alpha = -z_{\alpha/2} = z_{1-\alpha/2} > 0$ *and* z_p *is the generic* p*-th quantile of the standard normal distribution.*

Proof. According to the rationale of Fieller's theorem [19], a pivotal quantity for $\bar{o}(k)$ can be obtained considering the following linear combination of Y_1 and Y_2:

$$Y_1 - \bar{o}(k) Y_2 \sim N(0, 2\sigma_n^2 \left(\bar{o}(k)^2 + \varrho \bar{o}(k) + 1\right) \ . \tag{15}$$

Quantity $T\big([Y_1\ Y_2],\, \bar{o}(k)\big)$ defined in (12) is obtained standardizing the normal RV (15) and is therefore a pivotal quantity for $\bar{o}(k)$ since its distribution does

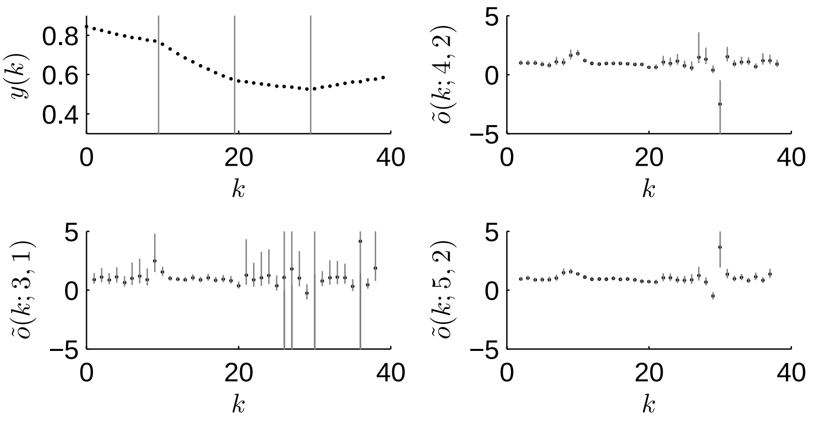

Fig. 3. Switching indexes $\tilde{o}(k;3,1)$, $\tilde{o}(k;4,2)$ and $\tilde{o}(k;5,2)$, computed on noisy data $y(k)$ with $\sigma_n = 10^{-3}$, and confidence sets for $\bar{o}(k)$.

not depend on $\bar{o}(k)$.
Since $T\big([Y_1\,Y_2],\,\bar{o}(k)\big) \sim N(0,1)$, it holds that

$$P\left(\left|T\big([Y_1\,Y_2],\,\bar{o}(k)\big)\right| \le \bar{z}_\alpha\right) = 1-\alpha\ . \tag{16}$$

Therefore a $(1-\alpha)$-level confidence set for $\bar{o}(k)$ is

$$S_\alpha\big([Y_1\,Y_2]\big) = \left\{\bar{o}(k) \in \mathbb{R} : \left|T\big([Y_1\,Y_2],\,\bar{o}(k)\big)\right| \le \bar{z}_\alpha\right\}\ . \tag{17}$$

To get an explicit expression of $S_\alpha\big([Y_1\,Y_2]\big)$, one has to find the values of $\bar{o}(k)$ verifying the inequality

$$T\big([Y_1\,Y_2],\,\bar{o}(k)\big)^2 = \frac{\big(Y_1 - \bar{o}(k)Y_2\big)^2}{2\sigma_n^2\left(\bar{o}(k)^2 + \varrho\bar{o}(k) + 1\right)} \le \bar{z}_\alpha^2\ , \tag{18}$$

which can be conveniently rewritten as

$$\left(Y_2^2 - 2\bar{z}_\alpha^2\sigma_n^2\right)\bar{o}(k)^2 - 2\left(Y_1Y_2 + \varrho\bar{z}_\alpha^2\sigma_n^2\right)\bar{o}(k) + \left(Y_1^2 - 2\bar{z}_\alpha^2\sigma_n^2\right) \le 0\ . \tag{19}$$

Solutions to (19) depend on the sign of $\left(Y_2^2 - 2\bar{z}_\alpha^2\sigma_n^2\right)$ and of the discriminant $\varDelta$ of the associated quadratic equation, given by:

$$\varDelta = 4\bar{z}_\alpha^2\sigma_n^2\left[2\left(Y_1^2 + \varrho Y_1Y_2 + Y_2^2\right) - (4-\varrho)\bar{z}_\alpha^2\sigma_n^2\right]\ . \tag{20}$$

Noticing that if $Y_2^2 - 2\bar{z}_\alpha^2\sigma_n^2 > 0$ then $\varDelta > 0$, confidence sets $S_\alpha\big([Y_1\,Y_2]\big)$ obtained as solutions to (19) are given by (13). □

From (13) it follows that the sets $S_\alpha\big([Y_1\,Y_2]\big)$ become smaller as W increases, as shown in Fig. 3.

5 Switch Detection Based on the Index $\bar{o}(k)$

In this section we introduce a first algorithm to detect switches between molecule domains from noisy data $y(k)$ of molecule concentrations. The proposed method uses the confidence sets for $\bar{o}(k)$ given in Theorem 1 in order to decide whether consecutive data points have been generated by the same mode or not. Moreover, the algorithm uses indexes with different values of W in order to exploit the statistical advantages of MA-based switching indexes.

For illustrating the core of the algorithm, we assume that W_M consecutive data up to $y(k_M)$ have been already aggregated, i.e. attributed to the same mode of operation. The question is whether to aggregate a subsequent data $y(k_a)$, with $k_a > k_M$. Aggregated data are characterized by a switching index involving all of them, namely $\tilde{o}(k_M; W_M, W_M - 1)$. To decide about the aggregation of $y(k_a)$, we consider the switching index involving W_a data up to $y(k_a)$, i.e. $\tilde{o}(k_a; W_a, W_a - 1)$, using a small value of W_a to avoid an excessive smoothing which could compromise switch detection.

In a noiseless setting one would perform the aggregation of $y(k_a)$ if $\bar{o}(k_a; W_a, W_a - 1) = \bar{o}(k_M; W_M, W_M - 1)$. However, from the noisy data $y(k)$ one can only compute the confidence sets I_M and I_a, for $\bar{o}(k_M; W_M, W_M - 1)$ and $\bar{o}(k_a; W_a, W_a - 1)$ respectively. As a decision rule, we detect a switch if $I_M \cap I_a = \emptyset$. It is easy to prove that, with such a rule, the probability of detecting an erroneous switch is lower than 2α, where $(1 - \alpha)$ is the confidence level. The main drawback of the proposed decision rule is that it does not allow one to detect switches if confidence sets are overlapping and this problem becomes critical for high noise levels.

Algorithm 1 shows the whole method for switch detection. The basic aggregation strategy is enhanced with a backtracking strategy, in lines 14–20, which double checks already aggregated data exploiting the new information carried by the most recent data. Moreover, infinite confidence sets are treated in a specific way and the aggregation test at the corresponding time instants is skipped (function firstFiniteConfidenceSet). The output of Algorithm 1 is the set E of time intervals corresponding to sets of data generated by the same mode. Such intervals could be overlapping, when switching instants are not precisely detected. Note also that the only piece of information about data needed in Algorithm 1 is the noise variance σ_n^2.

6 Switch Detection Based on Nonlinear Estimation

In this section we present an algorithm where data aggregation is based on the estimation of the model describing the concentration dynamics within a molecule domain.

Algorithm 1 Index based switch detection

```
Require: $Y = [y(1)\, y(2)\, \ldots\, y(N)]$, $W_a \geq 3$, $g_{\max} > 0$, $\sigma_n > 0$
initialize $E = \emptyset$ $W_M = W_a$, $k_M = W_M$
$k_M, I_M =$ firstFiniteConfidenceSet$(Y, k_M, W_M)$
$k_a, I_a =$ firstFiniteConfidenceSet$(Y, k_M + 1, W_a)$
while $(k_M < N)$ and $(k_a \leq N)$ do
    if $(I_M \cap I_a = \emptyset)$ or $(k_a - k_M > g_{\max})$ then
        if $W_M > W_a$ then
            $E = E \cup [k_M - W_M + 1, k_M]$
        end if
        $W_M = W_a$
        $k_M, I_M =$ firstFiniteConfidenceSet$(Y, k_M + 1, W_M)$
    else
        $W_M = W_M + k_a - k_M$
        $k_M = k_a$
        $I_M =$ confidence set for $\bar{o}(k_M - W_M + 2; W_M - 1, 0)$
        $I_a =$ confidence set for $\bar{o}(k_M - W_M + 1; W_a, 0)$
        while $(W_M > W_a)$ and $\big((I_M \cap I_a = \emptyset)$ or $(I_a$ is not a finite interval$)\big)$ do
            $W_M = W_M - 1$
            $I_M =$ confidence set for $\bar{o}(k_M - W_M + 2; W_M - 1, 0)$
            $I_a =$ confidence set for $\bar{o}(k_M - W_M + 1; W_a, 0)$
        end while
        $I_M =$ confidence set for $\bar{o}(k_M; W_M, W_M - 1)$
    end if
    $k_a, I_a =$ firstFiniteConfidenceSet$(Y, k_M + 1, W_a)$
end while
if $W_M > W_a$ then
    $E = E \cup [k_M - W_M + 1, k_M]$
end if
return $E$
function firstFiniteConfidenceSet$(Y, k, W)$
$I =$ confidence set for $\bar{o}(k; W, W - 1)$
while ($I$ is not a finite interval) and $(k \leq N)$ do
    $k = k + 1$
    if $k \leq N$ then
        $I =$ confidence set for $\bar{o}(k; W, W - 1)$
    end if
end while
return $k, I$
```

If $x(k_0), x(k_0 + 1), \ldots, x(\bar{k})$ have been all generated by the j-th mode of operation, the output measurements $y(k)$, $k = k_0, k_0 + 1, \ldots, \bar{k}$, are given by

$$\begin{aligned} y(k) &= x(k) + n(k) \\ &= \frac{\kappa^j}{\gamma^j} - \left(\frac{\kappa^j}{\gamma^j} - x_0^j\right) e^{-\gamma^j \left(kT - t_0^j\right)} + n(k) \ , \end{aligned} \qquad (21)$$

Algorithm 2 Nonlinear estimation based switch detection

```
Require: $Y = [y(1)\, y(2)\, \ldots\, y(N)]$, $W_a \geq 3$, $\sigma_n > 0$
initialize $E = \emptyset$, $W_M = W_a$, $k_M = W_M$
while $k_M < N$ do
    compute estimates $\widehat{\kappa^j}$, $\widehat{\gamma^j}$, $\widehat{x_0^j}$ on $y(k_M - W_M + 1), \ldots, y(k_M)$
    $I_a$ = confidence interval for $y(k_M + 1)$ under $\mathrm{H}_0$
    if $y(k_M + 1) \notin I_a$ then
        if $W_M > W_a$ then
            $E = E \cup [k_M - W_M + 1, k_M]$
        end if
        $W_M = W_a$
        $k_M = k_M + 1$
    else
        $W_M = W_M + 1$
        $k_M = k_M + 1$
        compute estimates $\widehat{\kappa^j}$, $\widehat{\gamma^j}$, $\widehat{x_0^j}$ on $y(k_M - W_M + 2), \ldots, y(k_M)$
        $I_a$ = confidence interval for $y(k_M - W_M + 1)$ under $\mathrm{H}_0$
        while $(W_M > W_a)$ and $(y(k_M - W_M + 1) \notin I_a)$ do
            $W_M = W_M - 1$
            compute estimates $\widehat{\kappa^j}$, $\widehat{\gamma^j}$, $\widehat{x_0^j}$ on $y(k_M - W_M + 2), \ldots, y(k_M)$
            $I_a$ = confidence interval for $y(k_M - W_M + 1)$ under $\mathrm{H}_0$
        end while
    end if
end while
if $W_M > W_a$ then
    $E = E \cup [k_M - W_M + 1, k_M]$
end if
return $E$
```

where $n(k) \sim WGN\left(0, \sigma_n^2\right)$, T and $t_0^j = k_0 T$ are known values and κ^j, γ^j, $x_0^j = x(k_0)$ represent the unknown model parameters. The estimates $\widehat{\kappa^j}$, $\widehat{\gamma^j}$ and $\widehat{x_0^j}$ of κ^j, γ^j and x_0^j can be obtained using nonlinear least squares [21].

In order to illustrate the decision rule for aggregation, assume that W_M consecutive data up to $y(k_M)$ have been aggregated and let $\widehat{\kappa^j}$, $\widehat{\gamma^j}$ and $\widehat{x_0^j}$ be the estimates produced by nonlinear least squares methods using these data. We can perform an hypothesis test to evaluate if $y(k_M + 1)$ can be described by the estimated model. The null hypothesis H_0 is that $y(k_M + 1)$ belongs to the same dynamics of the aggregated data. Under H_0, we consider the confidence interval I_a for the measurement at $k_M + 1$ generated according to the estimated model. Aggregation of $y(k_M + 1)$ is therefore performed if $y(k_M + 1) \in I_a$, detecting a switch otherwise. As described in [21], the interval I_a can be computed by linearizing model (21).

In Algorithm 1, the decision rule based on the comparisons of confidence sets can be replaced by the hypothesis test described above, thus obtaining Algorithm 2 listed below.

7 Algorithms Validation

In order to test the proposed algorithms, we considered a set of simulated time series. We generated 42 typical noiseless behaviors for gene expression, each containing two switches among different modes of operation. Then we produced 16800 time series by adding noise with $\sigma_n \in \left\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}\right\}$. The total number of noisy data was 702400.

The performances of Algorithm 1 and 2 can be evaluated introducing suitable performance indexes. Such quantities are computed comparing the algorithms output (the set E of estimated time intervals) with the true intervals collecting data generated by the same mode of operation in each time series. Comparisons are performed between corresponding intervals, where the correspondence criterion is chosen in term of maximum overlap. A formal introduction of performance indexes is based on a set description of estimated and true intervals.

Let $2^{\mathbb{N}}$ be the powerset of $\mathbb{N}$ and $\mathbb{I} \subset 2^{\mathbb{N}}$ be the set of all intervals of consecutive natural numbers, defined as follows:

$$\forall I \in \mathbb{I} \;\; \exists\, a(I), b(I) \in \mathbb{N} \text{ such that } a(I) \leq k \leq b(I), \forall k \in I \;\; , \tag{22}$$

where $a(I)$ and $b(I)$ represent the endpoints of a generic interval I. The set of true intervals, denoted by T, is characterized as follows:

$$\begin{aligned} T = &\left\{I_1^T, I_2^T, \ldots, I_{n_T}^T\right\} \subset \mathbb{I} \\ &\text{with } I_i^T \cap I_j^T = \emptyset, \forall i \neq j, \text{ and } \bigcup_{i=1}^{n_T} I_i^T \in \mathbb{I} \;\; . \end{aligned} \tag{23}$$

The set E of estimated intervals is defined assuming a particular order:

$$\begin{aligned} E = &\left\{I_1^E, I_2^E, \ldots, I_{n_E}^E\right\} \subset \mathbb{I} \\ &\text{with } a(I_i^E) < a(I_{i+1}^E) \text{ and } b(I_i^E) < b(I_{i+1}^E), \; \forall i = 1, \ldots, n_E - 1 \;\; . \end{aligned} \tag{24}$$

Since intervals in E could be overlapped, we decided to remove common points between intervals $I_i^E \in E$, obtaining the new set of intervals

$$E' = \left\{I_1^{E'}, I_2^{E'}, \ldots, I_{n_{E'}}^{E'}\right\} \subset \mathbb{I} \;\; , \tag{25}$$

constructed as follows:

$$\begin{aligned} E' = &\left\{I_i'^E \neq \emptyset, \; i = 1, \ldots, n_E\right\} \\ &\text{where } \;\; I_i'^E = I_i^E \setminus \left(I_{i-1}^E \cup I_{i+1}^E\right), \; \forall i = 1, \ldots, n_E, \\ &\text{and } I_0^E = I_{n_E+1}^E = \emptyset \;\; . \end{aligned} \tag{26}$$

Every $I_i^{E'} \in E'$ is associated to an interval $I_{t_i}^T \in T$, where $t_i \in \{1, \ldots, n_T\}$ is computed as follows:

$$t_i = \arg \max_{j \in \{1, \ldots, n_T\}} \left| I_i^{E'} \cap I_j^T \right|, \quad \forall i \in \{1, \ldots, n_{E'}\} \;\; . \tag{27}$$

Notice that more than one t_i might satisfy (27). In these situations, t_i is assigned as the index of the first true interval according to the temporal order.

In a similar way to (27), each true interval $I_j^T \in T$ is associated to the estimated interval $I_{m_j}^{E'} \in E'$ having the maximum overlap with I_j^T. This correspondence is achieved finding $m_j \in \{1, \ldots, n_{E'}\}$, $\forall j \in \{1, \ldots, n_T\}$, as:

$$m_j = \begin{cases} -1, & \text{if } \nexists i \in \{1, \ldots, n_{E'}\} \mid t_i = j \\ \arg \max\limits_{\substack{i \in \{1, \ldots n_{E'}\} \\ \text{s.t.} t_i = j}} \left| I_i^{E'} \cap I_j^T \right|, & \text{otherwise} \end{cases} . \qquad (28)$$

Possible ambiguities are solved as for t_i.

Starting from definitions (23)–(28), we introduce particular sets to formally define quantities useful for performance evaluation:

1. let $M = \{m_j \neq -1,\ j = 1, \ldots, n_T\}$, then:

$$(a) \quad n_M = |M| \ , \qquad (b) \quad N_M = \sum_{i \in M} \left| I_i^{E'} \cap I_{t_i}^T \right| \ ; \qquad (29)$$

2. let $S = \{1, \ldots, n_{E'}\} \setminus M$, then:

$$(a) \quad n_S = |S| = n_{E'} - |M| \ , \qquad (b) \quad N_S = \sum_{i \in S} \left| I_i^{E'} \cap I_{t_i}^T \right| \ . \qquad (30)$$

Quantity n_M is the number of true intervals having a correspondence with at least one estimated interval, while N_M is the total number of instants belonging to both true intervals and estimated intervals characterized by the maximum correspondence. On the other hand, n_S is the number of estimated intervals with a non-maximum correspondence with a true interval, and N_S is the total number of instants belonging to such intervals and to the corresponding true intervals.

Starting from quantities (29)–(30) it is possible to define performance indexes which highlight important characteristics of the algorithms. In particular, we can consider a classification accuracy index *acc* and a fragmentation index $frag$. Accuracy is evaluated considering N_M as the number of correctly classified data, thus defining

$$acc = \frac{N_M}{N_T} \ , \qquad (31)$$

where N_T is the total number of analyzed data. The fragmentation index expresses the tendency of an algorithm to overestimate the number of switches by fragmenting a true interval into different estimated intervals. Since such a fragmentation leads to estimated intervals characterized by a non-maximum correspondence, $frag$ is defined as

$$frag = \frac{n_S}{(n_S + n_M)} = \frac{n_S}{n_{E'}} \ . \qquad (32)$$

Table 1. Accuracy and fragmentation performances of Algorithm 1 and 2 at different noise levels.

	Algorithm 1			
σ_n	10^{-5}	10^{-4}	10^{-3}	10^{-2}
acc	97.1%	93.8%	69.7%	22.3%
$frag$	4.4%	5.2%	16.4%	34.3%

	Algorithm 2			
σ_n	10^{-5}	10^{-4}	10^{-3}	10^{-2}
acc	75.3%	80.7%	69.7%	63.8%
$frag$	34.4%	26.9%	30.7%	15.2%

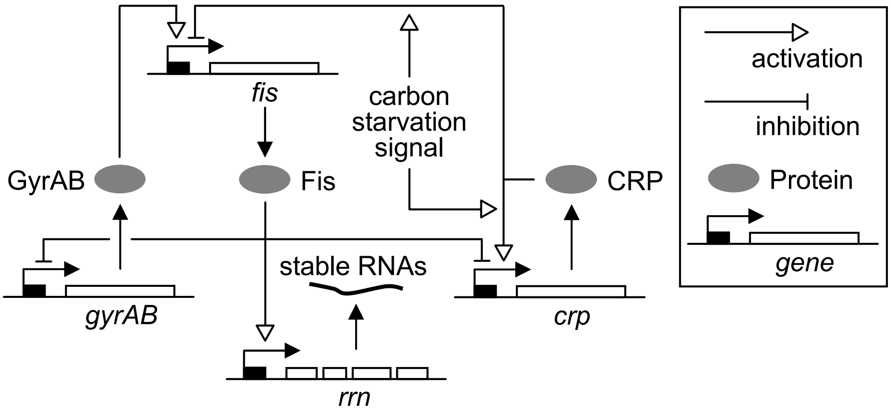

Fig. 4. Graphical representation of the simplified carbon starvation response network in *E. coli*.

Good performances are characterized by values of acc and $frag$ close to 1 and 0 respectively.

The results are reported in Table 1. We can observe that Algorithm 1 has good performances for noise levels $\sigma_n = 10^{-5}$ and $\sigma_n = 10^{-4}$. In these cases, Algorithm 2 has lower accuracy and exhibits the tendency to overestimate the number of switches thus fragmenting true intervals. As σ_n increases, we notice the sensitivity of Algorithm 1 to noise. Such a behavior is consistent with the statistical analysis of switching indexes presented in Section 4.1. On the contrary, Algorithm 2 exhibits lower performance degradation. Notice that at the highest noise level the use of Algorithm 1 is heavily compromised.

8 Switch Detection in a PWA Model of the Carbon Starvation Response of *E. coli*

In this section, we present the application of switch detection algorithms to data generated by the GRN regulating carbon starvation response in *E. coli*. A complete model of this network has been proposed in [22]. For this study we considered the simplified GRN used by Drulhe et al. in [1] and shown in Fig. 4.

The network involves interactions between genes *crp*, *fis*, *gyrAB* and their products (proteins CRP, Fis, GyrAB), regulating the synthesis of stable RNAs.

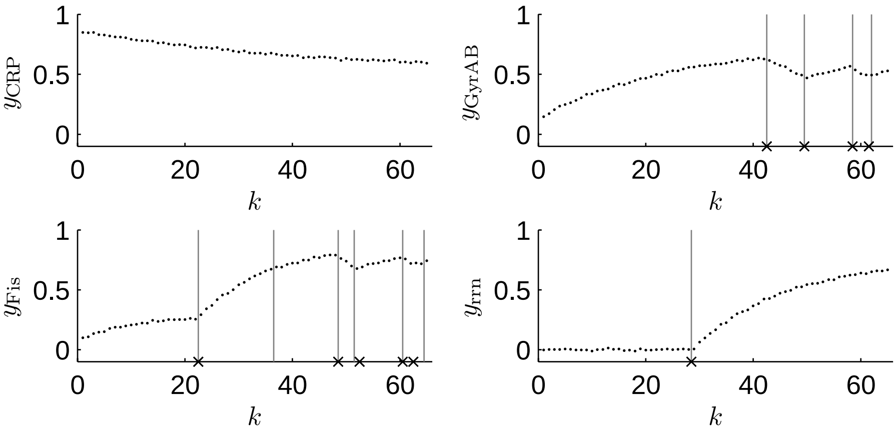

Fig. 5. Switch detection results on simulated data of the GRN depicted in Fig. 4. Variables y_{CRP}, y_{Fis}, y_{GyrAB}, and y_{rrn} denote the concentration measurements of proteins CRP, Fis, GyrAB, and stable RNAs. Vertical lines denote detected switches, while crosses correspond to real switching times.

In response to a carbon starvation signal, the regulatory mechanisms inhibit the synthesis of stable RNAs and then *E. coli* cells abandon their exponentially-growth state to enter a more resistent non-growth state called stationary phase.

Switch detection algorithms have been applied to time series of concentration data simulated with initial conditions, sampling time and noise level similar to the data used in [1]. In particular, data refer to the reentry into an exponential growth phase after a carbon upshift, while noise and sampling characteristics are close to ones that characterize measurements produced by gene reporter systems. Figure 5 shows switches detected using Algorithm 2. In this case, the noise level was too high to obtain reasonable results with Algorithm 1. All switches have been identified except for the concentration of protein Fis where a spurious switch has been introduced.

9 Conclusions

In this paper we considered the problem of detecting switches in gene expression profiles, by assuming that the underlying model of the GRN has a PW-OE structure. First, we proposed a method based on switching indexes that emphasizes the occurrence of switches without exploiting any information on the parameters of the PW-OE model. Then, we introduced a second algorithm based on nonlinear identification techniques and hypothesis testing. An extensive testing of the two methods highlighted that they are complementary since the first algorithm outperforms the second one for low noise levels while the second one produces better results for high noise levels. Future research will consider the generalization of the algorithms to the case of data sets including sliding-mode behaviors. We will also study methods based on the knowledge of switching times for attributing data points to the modes of operation of the whole network. In a broader perspective, these procedures will be integrated with the algorithm proposed in [1] with the goal of reconstructing all parameters characterizing PWA models of GRNs.

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

References

1. Drulhe, S., Ferrari-Trecate, G., de Jong, H., Viari, A.: Reconstruction of switching thresholds in piecewise-affine models of genetic regulatory networks. In Hespanha, J., Tiwari, A., eds.: Proc. Hybrid Systems: Computation and Control (HSCC 2006). Volume 3927 of LNCS. Sringer-Verlag, Berlin (2006) 184–199
2. Lockhart, D., Winzeler, E.: Genomics, gene expression and DNA arrays. Nature **405**(6788) (2000) 827–836
3. 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**(16) (2002) 10555–10560
4. Glass, L., Kauffman, S.: The logical analysis of continuous, non-linear biochemical control networks. J. Theor. Biol. **39**(1) (1973) 103–129
5. 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**(2) (2004) 301–340
6. 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
7. Casey, R., de Jong, H., Gouzé, J.L.: Piecewise-linear models of genetic regulatory networks: Equilibria and their stability. J. Math. Biol. **52**(1) (2005) 27–56
8. Ferrari-Trecate, G., Muselli, M., Liberati, D., Morari, M.: A clustering technique for the identification of piecewise affine systems. Automatica **39**(2) (2003) 205–217
9. Bemporad, A., Garulli, A., Paoletti, S., Vicino, A.: A bounded-error approach to piecewise affine system identification. IEEE Trans. Autom. Control **50**(10) (2005) 1567–1580
10. Juloski, A., Weiland, S., Heemels, W.: A bayesian approach to identification of hybrid systems. IEEE Trans. Autom. Control **50**(10) (2005) 1520–1533
11. Ma, Y., Vidal, R.: Identification of deterministic switched ARX systems via identification of algebraic varieties. In Morari, M., Thiele, L., eds.: Proc. Hybrid Systems: Computation and Control (HSCC 2005). Volume 3414 of LNCS. Springer-Verlag, Berlin (2005) 449–465
12. Gustafsson, F.: Adaptive Filtering and Change Detection. John Wiley & Sons, Chichester, West Sussex, England (2000)
13. Juloski, A., Weiland, S.: A bayesian approach to the identification of piecewise linear output error models. IFAC Symposium on System Identification (SYSID) 2006 (to appear)

14. Rosenqvist, F., Karlströmb, A.: Realisation and estimation of piecewise-linear output-error models. Automatica **41**(3) (2005) 545–551
15. Fieller, E.: The distribution of the index in a normal bivariate population. Biometrika **24**(3-4) (1932) 428–440
16. Marsaglia, G.: Ratios of normal variables and ratios of sums of uniform variables. J. Amer. Stat. Assoc. **60**(309) (1965) 193–204
17. Hinkley, D.: On the ratio of two correlated normal random variables. Biometrika **56**(3) (1969) 635–639
18. Pham-Gia, T., Turkkan, N., Marchand, E.: Density of the ratio of two normal random variables. Technical Report 8, Université de Moncton and University of New Brunswick, Canada (2004)
19. Fieller, E.: A fundamental formula in the statistics of biological assay, and some applications. Quart. J. Pharm. Pharmacol. **17**(2) (1944) 117–123
20. Rohatgi, V., Saleh, A.: An Introduction to Probability and Statistics. 2nd edn. John Wiley & Sons, New York (2001)
21. Ljung, L.: System Identification: Theory for the User. 2nd edn. Prentice-Hall, Upper Saddle River, NJ (1999)
22. Ropers, D., de Jong, H., Page, M., Schneider, D., Geiselmann, J.: Qualitative simulation of the carbon starvation response in *Escherichia coli.* BioSystems **84**(2) (2006) 124–152