Structural Identification of Piecewise-Linear Models of Genetic Regulatory Networks

Riccardo Porreca  Samuel Drulhe  Hidde de Jong
Giancarlo Ferrari-Trecate *

Keywords: Genetic Networks, Reverse Engineering, Differential Equations, Hybrid Systems.

*Corresponding author.
Contact Information

Riccardo Porreca
Dipartimento di Informatica e Sistemistica, Università degli Studi di Pavia, via Ferrata 1, 27100 Pavia, Italy.
Phone: +39 0382 985756, Fax: +39 0382 985373/525638
E-mail: riccardo.porreca@unipv.it

Samuel Drulhe
INRIA Grenoble-Rhône-Alpes,
655 avenue de l’Europe, Montbonnot, 38334 Saint-Ismier Cedex, France.
Phone: 33 4 76 61 53 24, Fax: +33 4 76 61 54 08
E-mail: Samuel.Drulhe@inria.fr

Hidde de Jong
INRIA Grenoble-Rhône-Alpes,
655 avenue de l’Europe, Montbonnot, 38334 Saint-Ismier Cedex, France.
Phone: +33 4 76 61 53 35, Fax: +33 4 76 61 54 08
E-mail: Hidde.de-Jong@inria.fr

Giancarlo Ferrari-Trecate
Dipartimento di Informatica e Sistemistica, Università degli Studi di Pavia, via Ferrata 1, 27100 Pavia, Italy.
Phone: +39 0382 985791, Fax: +39 0382 985373/525638
E-mail: giancarlo.ferrari@unipv.it
Abstract

We present a method for the structural identification of genetic regulatory networks (GRNs), based on the use of a class of Piecewise-Linear (PL) models. These models consist of a set of decoupled linear models describing the different modes of operation of the GRN and discrete switches between the modes accounting for the nonlinear character of gene regulation. They thus form a compromise between the mathematical simplicity of linear models and the biological expressiveness of nonlinear models. The input of the PL identification method consists of time-series measurements of concentrations of gene products. As output it produces estimates of the modes of operation of the GRN as well as all possible minimal combinations of threshold concentrations of the gene products accounting for switches between the modes of operation. The applicability of the PL identification method has been evaluated using simulated data obtained from a model of the carbon starvation response in the bacterium *Escherichia coli*. This has allowed us to systematically test the performance of the method under different data characteristics, notably variations in the noise level and the sampling density.
1 Introduction

The large amounts of gene expression data that have become available provide a rich source of information for the identification of the structure of genetic regulatory networks (GRNs). Various so-called reverse engineering approaches have been proposed in the literature (see Bansal et al., 2007; Cho et al., 2007; Gardner and Faith, 2005; Markowetz and Spang, 2007; van Riel, 2006, for recent reviews). The quality and quantity of currently available data have often limited the approaches to linear models (e.g., Chen et al., 1999; D’haeseleer et al., 2000; Gardner et al., 2003; Guthke et al., 2005; Lemeille et al., 2005; van Someren et al., 2000). Unfortunately, these models are only valid around a steady state and are not able to take into account the globally nonlinear character of most regulatory interactions. As an alternative, a variety of nonlinear models of gene regulation have been proposed in the literature (see de Jong, 2002, for a review). Their application in the context of reverse engineering is hampered, however, by the inherent computational complexity of the structural identification problem and the large amounts of data required.

In this paper, we present a method for the structural identification of GRNs, based on the use of a class of Piecewise-Linear (PL) models. These models, originally introduced by Glass and Kauffman (1973), consist of a set of decoupled linear models that each describe a particular mode of operation of the GRN. The system may switch from one linear model to another when the concentration of a regulator, e.g. a transcription factor, crosses a threshold concentration. PL models strike a compromise between linear and nonlinear models. Locally, the PL models have a particularly simple linear form, which makes them tractable for mathematical analysis. Globally, however, they are able to capture the nonlinear character of gene regulation, as demonstrated by
several modeling studies of regulatory networks of biological interest (e.g., Ghosh and Tomlin, 2004; Ropers et al., 2006). The study of PL models and their generalizations has been an active research area in both mathematical biology and hybrid systems theory (e.g., Batt et al., 2008; Belta et al., 2004; de Jong et al., 2004; Edwards et al., 2001; Ghosh and Tomlin, 2004; Mestl et al., 1995). In particular, powerful methods for the (structural) identification of PL systems have been developed in the hybrid systems community (see Paoletti et al., 2007, and the references therein), which might be profitably applied to the reconstruction of GRNs from experimental data.

Inspired by these hybrid-systems identification methods, we presented in previous work an algorithm that is capable of inferring from time-series data all possible minimal combinations of threshold concentrations of the regulators. By analyzing changes in the mode of operation following a threshold crossing, the possible regulatory interactions of the network can be identified (Drulhe et al., 2008). The threshold reconstruction algorithm is based on the simplifying assumptions that the data are noiseless and that the data points have already been assigned to the different modes of operation of the GRN. In other words, it is assumed that for each data point we know the linear mode from which it is generated. For the practical application of the algorithm to gene expression data these assumptions are not satisfactory. First, measurements of gene expression are noisy, and second, it is not obvious how to determine the modes of operation of a GRN and to classify the data points by attributing them to a mode.

In this paper, we therefore generalize the threshold reconstruction algorithm to noisy data, and present novel algorithms for inferring the modes of operation and classifying the data-points. These algorithms have been combined into an integrated method for the identification of PL models of GRNs from time-series data. In addition,

---

1In the hybrid systems community, the models under study are usually called *piecewise-affine* instead of piecewise-linear. We will follow the latter terminology in this paper, in order to adhere to standard usage in bioinformatics and mathematical biology.
the applicability of this method has been evaluated using simulated data obtained from a model of the carbon starvation response in the bacterium *Escherichia coli*. This has allowed us to systematically test the performance of the method under different data characteristics, notably variations in the noise level and the sampling density. For simulated data with characteristics similar to those obtained by means of fluorescent and luminescent reporter gene systems (Ronen *et al.*, 2002), the PL identification method is shown to yield good results. Most of the time, it generates a limited number of alternative PL models compatible with the data. Moreover, these models have high recall and precision scores, in the sense that they include most of the real and only a limited number of spurious threshold concentrations.

In Sec. 2 we briefly review PL models of genetic regulatory networks, followed by a description of the PL identification method in Sec. 3. The results of the application of the method to simulated data are presented in Sec. 4 and the discussion in Sec. 5 places our method in the context of related work.
2 PL models of genetic regulatory networks

We consider a GRN composed of \(n\) genes, each coding for a protein and/or an RNA, briefly referred to as a gene product. The products of a gene may regulate the expression of the gene itself or of other genes in the network. For example, in the network represented in Fig. 1(a) each of the two genes codes for a protein that functions as an activator or inhibitor of the expression of both genes. The concentrations of the gene products at time \(t \in \mathbb{R}_{\geq 0}\) are denoted by \(x_i(t), i \in \{1, \ldots, n\}\). In what follows, the gene products are also referred to as the regulators of the network, to emphasize their role in the dynamics of the GRN.

Following the modeling formalism originally proposed in (Glass and Kauffman, 1973), the dynamics of the gene products can be described by PL differential equations using step functions to account for the regulatory interactions. This is illustrated for the example network in Fig. 1(b). Below we generally define the models, in a way that highlights their PL structure.

We denote by \(x = (x_1, \ldots, x_n)' \in \Omega\) a vector of concentrations of gene products, where \(\Omega \subset \mathbb{R}_{\geq 0}^n\) is a bounded hyperrectangle that includes the origin as a vertex. For each concentration \(x_i, i \in \{1, \ldots, n\}\), we distinguish a set of constant, strictly positive threshold concentrations \(\{\theta_{i1}^1, \ldots, \theta_{ip_i}^p\}, p_i \geq 0\). At its threshold concentrations a protein or RNA may affect the expression of its own gene or the expression of genes coding for other regulators, thus changing the dynamics of the GRN. The grid of threshold hyperplanes, defined as

\[
\Theta = \bigcup_{i \in \{1, \ldots, n\}, l_i \in \{1, \ldots, p_i\}} \{x \in \Omega, x_i = \theta_{i l_i}^i\},
\]

splits \(\Omega\) into open hyperrectangular regions \(D^j, j \in \{1, \ldots, s\}\), called regulatory do-
mains. Notice that \( s = \prod_{i=1}^{n} (p_i + 1) \). Figure 2(a) shows the regulatory domains for the network in Fig. 1. In each regulatory domain, the system can be written in the PL form

\[
\dot{x} = \mu^j - \nu^j x, \quad \text{if} \quad \lambda(x) = j ,
\]

where \( \mu^j \) is a vector of \( n \) positive constants, \( \nu^j = \text{diag}(\nu^j_1, \ldots, \nu^j_n) \) a diagonal matrix of strictly positive constants, and \( \lambda \) a switching function, such that \( \lambda(x) = j \) iff \( x \in D^j \). That is, within each regulatory domain, the evolution of the concentrations is a balance between a synthesis term \( \mu^j \) and a degradation term \( \nu^j x \). Because \( \nu^j \) is a diagonal matrix, (2) is a system of decoupled linear differential equations. The set \( M = \{ (\mu^j, \nu^j), j = 1, \ldots, s \} \) collects the modes of operation of the network. Figure 2(b) gives the rate parameters of the linear model associated with regulatory domains in the example network.

We assume that the gene expression data that can be obtained are direct or indirect measurements of the concentrations of gene products at times \( t_k, k = 0, \ldots, m \). In order to be able to reconstruct the network dynamics we further assume that all sampling intervals \( T(k) = t_{k+1} - t_k \) are sufficiently small with respect to the time constants of the network. Moreover, we assume implicitly that for \( t \) ranging over \( [t_k, t_{k+1}] \), the pair \( (\mu^j(x(t)), \nu^j(x(t))) \) changes at most once. Notice that this is different from requiring that \( \lambda(x(t)) \) changes at most once: indeed, a change in the value of \( \lambda(x(t)) \) does not necessarily imply that the pair \( (\mu^j(x(t)), \nu^j(x(t))) \) changes.

By defining \( x_i(k) = x_i(t_k) \), a model of the data is provided by the following discrete-
time PL system

\[ x_i(k + 1) = \phi(\mu_i^j, \nu_i^j, x_i(k), T(k)), \quad \text{if } \lambda(x(k)) = j \quad (3) \]

\[ \phi(\mu_i^j, \nu_i^j, x_i(k), T(k)) = \frac{\mu_i^j}{\nu_i^j} - \left( \frac{\mu_i^j}{\nu_i^j} - x_i(k) \right) e^{-\nu_i^j T(k)} \]

\[ y_i(k) = x_i(k) + \xi_i(k) \quad (4) \]

where \( \xi_i(\cdot) \) is a white Gaussian noise with zero mean and variance \( \sigma_i^2 \), and \( y_i(k) \) are the measurements for the \( i \)th regulator. Since noise enters only the output equation (4), system (3)-(4) is an output error model (Ljung, 1999).

Using the discrete-time PL model, we can precisely define when the GRN switches from one mode of operation to another. Let \( \lambda_k = \lambda(x(k)) \) for all \( k = 0, \ldots, m \). \( k \) is a switching time if \( (\mu_\lambda^k, \nu_\lambda^k) \neq (\mu_\lambda^{k+1}, \nu_\lambda^{k+1}) \). By extension, \( k \) is a switching time for regulator \( i \) if \( (\mu_i^\lambda^k, \nu_i^\lambda^k) \neq (\mu_i^\lambda^{k+1}, \nu_i^\lambda^{k+1}) \).

The switching times can be used to define data segments generated by the same mode of operation of the GRN or of a particular gene. Let \( \tau \) be an interval \([k_0, k_1]\), such that \( k_0, k_1 \in \{0, \ldots, m\} \).\(^2\) Moreover, define the restriction \( y|_\tau \) as the data set \( \{y(k), k \in [k_0, k_1]\} \). We call \( y|_\tau \) a segment if no \( k \in [k_0, k_1 - 1] \) is a switching time. A segment is maximal if \( k_0 - 1 \) and \( k_1 \) are switching times (or equal \(-1\) and \( m \), respectively). Analogously, \( y_i|_\tau \) is a segment for the \( i \)th regulator if no \( k \in [k_0, k_1 - 1] \) is a switching time for regulator \( i \) and the segment is maximal if \( k_0 - 1 \) and \( k_1 \) are switching times for \( i \) (or equal \(-1\) and \( m \), respectively).

From the definition of a segment for the \( i \)th regulator and the output error model (3)-(4) it follows that the measurements \( y_i|_\tau \) over such a segment \( \tau = [k_0, k_1] \) verify the

\(^2\)We set \([k_0, k_1] = \emptyset\) if \( k_1 < k_0\).
\[ y_i(k) = \phi(\bar{\mu}_i, \bar{\nu}_i, x_i(k_0), t_k - t_{k_0}) + \xi_i(k), \]  

(5)

where \((\bar{\mu}_i, \bar{\nu}_i) = (\mu^\lambda_i, \nu^\lambda_i)\), for all \(k \in \tau\).

Let \(Y\) denote the set of maximal data segments on \(y|_{[0,m]}\). Notice that two different segments in \(Y\) may originate from the same mode of operation of the GRN. We can formally define this by means of an equivalence relation on data segments. Let \(\tau_0, \tau_1\) be two intervals and let \(y|_{\tau_0}, y|_{\tau_1}\) be two data segments. Moreover, let \((\bar{\mu}, \bar{\nu}) = (\mu^\lambda, \nu^\lambda)\), for all \(k \in \tau_0\) and \((\tilde{\mu}, \tilde{\nu}) = (\mu^\lambda, \nu^\lambda)\), for all \(k \in \tau_1\) (remind that, by definition, no mode switches occur inside a data segment). The data segments belong to the same equivalence class, written as \(y|_{\tau_0} \sim y|_{\tau_1}\), if \((\bar{\mu}, \bar{\nu}) = (\tilde{\mu}, \tilde{\nu})\). The partition of \(Y\) induced by the equivalence classes is denoted by \(Y = Y/\sim\). In a similar way, a partition of \(Y_i\), the set of maximal data segments of the \(i\)th regulator, can be defined. In the latter case, two data segments belong to the same equivalence class, written as \(y_i|_{\tau_0} \sim y_i|_{\tau_1}\), if \((\bar{\mu}_i, \bar{\nu}_i) = (\tilde{\mu}_i, \tilde{\nu}_i)\). The partition is denoted by \(Y_i = Y_i/\sim\).
3 PL identification method

3.1 Overview of the identification method

The identification of model (2) amounts to estimating the following quantities

- the number of different modes of operation of the GRN;
- the regulatory domains $D^j$, $j \in \{1, \ldots, s\}$, or equivalently, the thresholds on concentration variables defining the grid $\Theta$;
- the matrices $\mu^j$, $\nu^j$, $j \in \{1, \ldots, s\}$.

In practice, since it never happens that data are collected in all domains $D^j$, the best we can hope for is an estimation of the number of different modes of operation occurring in the data. Similarly, we only aim at estimating the threshold boundaries of these domains and the rate parameters associated to them.

As explained in (Drulhe et al., 2008), it is not convenient to solve the identification problem using standard techniques for the identification of hybrid systems. Rather, we propose the following gray-box procedure for reconstructing the above-mentioned quantities.

1. **Switch detection.** Estimate the maximal data segments of each regulator in the GRN.

2. **Data classification.** From the estimated data segments of the regulators, construct the partition consisting of blocks of data segments generated by the same mode of operation of the GRN. This yields the rate parameters defining the modes of operation as a by-product.
3. **Threshold reconstruction.** From the partitioned data segments, estimate possible thresholds on concentration variables and find all minimal combinations of thresholds that account for observed changes in the mode of operation.

In the remainder of this section, we summarize the algorithms developed for each step of the method. For lack of space we cannot specify these in full, but we give the main idea and refer to technical reports (Porreca *et al.*, 2006; Porreca and Ferrari-Trecate, 2007) for details. The method has been completely implemented in MatLab (The MathWorks).

### 3.2 Switch detection

The problem of finding maximal segments for each of the regulators in the data set can be solved by switch detection based on nonlinear estimation. In what follows, we outline the procedure used for generating a generic data segment for regulator $i$. Assume that $y_i|_{[k_0,k_1]}$, $k_0, k_1 \in \{0, \ldots, m-1\}$, is a segment for regulator $i$. We want to test the null hypothesis

$$H_0: y_i|_{[k_0,k_1+1]} \text{ is a segment for regulator } i$$

One way to achieve this, is to use data $y_i|_{[k_0,k_1]}$ for estimating by means of nonlinear least squares (Ljung, 1999) the parameters in (5). From the estimated parameters and the noise variance $\sigma_i^2$, we can compute the $(1-\alpha)$-level confidence interval $I_{k_1+1}$ for the measurement $y_i(k_1 + 1)$ under $H_0$. Then, $H_0$ is accepted if $y_i(k_1 + 1) \in I_{k_1+1}$, whereas a switch is detected otherwise. A variation of this idea is to compute the confidence intervals $I_{k_1+1}$ and $I_{k_1+2}$ (under $H_0$) for $y_i(k_1 + 1)$ and $y_i(k_1 + 2)$, respectively, and to accept the occurrence of a switch if both $y_i(k_1 + 1) \notin I_{k_1+1}$ and $y_i(k_1 + 2) \notin I_{k_1+2}$. On
the basis of extensive simulations we showed that the latter test is more effective for maximizing the length of the segments.

The complete switch detection algorithm is complemented by a procedure for initializing a segment and a backtracking strategy, as described in (Porreca et al., 2006). We remark that initialization and backtracking require the specification of a minimal length $l_{\text{min}} \geq 3$ for segments. Moreover, because of initialization and backtracking, it can happen that some data are not attributed to any segment. This can happen, for instance, when multiple switches are detected in less than $l_{\text{min}}$ consecutive time instants.

In summary, the input for the switch detection algorithm consists of the gene expression data $y_{|[0,m]}$, the noise variances $\sigma_i^2$, the confidence parameter $\alpha \in (0,1)$ and the minimal length $l_{\text{min}}$. The output is, for each regulator $i$, an estimated set of data segments $\hat{Y}_i$. Intuitively, each segment is an estimate of a maximal segment for regulator $i$. Figure 3(a) shows simulated data for the example network of two genes in Fig. 1. In the same figure the real switching times and the segments reconstructed by means of the switch detection procedure are shown. In the sequel, the set $D_i = \bigcup_{P \in \hat{Y}_i} P$ collects all measurements of the $i$th regulator attributed to some segment.

### 3.3 Data classification

The problem of obtaining an estimate of $Y$, the partition consisting of blocks of data segments generated by the same mode of operation of the GRN, is solved in two stages. In the first stage, we search valid partitions of the data segments for each of the regulators, and in the second stage we combine these into partitions of data segments for the entire network.

---

3Note that for estimating parameters in model (5) we need at least 3 consecutive data points.
In order to evaluate whether a proposed partition $\hat{Y}_i$ is acceptable, we need to formulate an appropriate statistical test. Consider the following hypotheses

$$H_0 : \hat{Y}_i \text{ is a valid partition of the data segments}$$

$$H_1 : \{\{P\}, P \in \hat{Y}_i\} \text{ is a valid partition of the data segments}$$

Notice that the partition under $H_1$ is based on the assumption that no two data segments of regulator $i$ can be grouped together, that is, they all originate from a different linear dynamics of the gene. Let $J_0$ be the sum of the square residuals obtained by fitting the model (5) to the data $y_i \in D_i$ under $H_0$, and let $J_1$ be the sum of the square residuals obtained by fitting the same model to the data under $H_1$. According to Rohatgi and Saleh (2000), the Generalized Likelihood Ratio (GLR) $(1 - \bar{\alpha})$-level test$^4$ will reject $H_0$, and the corresponding partition of data segments, if

$$J_0 \geq (1 + \Delta_{\bar{\alpha}}(|\hat{Y}_i|))J_1$$

$$\Delta_{\bar{\alpha}}(|\hat{Y}_i|) = \frac{2(|\hat{Y}_i| - |\hat{Y}_i'|)}{|D_i| - 3|\hat{Y}_i|} F_{\bar{\alpha}}(2(|\hat{Y}_i| - |\hat{Y}_i'|), |D_i| - 3|\hat{Y}_i|)$$

where $|A|$ denotes the cardinality of the finite set $A$ and $F_{\bar{\alpha}}(v_1, v_2)$ represents the $(1 - \bar{\alpha})$th quantile of the $F$ distribution with $(v_1, v_2)$ degrees of freedom.

The statistical test indicates which partitions are supported by the data. We are particularly interested in those valid partitions that are maximal. A valid partition $\hat{Y}_i$ is maximal if every partition $\hat{Y}_i'$ such that $\hat{Y}_i' > \hat{Y}_i$ is not valid.$^5$ Maximal partitions are most informative, as they fulfill the constraints imposed by any subsumed partition.

In principle, one can find maximal valid partitions by generating all partitions

---

$^4$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 $(1 - \bar{\alpha})$ under some approximation (Gallant, 1987).

$^5$Given two partitions $P, Q$, the partial order relation $P \geq Q$ means that for each block $Q \in Q$ there exists a block $P \in P$ such that $Q \subseteq P$. Moreover, $P > Q$ means that $P \geq Q$ and $P \neq Q$. 
and discarding those that are invalid or not maximal. However, this procedure is computationally prohibitive, because the number of partitions grows exponentially with the size of $\hat{Y}_i$. In (Porreca and Ferrari-Trecate, 2007) an efficient algorithm based on a pruning strategy has been proposed for solving the problem without generating all partitions. We emphasize that, because of the statistical nature of the test (6), more than one maximal valid partition can be produced by the algorithm. In this case the remainder of the identification procedure is applied to each of these partitions.

Maximal valid partitions of $\hat{Y}_1, \ldots, \hat{Y}_n$ produced by the above algorithm are estimates of $Y_1, \ldots, Y_n$, respectively. In order to combine these into estimates of $Y$, we need to systematically intersect the time intervals of the data segments of each of the $n$ regulators and determine the mode of operation of the GRN for the resulting intersections. The size of the estimated partition $\hat{Y}$ thus obtained gives the number of different modes of operation of the GRN. The algorithm is rather straightforward, so the details are omitted here and can be found in (Porreca and Ferrari-Trecate, 2007).

In summary, the inputs to the classification algorithm are the data segments $\hat{Y}_i$ and the confidence parameter $\bar{\alpha} \in (0, 1)$. The output consists of different estimations of $Y$. Figure 3(b) illustrates this on simulated data for the example network of two genes.

### 3.4 Threshold reconstruction

The blocks of the partition $\hat{Y}$ consist of disjoint data sets in $\Omega$, each corresponding to a particular mode of operation of the GRN. The basic idea of threshold reconstruction is to find hyperplanes parallel to the linear combination of $n - 1$ axes that separate data in different blocks of $\hat{Y}$. This section follows the approach described in (Drulhe et al., 2008), but generalizes it to the case of noisy data $y(k)$.

We associate to each block $\hat{Y}^j \in \hat{Y}, j = 1, \ldots, |\hat{Y}|$, the data set $F^j = \bigcup_{P \in \hat{Y}^j} P$. 
Moreover, we define $\mathcal{F} = \{F^i, j = 1, \ldots, |\hat{Y}|\}$. Given a pair of distinct sets $F^p$, $F^q$, and a hyperplane $x_i = \hat{\theta}, \hat{\theta} \in \mathbb{R}_{>0}$, we would like to test the following null hypothesis

$$H_0: \text{The hyperplane } x_i = \hat{\theta} \text{ separates } F^p \text{ and } F^q$$

For $j = 1, \ldots, |\mathcal{F}|$, let $m^j_i = \min_{y \in F^j} y_i$ and $M^j_i = \max_{y \in F^j} y_i$. Classical hypothesis testing leads to the following condition: $H_0$ is accepted with confidence level $(1 - \tilde{\alpha})$ if

$$\left( [m^p_i - \hat{\theta} \geq -z_{\tilde{\alpha}} \sigma_i] \wedge [M^q_i - \hat{\theta} \leq z_{\tilde{\alpha}} \sigma_i] \right) \vee \left( [M^p_i - \hat{\theta} \leq z_{\tilde{\alpha}} \sigma_i] \wedge [m^q_i - \hat{\theta} \geq -z_{\tilde{\alpha}} \sigma_i] \right),$$

(8)

with $z_{\tilde{\alpha}}$ such that $\Phi(-z_{\tilde{\alpha}}) = \tilde{\alpha}$, where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function (Rohatgi and Saleh, 2000). Based on (8), one can write a simple condition for the existence of a hyperplane separating $F^p$ and $F^q$ along the $i$th dimension:

$$[m^p_i - M^q_i \geq -2z_{\tilde{\alpha}} \sigma_i] \vee [m^q_i - M^p_i \geq -2z_{\tilde{\alpha}} \sigma_i].$$

(9)

Two hyperplanes along the same dimension are said to be equivalent if they separate the same pairs of data sets in $\mathcal{F}$. Among equivalent hyperplanes only one is optimal in a statistical sense (Vapnik, 1998). This hyperplane, which lies in the middle of the equivalence class, will be called a cut. In general, several cuts will be required to separate all sets in $\mathcal{F}$. This motivates the introduction of multicuts. A multicut $M$ of $\mathcal{F}$ is a finite set of cuts such that for all $F^p, F^q \in \mathcal{F}$ there exists a $\hat{\theta} \in M$, separating $F^p$ and $F^q$. $\mathcal{F}$ is said to be $m$-separable if there exists a multicut of $\mathcal{F}$.

Each cut $x_i = \hat{\theta}$ corresponds to a switching threshold for $x_i$. When the concentration crosses this threshold, the dynamics of the PL system may switch from one mode of operation to another. By extension, a multicut corresponds to a set of switching thresholds. These thresholds have the property that they allow all sets in $\mathcal{F}$, associated
with different modes of operation of the system, to be separated.

In general, the available data are consistent with a large number of multicuts, and thus a large number of PL models of the GRN. A priori there is no reason to prefer one of these models above the others. However, in practice we are most interested in minimal models that account for the available data, that is, models defined by multicuts of minimal cardinality. It is not difficult to compute the set of all possible cuts (Drulhe et al., 2008). In principle minimal multicuts can be found by generating all possible multicuts from this set. The obvious problem with this approach is that it leads to a combinatorial explosion of the number of multicuts. In (Drulhe et al., 2008) we have developed algorithms that compute the minimal multicuts without generating all possible multicuts. The algorithms are based on partial order relations on cuts and multicuts, which among other things allows cuts with a weak separation power to be eliminated beforehand. Moreover, we use a branch-and-bound algorithm to efficiently explore the search space.

In summary, the inputs to the threshold reconstruction algorithm are the partition \( \hat{Y} \) of the time-series data, the noise variances \( \sigma_i^2 \) and the confidence parameter \( \tilde{\alpha} \in (0, 1) \). The output are all minimal multicuts compatible with the data. Figure 4 shows the possible cuts and a minimal multicut generated from simulated data for the model of the two-gene example network.
4 Results

4.1 Application to E. coli carbon starvation response

In order to illustrate the PL identification method, we have generated synthetic gene expression data by means of a model of the GRN controlling the carbon starvation response in the bacterium *Escherichia coli* (Ropers *et al.*, 2006). In the absence of essential carbon sources, an *E. coli* population abandons exponential growth and enters a non-growth state called stationary phase. This growth-phase transition is accompanied by numerous physiological changes in the individual bacteria, concerning among other things cellular morphology and metabolism, as well as the expression of stress response genes. On the molecular level, the carbon starvation response is controlled by a complex GRN, a key part of which is shown in Fig. 5. The PL model in (Ropers *et al.*, 2006), reproduced in the Supplemental Materials of this paper, describes how a carbon stress signal is propagated through the network of global regulators of the bacterium, so as to downregulate the synthesis of stable RNAs and thereby adapt the growth of the cell. In particular, it predicts that in response to a carbon starvation signal, the system switches from a steady state characteristic for exponential growth to another steady state, corresponding to stationary phase. Reentry into exponential phase after a carbon upshift gives rise to a damped oscillation towards the exponential-phase steady state.

The PL model in (Ropers *et al.*, 2006) has been analyzed in a qualitative way, using inequality constraints on the parameters and initial conditions instead of exact numerical values (Batt *et al.*, 2008; de Jong *et al.*, 2004). For the generation of synthetic gene expression data by numerical simulation we need such values though. We have therefore assigned arbitrary, but biologically realistic values to the parameters and
initial conditions that respect the inequality constraints used in the qualitative model (see Supplemental Materials). The data generation process proceeds in two steps. First, given initial concentrations $x(0)$, we numerically compute the solution $x(t)$, $t \in \mathbb{R}_{\geq 0}$, of model (2). Second, we sample $x(t)$ in order to obtain measurements according to model (3)-(4), with a uniform sampling interval $T$ and fixed noise standard deviations $\sigma_i$. Figure 6 shows a data set $y|_{[0,72]}$ obtained by simulating the entry of the bacteria into exponential phase after a carbon upshift, for $T = 10$ min and $\sigma_i = 0.01 \left( \max x_i|_{[0,72]} - \min x_i|_{[0,72]} \right)$. The $\sigma_i$ values yield a signal-to-noise ratio $\text{SNR} = \frac{\max x_i|_{[0,72]} - \min x_i|_{[0,72]}}{\sigma_i} = 100$ for each gene product. We emphasize that sampling densities and noise levels are similar to those characterizing data obtained by means of gene reporter systems.

The PL identification method has been applied to the simulated data set with confidence parameters $\alpha = 0.01$, $\bar{\alpha} = 0.01$, and $\hat{\alpha} = 0.01$ for the switch detection, data classification, and threshold reconstruction steps, respectively. The parameter $l_{\text{min}} = 4$ has been chosen for switch detection. Figures showing the data segments for each of the regulators as well as the partition of the data segments into equivalence classes with the same mode of operation are included in the Supplemental Materials. The output of the method consists of 24 minimal multicuts, each one consisting of 3 cuts. The application of the whole procedure took about 30 seconds on a PC equipped with a 3.2 GHz CPU, 1 GB RAM and running MatLab 7. A representative minimal multicut is

$$\{\hat{\theta}^1_{F_{is}}, \hat{\theta}^2_{F_{is}}, \hat{\theta}^1_{G_{yrb}}\},$$

where the numerical values are given in the following table.

Figure 6
<table>
<thead>
<tr>
<th>Cut</th>
<th>Regulator</th>
<th>Value [M]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\theta}^1_{\text{Fis}}$</td>
<td>Fis</td>
<td>$3.12 \cdot 10^{-7}$</td>
</tr>
<tr>
<td>$\hat{\theta}^2_{\text{Fis}}$</td>
<td>Fis</td>
<td>$7.11 \cdot 10^{-7}$</td>
</tr>
<tr>
<td>$\hat{\theta}^1_{\text{GyrAB}}$</td>
<td>GyrAB</td>
<td>$3.98 \cdot 10^{-8}$</td>
</tr>
</tbody>
</table>

A complete list of all multicuts found is reported in the Supplemental Materials.

Figure 5 projects the minimal multicut (10) on the *E. coli* carbon starvation network. The interactions between the regulators in the network can be reconstructed by verifying which of the estimated values of the matrices $\mu$ and $\nu$ in (5) change when the system switches from one identified mode of operation to another during the simulation. For instance, the value of $\mu_{\text{rrn}}/\nu_{\text{rrn}}$ increases when $x_{\text{Fis}}$ crosses the cut $\hat{\theta}^1_{\text{Fis}}$, thus supporting the inference that $\hat{\theta}^1_{\text{Fis}}$ is associated with the activation of *rrn* expression by Fis.

### 4.2 Performance measures: recall, precision, and F-measures

The previous section illustrated the PL identification method on a synthetic data set with given noise level and sampling density. In order to more rigorously test the performance of the algorithms, we need to assess the quality of the identification results when the data characteristics vary. A prerequisite for this evaluation study is the definition of appropriate performance measures, which compare the identified PL models with the models used to generate the data. In particular, performance criteria need to answer the following questions:

1. Which thresholds are identifiable from the data?

2. Do all cuts found by the method agree with identifiable thresholds? Are all identifiable thresholds recovered by the method?
The answer to the first question establishes a baseline for the performance of the method, in the sense that no method can be expected to do better than reconstruct all and only identifiable thresholds. The second question addresses the capability of the method to minimize the number of false positives and false negatives.

In order to determine the thresholds identifiable from the data, a first approach would be to equate these with the thresholds crossed during the simulation. This is not sufficient for our purpose though, because not all threshold crossings lead to mode changes of the GRN, and only the latter can be recognized from the data. This leads us to define a threshold \( x_i = \theta \) as identifiable from a simulation \( x(t), t \in \mathbb{R}_{\geq 0} \), if there exists \( \bar{t} \in \mathbb{R}_{>0} \) such that \( x_i(\bar{t}) = \theta \) and \( (\mu^{\lambda(x(\bar{t}^-))), \nu^{\lambda(x(\bar{t}^-))}} \neq (\mu^{\lambda(x(\bar{t}^+))), \nu^{\lambda(x(\bar{t}^+))}} \). The set of identifiable thresholds is denoted by \( \Theta_{id} \) and can be easily computed from the model simulations. The set \( \Theta_{id} \) for the simulations of the E. coli carbon starvation network considered in our study can be found in the Supplemental Materials. In particular, only 4 out of 15 thresholds appearing in the model are identifiable.

The PL identification method leads to minimal multicuts, each consisting of a number of cuts. How can we assess if a cut corresponds to some identifiable threshold, that is, if it is correct? The intuition that we followed is to call a cut \( x_i = \hat{\theta} \) correct if an identifiable threshold \( x_i = \theta \) separates exactly the same pairs of data sets in \( \mathcal{F} \) as \( x_i = \hat{\theta} \), taking into account noise in the measurements. In other words, the two are equivalent in the sense of Sec. 3.4. This leads us to introduce the equivalence class
\[ I_{eq}(\hat{\theta}) = [\ell_{\hat{\theta}}, \bar{\epsilon}_{\hat{\theta}}], \]

where

\[ \ell_{\hat{\theta}} = \max \left\{ m_{i}^{j} + 2 z_{\hat{\theta}} \sigma_{i} \leq \hat{\theta}, j = 1, \ldots, |\mathcal{F}| \right\} \cup \{ M_{i}^{j} - 2 z_{\hat{\theta}} \sigma_{i} \leq \hat{\theta}, j = 1, \ldots, |\mathcal{F}| \} , \]

\[ \bar{\epsilon}_{\hat{\theta}} = \min \left\{ m_{i}^{j} + 2 z_{\hat{\theta}} \sigma_{i} \geq \hat{\theta}, j = 1, \ldots, |\mathcal{F}| \right\} \cup \{ M_{i}^{j} - 2 z_{\hat{\theta}} \sigma_{i} \geq \hat{\theta}, j = 1, \ldots, |\mathcal{F}| \} . \]

It is not difficult to verify that every pair of data sets separated by \( x_i = \hat{\theta} \) is also separated by any other threshold \( x_i = \tilde{\theta}, \tilde{\theta} \in I_{eq}(\hat{\theta}) \), and vice-versa. The set of correct cuts for a minimal multicut \( M \) is then defined as

\[ C(M) = \{ \hat{\theta} \in M, \exists \theta \in \Theta_{id} : \theta \in I_{eq}(\hat{\theta}) \} . \quad (11) \]

Note that a correct cut may correspond to several identifiable thresholds but an identifiable threshold can make only one cut correct.

In order to characterize the quality of a minimal multicut, we introduce the notions of recall and precision. These measures are typically used in the field of information retrieval for evaluating the effectiveness of a search procedure (Singhal, 2001). The notions of recall \( \rho(M, \Theta_{id}) \) and precision \( \pi(M, \Theta_{id}) \) are defined in our context as follows:

\[ \rho(M, \Theta_{id}) = \frac{|C(M)|}{|\Theta_{id}|} , \]

\[ \pi(M, \Theta_{id}) = \frac{|C(M)|}{|M|} . \]

Intuitively, recall measures the fraction of identifiable thresholds that is correctly identified, while precision measures the fraction of correct cuts in the minimal multicut. Note that \( M \) is optimal if it is composed of correct cuts only and \( |M| = |\Theta_{id}| \). This
yields values of recall and precision equal to 1. Considering the multicut (10), all three
cuts are correct, yielding recall and precision values equal to 0.75 and 1, respectively.

A simple way to associate to a multicut a scalar measure of performance is to
compute the harmonic mean of recall and precision, which results in the so-called

$$F\text{-measure}$$

$$f(M, \Theta_{id}) = \frac{2\pi(M, \Theta_{id}) \rho(M, \Theta_{id})}{\pi(M, \Theta_{id}) + \rho(M, \Theta_{id})}.$$  \hspace{1cm} (12)

An optimal multicut yields an F-measure equal to 1. The value of the F-measure
characterizing multicut (10) is $6/7 \approx 0.86$.

We emphasize that rather than producing a single minimal multicut, the thresh-
old reconstruction procedure computes the set of all minimal multicuts. We measure
the performance of the method by considering the maximum and the average of the
F-measures associated to the alternative minimal multicuts. In the first case, we eval-
uate the quality of the best hypothesis on the network model provided by the PL
identification method, whereas in the second case we measure the average quality of
the hypotheses. We emphasize that cuts will typically appear in different minimal
multicuts produced by our method. Therefore, it is possible to rank cuts, and the
corresponding regulatory interactions, based on the number of multicuts in which they
occur. Similarly to what is done in the case of Bayesian networks (Pe’er et al., 2001),
this allows us to highlight the reconstructed interactions that are most supported by
the data.

### 4.3 Performance of PL identification method

In this section we study the performance of the identification procedure when tested
in multiple experiments with different noise levels and sampling densities. Data in
each experiment have been sampled from the continuous-time concentrations $x(t)$, $0 \leq$
$t \leq 12\text{ h}$, computed as described in Sec. 4.2. We considered sampling intervals $T \in \{5\text{ min}, 10\text{ min}\}$ and noise levels yielding $\text{SNR} \in \{1000, 200, 100\}$ for each regulator in the network. For each combination of $T$ and $\text{SNR}$ we generated 100 data sets and evaluated the performance of each identification experiment computing the maximum and average F-measure (12). In all experiments we set $\alpha = \bar{\alpha} = \hat{\alpha} = 0.01$ and $l_{\text{min}} = 4$. The results are shown in Fig. 7.

We conclude from the experiments that the identification procedure performs very well even in the least favorable case given by $T = 10\text{ min}$ and $\text{SNR} = 100$. In all cases at least 85 out of 100 data sets yield, for the best hypothesis produced by the method, an F-measure above 0.75. This result should be compared with the relatively low total number of hypotheses produced. In fact, the distribution of the number of minimal multicuts in Fig. 8 reveals that our method produces, for most data sets, less than 10 and 30 minimal multicuts for the shortest and longest sampling interval, respectively. The presence of outliers higher than 100 in the distributions is mainly due to experiments producing multiple partitions of data segments in the classification step. The average F-measure for $T = 10\text{ min}$ falls to about 60% of that obtained with the shortest sampling interval $T = 5\text{ min}$. This can be explained by the higher number of minimal multicuts generated in the former case, since a lower number of data points allows more hypotheses to be consistent with the data. As a consequence, spurious cuts are introduced that lower the precision of the results obtained by the method.

Figure 7

Figure 8
5 Discussion

The structural identification of GRNs is a difficult problem due to its inherent computational complexity and the large amounts of data required. Over the last decade a large number of approaches have been proposed in the literature, reviewed in (Bansal et al., 2007; Cho et al., 2007; de Jong, 2002; Gardner and Faith, 2005; Markowetz and Spang, 2007; van Riel, 2006). Following the classification by Bansal et al. (2007), three families of approaches can be distinguished, each based on a different type of model: correlation or mutual information networks, (dynamic) Bayesian networks, and differential equations. In this paper we have focused on differential equation models, since unlike the other two, they do not only capture the network structure, but also allow the time evolution of the system to be predicted by numerical simulation. This greatly contributes to a better understanding of GRN functioning and enables the design of further experiments to discriminate between alternative models.

Most examples of structural identification using differential equations concern linear models, either globally linear or linearized around a steady state (Chen et al., 1999; D’haeseleer et al., 2000; Gardner et al., 2003; Guthke et al., 2005; Lemeille et al., 2005; van Someren et al., 2000). Whereas the estimation of parameter values of nonlinear models, given a fixed network structure, has been an active research topic in systems biology (e.g., Kuepfer et al., 2007; Quach et al., 2007; Rodriguez-Fernandez et al., 2006; Ronen et al., 2002; van Riel and Sontag, 2006; Zwolak et al., 2005), not much work seems to have been done on nonlinear structural identification. While there have been some approaches based on nonlinear models, their practical applicability is often compromised by the intrinsic mathematical and computational difficulty of the structural identification problem. Not surprisingly, some authors have therefore
focused on specific classes of nonlinear models, with restrictions that reduce the number of parameters and simplify the equations (e.g., Jaeger et al., 2004; Kikuchi et al., 2003; Vilela et al., 2008).

The original contribution of this paper consists in the development of a structural identification method using a class of nonlinear models specifically adapted to GRNs. These so-called PL models consist of a set of decoupled linear models describing the different modes of operation of the GRN and switches between these modes accounting for the nonlinear character of gene regulation (Glass and Kauffman, 1973). PL models thus take a middle ground between the mathematical simplicity of linear models and the biological expressiveness of nonlinear models. The input of the method consists of time-series measurements of concentrations of gene products, while the output is made up of estimates of the modes of operation occurring in the data and all possible minimal combinations of thresholds concentrations accounting for observed switches between modes of operation of the GRN. The PL identification method involves three consecutive steps: switch detection, data classification, and threshold reconstruction. Each step comes with tailored, efficient algorithms and includes appropriate statistical tests to account for the noisy character of the data.

Our approach generalizes and extends the approach described in (Druelle et al., 2006, 2008), in that it makes the threshold reconstruction step robust to noisy data and combines it with switch detection and data classification into an integrated PL identification method. In comparison with the approach suggested in (Westra et al., 2007), and following ideas developed for the identification of hybrid systems (Paoletti et al., 2007), we avoid the transformation of PL identification into a large global optimization problem. Moreover, our method does not need derivatives of observed variables, and produces all network structures consistent with the data instead of a
single solution. We highlight that our method is complementary to the approach presented in (Perkins et al., 2004), in the sense that the latter focuses on the problem of inferring the logic of gene regulation once the switching thresholds are known. The combination of the two methods would be an interesting direction for further research, as it would allow the generation of PL models with step functions (which are easier to interpret biologically) from the flat PL models (2) generated by our method.

The PL identification method infers minimal combinations of switching thresholds from the data, corresponding to different sets of regulatory interactions in the GRN. In order to systematically test the capability of our method to include correct thresholds in the minimal multicuts and avoid spurious thresholds, we have performed simulation experiments with a model of the carbon starvation response in *E. coli*. For sampling intervals up to 10 min and signal-to-noise ratios up to 0.01, the method remains capable of reconstructing the identifiable part of the network from the synthetic data. Although real data of this quantity and quality may be difficult to obtain with DNA microarrays, the data requirements fall within the limits of recent measurement techniques like fluorescent and luminescent reporter gene systems (Ronen et al., 2002). We have assumed in our tests that all concentration variables are observable. If this condition is not satisfied, then spurious switching thresholds may arise, accounting for indirect instead of direct regulatory interactions.

A potential upscaling problem of our method is the generation of all minimal combinations of switching thresholds consistent with the data. The simulation experiments show, however, that the number of minimal combinations of thresholds generated by the method is relatively low, even for the relatively complex network in Fig. 5. Several heuristics that are commonly used for GRN identification can be easily integrated, so as to further improve the performance of the method. For instance, known interac-
tions can be taken into account by imposing certain cuts \textit{a priori}, which will reduce the number of minimal multicuts. Moreover, it will eliminate spurious thresholds and therefore increase the precision of the remaining minimal multicuts. A similar effect is obtained by assuming a sparse network structure (Chen \textit{et al.}, 1999; Yeung \textit{et al.}, 2002). Finally, like for any system identification method, the possibility to generate several time-series data sets displaying a variety of dynamical behaviors of the system, will be critical for the application of the method to larger systems. From this perspective, the growing availability of experimental resources like mutant collections (Baba \textit{et al.}, 2006) and libraries of inducible plasmids (Kitagawa \textit{et al.}, 2005), are extremely important developments for systems biology in general, and reverse engineering efforts in particular.
Funding

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

The authors would like to thank Delphine Ropers for her help with the generation of simulated data from the *E. coli* model.
Author Disclosure Statement

No competing financial interests exist.
References


Rodriguez-Fernandez, M., Mendes, P., and Banga, J.R. 2006. A hybrid approach for


List of Figures

Figure 1
(a) Example of a simple GRN, composed of gene 1 and gene 2, Protein 1 and Protein 2, and their regulatory interactions. (b) PL model of the network in (a). \( x_1, x_2 \geq 0 \) are protein concentrations, \( \kappa_1, \kappa_2 > 0 \) are synthesis parameters, \( \gamma_1, \gamma_2 > 0 \) are degradation parameters, and \( \theta_1^1, \theta_2^1, \theta_1^2, \theta_2^2 > 0 \) are threshold parameters. Protein 1 is synthesized at a rate \( \kappa_1 \), if the concentration of Protein 1 is below its threshold \( \theta_1^1 \) \( (x_1 < \theta_1^1) \) and the concentration of Protein 2 above its threshold \( \theta_2^1 \) \( (x_2 > \theta_2^1) \). The degradation of proteins 1 and 2 occur at a rate proportional to the concentration of the protein itself \( (\gamma_1 x_1, \gamma_2 x_2) \).

Figure 2
(a) Set \( \Omega \) and regulatory domains for the GRN in Fig. 1. It is assumed that the thresholds verify \( \theta_1^2 > \theta_1^1 \) and \( \theta_2^2 > \theta_2^1 \). (b) Rate parameters of the linear model associated with the regulatory domains in (a) are given by the pairs \( (\mu^j, \nu^j) \), \( j = 1, \ldots, 9 \).

Figure 3
(a) Switch detection results on simulated data for the two-gene network in Fig. 1 with \( \kappa_1 = 0.025, \kappa_2 = 0.04, \gamma_1 = 0.01, \gamma_2 = 0.015, \theta_1^1 = \theta_2^1 = 1, \theta_1^2 = \theta_2^2 = 2, x(0) = (0, 0)' \). The standard deviation of the noise is \( \sigma_1 = \sigma_2 = 0.005 \). Vertical lines: detected switching times \( (\alpha = 0.01, l_{min} = 4) \) defining the segments. Crosses on the time axis: real switching times. Data-points represented by small crosses were not attributed to any segment. (b) Classification results for \( \bar{\alpha} = 0.01 \). Same marker: data belonging to the same block in the estimated partition \( \hat{Y} \).
Figure 4
(a) Classified data for the network in Fig. 1. Bold lines: identifiable thresholds. (b) Cuts found using $\tilde{\alpha} = 0.01$. (c) The only minimal multicut found by the threshold reconstruction procedure.

Figure 5
Network of key genes, proteins, and regulatory interactions involved in the GRN controlling the carbon starvation response in *E. coli* (Ropers *et al.*, 2006). Bold and solid lines: identifiable interactions correctly reconstructed by the PL identification method. Bold and dashed line: identifiable interaction not reconstructed. Normal and solid lines: non-identifiable interactions.

Figure 6
Simulated data of the *E. coli* carbon starvation network for $T = 10$ min and $SNR = 100$. The variables $y_{Cya}$, $y_{CRP}$, $y_{Fis}$, $y_{GyrAB}$, $y_{TopA}$, and $y_{rrn}$ denote the concentration measurements [M] of Cya, CRP, Fis, GyrAB, TopA, and stable RNAs, respectively. The simulations have been carried out by setting the carbon starvation signal to 0 (absence of stress signal).

Figure 7
Distribution of the (a)-(b) maximum and (c)-(d) average F-measure for multiple experiments with sampling intervals $T = 5$ min in (a) and (c), and $T = 10$ min in (b) and (d).

Figure 8
Distribution of the number of minimal multicut for multiple experiments with sampling intervals of (a) $T = 5$ min and (b) $T = 10$ min.
\[ \dot{x}_1 = \kappa_1 s^+(x_2, \theta_2^1) s^-(x_1, \theta_1^2) - \gamma_1 x_1 \]
\[ \dot{x}_2 = \kappa_2 s^-(x_1, \theta_1^1) s^-(x_2, \theta_2^2) - \gamma_2 x_2 \]
\[ s^+(x, \theta) = \begin{cases} 1 & \text{if } x > \theta \\ 0 & \text{if } x < \theta \end{cases} \]
\[ s^-(x, \theta) = 1 - s^+(x, \theta) \]

Figure 1: (a) Example of a simple GRN, composed of gene 1 and gene 2, Protein 1 and Protein 2, and their regulatory interactions. (b) PL model of the network in (a). $x_1, x_2 \geq 0$ are protein concentrations, $\kappa_1, \kappa_2 > 0$ are synthesis parameters, $\gamma_1, \gamma_2 > 0$ are degradation parameters, and $\theta_1^1, \theta_1^2, \theta_2^1, \theta_2^2 > 0$ are threshold parameters. Protein 1 is synthesized at a rate $\kappa_1$, if the concentration of Protein 1 is below its threshold $\theta_1^2$ ($x_1 < \theta_1^2$) and the concentration of Protein 2 above its threshold $\theta_2^1$ ($x_2 > \theta_2^1$). The degradation of proteins 1 and 2 occur at a rate proportional to the concentration of the protein itself ($\gamma_1 x_1$, $\gamma_2 x_2$).
Figure 2: (a) Set $\Omega$ and regulatory domains for the GRN in Fig. 1. It is assumed that the thresholds verify $\theta^2_2 > \theta^1_1$ and $\theta^2_1 > \theta^1_2$. (b) Rate parameters of the linear model associated with the regulatory domains in (a) are given by the pairs $(\mu^j, \nu^j)$, $j = 1, \ldots, 9$. 

\[ \begin{align*} 
\mu^1 &= (0, \kappa_2)' \\
\mu^2 &= \mu^3 = \mu^6 = \mu^9 = (0, 0)' \\
\mu^4 &= (\kappa_1, \kappa_2)' \\
\mu^5 &= \mu^7 = \mu^8 = (\kappa_1, 0)' \\
\nu^j &= \text{diag}(\gamma_1, \gamma_2), \ j = 1, \ldots, 9 
\end{align*} \]
Figure 3: (a) Switch detection results on simulated data for the two-gene network in Fig. 1 with $\kappa_1 = 0.025$, $\kappa_2 = 0.04$, $\gamma_1 = 0.01$, $\gamma_2 = 0.015$, $\theta_1^1 = \theta_2^1 = 1$, $\theta_1^2 = \theta_2^2 = 2$, $x(0) = (0, 0)'$. The standard deviation of the noise is $\sigma_1 = \sigma_2 = 0.005$. Vertical lines: detected switching times ($\alpha = 0.01$, $l_{\text{min}} = 4$) defining the segments. Crosses on the time axis: real switching times. Data-points represented by small crosses were not attributed to any segment. (b) Classification results for $\bar{\alpha} = 0.01$. Same marker: data belonging to the same block in the estimated partition $\hat{Y}$. 
Figure 4: (a) Classified data for the network in Fig. 1. Bold lines: identifiable thresholds. (b) Cuts found using $\tilde{\alpha} = 0.01$. (c) The only minimal multicut found by the threshold reconstruction procedure.
Figure 5: Network of key genes, proteins, and regulatory interactions involved in the GRN controlling the carbon starvation response in *E. coli* (Ropers *et al.*, 2006). Bold and solid lines: identifiable interactions correctly reconstructed by the PL identification method. Bold and dashed line: identifiable interaction not reconstructed. Normal and solid lines: non-identifiable interactions.
Figure 6: Simulated data of the *E. coli* carbon starvation network for $T = 10$ min and $SNR = 100$. The variables $y_{Cya}$, $y_{CRP}$, $y_{Fis}$, $y_{GyrAB}$, $y_{TopA}$, and $y_{rrn}$ denote the concentration measurements [M] of Cya, CRP, Fis, GyrAB, TopA, and stable RNAs, respectively. The simulations have been carried out by setting the carbon starvation signal to 0 (absence of stress signal).
Figure 7: Distribution of the (a)-(b) maximum and (c)-(d) average F-measure for multiple experiments with sampling intervals $T = 5$ min in (a) and (c), and $T = 10$ min in (b) and (d).
Figure 8: Distribution of the number of minimal multicuts for multiple experiments with sampling intervals of (a) $T = 5$ min and (b) $T = 10$ min.