Identification of parameters and structure of piecewise affine models of genetic networks

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

∗ Dipartimento di Informatica e Sistemistica - Università degli Studi di Pavia - Via Ferrata 1, 27100 Pavia, Italy
(e-mail: riccardo.porreca,giancarlo.ferrari@unipv.it)
∗∗ INRIA Grenoble-Rhône-Alpes, 655 avenue de l’Europe, Montbonnot, 38334 Saint-Ismier Cedex, France.
(e-mail: Samuel.Drulhe, Hidde.de-Jong@inria.fr)

Abstract: In this paper we consider piecewise affine models of genetic regulatory networks proposed by Glass and Kauffman in the 70’s and discuss a method for their identification. These systems have specific features that must be preserved by the identification procedure in order to obtain biologically meaningful models. Moreover, rather than producing a single hypothesis about the way genes interact, identification should produce alternatives consistent with the available data so as to provide biologists with multiple hypothesis about the network functioning. The input of the identification method consists of time-series measurements of concentrations of gene products. As outputs, 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 are provided. Individual steps of the identification method have been described separately in previous publications and the main aim of this paper is to test the applicability of the whole procedure. To this purpose, we use simulated data obtained from a model of the carbon starvation response in the bacterium E. coli. In particular, performance of the method under different data characteristics, notably variations in the noise level and the sampling density, are discussed.

Keywords: Genetic Networks, Reverse Engineering, Systems Biology, Hybrid System Identification.

1. INTRODUCTION

Genetic Regulatory Networks (GRNs) govern many cellular processes like the response of cells to environmental perturbations, and hence play a key role in cell functioning. The fundamental units of GRNs are genes that typically code for a protein and/or RNA, briefly referred to as gene products. Synthesis of a product amounts to the expression of the corresponding gene, which can be activated or inhibited by the presence in the cell of (other) gene products at different concentration levels. This leads to possibly complex regulatory interactions between genes and their products, the latter playing the role of regulators.

The problem of reverse engineering GRNs from experimental data has been an active research field in systems biology over the last years [Bansal et al., 2007, Cho et al., 2007, Gardner and Faith, 2005, Markowetz and Spang, 2007, van Riel, 2006]. This growing interest is motivated by the availability of experimental techniques, such as gene reporter systems [Zaslaver et al., 2006], that allow one to measure gene expression at a sampling rate sufficiently high for capturing the dynamics of regulatory interactions with good accuracy. Among the modeling formalisms for GRNs proposed in the literature, PieceWise Affine (PWA) systems, introduced by Glass and Kauffman [1973], strike a compromise between simple linear and complex nonlinear models, and hence are attractive from the identification viewpoint. Locally, PWA models have a special 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 studies of regulatory networks of biological interest [e.g., Ghosh and Tomlin, 2004, Ropers et al., 2006].

PWA models of GRNs have specific features. First, the space of product concentrations is partitioned into hyper-rectangular regions by threshold concentrations at which a product can switch on/off genes in the network. Second, affine ODEs describing the evolution of products in each region are decoupled. As explained in [Ferrari-Trecate, 2007, Drulhe et al., 2008], most techniques for the identification of hybrid systems [Paoletti et al., 2007] fail to take into account these features and hence produce models that have little biological meaning. Moreover, since in biological experiments just a fraction of modes is usually sampled, different combinations of thresholds may be consistent with the data. In this case, it is preferable to produce all different models that fit the data with the goal of providing biologists with multiple hypothesis about the
network functioning. General-purpose techniques for the identification of hybrid systems are unsatisfactory from this perspective, as they produce a single model.

For these reasons, we propose a gray-box procedure for reconstructing PWA models of GRNs that is structured in three steps: switch detection, data classification and threshold reconstruction. Previous contributions by the authors of the present paper focused on statistically-grounded methods for solving each step separately [Porreca et al., 2007, Druhlhe et al., 2008, Porreca and Ferrari-Trecate, 2008, Porreca and Ferrari-Trecate, 2007]. The main purpose of this paper is to test the applicability of the whole identification procedure. Moreover, the algorithm in Druhlhe et al. [2008] for computing sets of thresholds assumed noiseless measurements. Therefore, we present a generalization of this procedure to the case of noisy data.

The paper is structured as follows: PW A models of GRNs are briefly recalled in Section 2. Section 3 summarizes the generalization of this procedure to the case of noisy data. The authors of the present paper focused on statistically-grounded methods for solving each step separately [Porreca et al., 2007, Drulhe et al., 2008].

2. PIECEWISE AFFINE MODELS OF GENETIC NETWORKS

Here we resume the main features of PWA models of GRNs deferring the interested reader to [de Jong et al., 2004] for a more detailed description. We denote by \( x = (x_1, \ldots, x_n)^T \in \Omega \) a vector of concentrations of products, where \( \Omega \subset \mathbb{R}_+^n \) is a bounded hyperrectangle that includes the origin and \( n \) is the number of genes composing the network. To each concentration \( x_i \), \( i \in \{1, \ldots, n\} \), we associate a set of strictly positive threshold concentrations \( \{\theta_i^{(1)}, \ldots, \theta_i^{(p_i)}\} \), \( p_i \geq 0 \). At its threshold concentrations a product 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

\[
\Theta = \bigcup_{i\in\{1,\ldots,n\},\ell\in\{1,\ldots,p_i\}} \{x \in \Omega, x_i = \theta_i^{(\ell)}\},
\]

splits \( \Omega \) into open hyperrectangular regions \( D^j, j \in \{1, \ldots, s\} \), called regulatory domains. The PWA dynamics governing the evolution of products is given by

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

where \( \mu^j \) is a vector of \( n \) positive synthesis parameters, \( \nu^j = \text{diag}(\nu_1^j, \ldots, \nu_n^j) \) a diagonal matrix of strictly positive degradation parameters, and \( \lambda \) a switching function, such that \( \lambda(x) = j \) iff \( x \in D^j \). Note that \( \nu^j \) is a diagonal matrix and hence (2) is a system of decoupled differential equations. The pairs \((\mu^j, \nu^j), j = 1, \ldots, s\), represent the modes of operation of the network.

We assume that gene expression data are 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 and that at most one switch can occur within a sampling interval.

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

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

\[
\phi(\mu^j_i, \nu^j_i, x_i(k), T(k)) = \frac{\mu^j_i}{\nu^j_i} x_i(k) - x_i(k) e^{-\nu^j_i T(k)}
\]

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

where \( \xi_i(\cdot) \) is a white Gaussian noise with zero mean and variance \( \sigma^2_i \), and \( y_i(k) \) are the measurements for the \( i \)th regulator. We define \( y(k) = [y_1(k), \ldots, y_n(k)]^T \).

Letting \( \lambda_k = \lambda(x(k)) \) we say that \( k \) is a switching time [resp. switching time for the \( i \)th product] if \( (\mu^{\lambda_k}, \nu^{\lambda_k}) \neq (\mu^{\lambda_{k+1}}, \nu^{\lambda_{k+1}}) \) [resp. \( (\mu^{\lambda_k}_i, \nu^{\lambda_k}_i) \neq (\mu^{\lambda_{k+1}}_i, \nu^{\lambda_{k+1}}_i) \)].

Consecutive switching times for the \( i \)th product can be used to define consecutive segments (for the \( i \)th product), by collecting data generated by the same synthesis and degradation parameters. Similarly, consecutive switching times define a set \( Y = \{\tau_1, \ldots, \tau_N\} \) of consecutive segments \( \tau_j \), by collecting data generated by the same mode of operation of the GRN.

Notice that two different segments may originate from the same mode of operation of the GRN. This happens when the sequence \( x(k) \) enters a regulatory domain, leaves it and re-enters the same domain at a later time. Such segments are termed equivalent and the corresponding equivalence relations will be denoted with \( \sim \). The partitions of \( Y \) induced by \( \sim \) is denoted by \( \mathcal{Y} = Y/\sim \).

3. THE IDENTIFICATION PROBLEM: DEFINITION AND MAIN CONTRIBUTIONS

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.

The gray-box procedure for reconstructing PWA models of GRNs we propose is structured in the following steps.

S1) **Switch detection.** Estimate the data segments of each product in the GRN.

S2) **Data classification.** From data segments estimated in step (S1), provide an estimate \( \hat{\mathcal{Y}} \) of the partition \( \mathcal{Y} \). This yields the rate parameters defining the modes of operation as a by-product.

S3) **Threshold reconstruction.** Using \( \hat{\mathcal{Y}} \), estimate possible thresholds on concentration variables and find all minimal combinations of thresholds that account for observed changes in the mode of operation. These sets of thresholds will be termed *multicuts.*

Methods for solving each step separately have been proposed in [Porreca et al., 2007, Druhlhe et al., 2008, Porreca et al., 2008].
and Ferrari-Trecate, 2008, Porreca and Ferrari-Trecate, 2007. However, the algorithm in Druhlle et al. [2008] for reconstructing thresholds assumed noiseless output measurements. The next section describes a generalization to the noisy case.

3.1 Threshold reconstruction

The blocks of the partition \( \tilde{Y} \) are disjoint data sets, 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 \( \tilde{Y} \).

We associate to each block \( \tilde{Y}^j \subseteq \tilde{Y}, j = 1, \ldots, |\tilde{Y}| \), the data set \( F^j = \cup_{\tau \in \tilde{Y}^j} \tau \) and define \( F = \{ F^j, j = 1, \ldots, |\tilde{Y}| \} \). Given a pair of distinct sets \( F^p, F^q \), and a hyperplane \( x_i = \theta, \theta \in \mathbb{R}_{>0} \), we would like to test the following null hypothesis

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

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

\[
\left( [m^p_j - \theta \geq -z_\sigma \sigma_j] \land [M^q_j - \theta \leq z_\sigma \sigma_j] \right) \lor \left( [M^p_j - \theta \leq z_\sigma \sigma_j] \land [m^q_j - \theta \geq -z_\sigma \sigma_j] \right),
\]

(5)

with \( z_\sigma \) such that \( \Phi(-z_\sigma) = \alpha \), where \( \Phi(\cdot) \) denotes the standard normal cumulative distribution function. Based on (5), a hyperplane separating \( F^p \) and \( F^q \) along the \( i \)-th dimension exists if

\[
[m^p_j - M^q_j \geq 2z_\sigma \sigma_j] \lor [m^q_j - M^p_j \geq -2z_\sigma \sigma_j] .
\]

(6)

Two hyperplanes along the same dimension are said to be equivalent if they separate the same pairs of data sets in \( F \). Among equivalent hyperplanes, the one 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 \( F \). This motivates the introduction of multicuts. A multicut \( M \) of \( F \) is a finite set of cuts such that for all \( F^p, F^q \in F \) there exists a \( \hat{\theta} \in M \), separating \( F^p \) and \( F^q \). \( F \) is said to be \( m \)-separable if there exists a multicut of \( F \).

Each cut \( x_i = \hat{\theta} \) corresponds to a threshold for \( x_i \). When the concentration \( x_i(t) \) 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 thresholds that allow all sets in \( F \) to be separated.

In general, the available data are consistent with multiple multicuts, and thus different PWA models of the GRN. 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 [Druhlle et al., 2008]. This set is given as input to the methods developed in [Druhlle et al., 2008] that compute all minimal multicuts without generating all possible multicuts. The algorithms are based on partial order relations on cuts and multicuts and use a branch-and-bound algorithm to efficiently explore the search space.

4. PERFORMANCE OF THE PWA IDENTIFICATION PROCEDURE

In order to evaluate the performance of the identification procedure, we have generated synthetic gene expression data by means of a PWA 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 undergoes exponential growth and enters a non-growing state called stationary phase. On the molecular level this growth-phase transition is controlled by a complex GRN, a key part of which is shown in Fig. 1. Using a PWA model of the GRN in Fig. 1, we systematically tested the identification method under variations in the noise level and the sampling density.

Particular care must be taken in defining suitable measures for assessing the quality of the reconstructed thresholds. To this purpose we call a threshold \( x_i = \theta \) identifiable from a simulation \( x(t), t \in \mathbb{R}_{>0} \), if there exists \( t \in \mathbb{R}_{>0} \) such that \( x_i(t) = \theta \) and \( (\mu^l(\theta(t)), \mu^u(\theta(t))) \neq (\mu^l(\hat{\theta}(t)), \mu^u(\hat{\theta}(t))) \). The set of identifiable thresholds is denoted by \( \Theta_{id} \) and can be easily computed from the model simulation. Next, we describe how to compute the set \( C(M) \) of thresholds in the multicut \( M \) corresponding to identifiable thresholds.

We call a cut \( x_i = \hat{\theta} \) correct if an identifiable threshold \( x_i = \hat{\theta} \) separates exactly the same pairs of data sets in \( F \) as \( x_i = \theta \), taking into account noise in the measurements. In other words, the two are equivalent in the sense of Sec. 3.1. This leads us to introduce the equivalence class

\[ I_{eq}(\theta) = \{ \hat{\theta}, \epsilon_\theta \}, \]

where

\[ \epsilon_\theta = \max \{ m^q_j + 2z_\sigma \sigma_j \leq \hat{\theta}, j = 1, \ldots, |F| \} \]

\[ \cup \{ M^p_j - 2z_\sigma \sigma_j \leq \hat{\theta}, j = 1, \ldots, |F| \}, \]

\[ \epsilon_\theta = \min \{ m^q_j + 2z_\sigma \sigma_j \geq \hat{\theta}, j = 1, \ldots, |F| \} \]

\[ \cup \{ M^p_j - 2z_\sigma \sigma_j \geq \hat{\theta}, j = 1, \ldots, |F| \} . \]

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 \( x_i = \theta, \theta \in I_{eq}(\hat{\theta}) \), and vice-versa. The set of correct cuts for a minimal multicut \( M \) is then defined as

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

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:

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

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

where \(|\mathcal{A}|\) denotes the cardinality of the set \( \mathcal{A} \). Intuitively, recall measures the fraction of identifiable thresholds that is correctly identified, while precision measures the fraction of correct cuts in the 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.

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

\[
  f(M, \Theta_{id}) = \frac{2\pi(M, \Theta_{id}) \rho(M, \Theta_{id})}{\pi(M, \Theta_{id}) + \rho(M, \Theta_{id})},
\]

(8)

An optimal multicut yields an F-measure equal to 1.

As an example of the application of the proposed identification procedure, we used the data set shown in Fig. 2. Defining the Signal-To-Noise ratio (SNR) as \(\frac{\sigma_i}{\bar{x}_i}\), where \(\sigma_i\) is the sample standard deviation of \(x_i(\cdot)\), the standard deviations \(\sigma_i\) have been chosen in order to obtain an SNR equal to 30 for each product. In the same figure, estimated and true segments for each regulator are shown.

Only four thresholds are identifiable from this simulation and they correspond to the identifiable interactions presented in Fig. 1. At the end of the identification procedure we obtained 30 multicut, each one consisting of three cuts. The best multicut corresponds to the three identifiable interaction shown with bold and solid lines in Fig. 1 and its F-measure is 6/7.

We now study the performance of the identification procedure when tested in multiple experiments with different noise levels and sampling intervals. Data in each experiment have been sampled from the continuous-time concentration of representative RNAs and sampling intervals. Data in each experiment when tested in multiple experiments with different parameters (see Porreca et al. [2007], Drulhe et al. [2008], Porreca and Ferrari-Trecate. [2008], Porreca and Ferrari-Trecate [2007]) that have been kept constant over the experiments. In particular, the confidence levels (such as \(1 - \alpha\) in Section 3.1) of the statistical tests used in all steps have been set to 0.99. The results are shown in Fig. 3.

We conclude from the experiments that the identification procedure performs very well for SNRs 40 and 30. In these cases at least 95 out of 100 data sets yield, for the best hypothesis produced by the method, an F-measure above 0.75. In the worst case, given by SNR=20 and \(T = 5\), 69 experiments gave an F-measure above 0.75. In particular, the bars close to zero in Fig. 3(a) and (c) represent cases where the set \(\mathcal{F}\) was not m-separable. Note that the shorter sampling periods the points before and after a switch are closer. Hence, it is more likely that m-separability breaks down because of noise for \(T = 5\) rather than for \(T = 10\), as confirmed by Fig. 3.

The total number of hypotheses produced is relatively low (see Fig. 4) since our method produces, for most data sets, less than 10 and 30 minimal multicut for the shortest and longest sampling interval, respectively.

The average F-measures for \(T = 10\) min are lower than the ones obtained with the shortest sampling interval \(T = 5\) min. This can be explained by the higher number of minimal multicut 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.

5. CONCLUSIONS

We presented an experimental evaluation of a method developed for the identification of GRNs models in the PWA form. In particular, we generalized the methods described in [Drulhe et al., 2008], in order to make the threshold reconstruction step robust to noise. In comparison with the approach suggested in [Westra 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.

The PWA identification method infers minimal combinations of 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 multicut and avoid spurious thresholds, we have introduced suitable performance measures and conducted 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 30, 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 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 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. 1. Several heuristics that are commonly used for GRN identification can be
Fig. 4. Distribution of the number of minimal multicuts for multiple experiments with sampling intervals of (a) $T = 5$ min and (b) $T = 10$ min.

easily integrated, so as to further improve the performance of the method. For instance, known interactions can be taken into account by imposing certain cuts 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.

**ACKNOWLEDGEMENTS**

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

**REFERENCES**


