ISTITUTO DI ANALISI DEI SISTEMI ED INFORMATICA
"Antonio Ruberti"
CONSIGLIO NAZIONALE DELLE RICERCHE

L. Farina, A. Germani, G. Mavelli, P. Palumbo

IDENTIFICATION OF REGULATORY NETWORK MOTIFS FROM GENE EXPRESSION DATA

R. 08-07, 2008

Lorenzo Farina – Dipartimento di Informatica e Sistemistica, Universitá "La Sapienza", Roma, Italy. Email: bertola@iasi.rm.cnr.it.

Alfredo Germani – Dipartimento di Ingegneria Elettrica e dell'Informazione, Università degli Studi dell'Aquila, Poggio di Roio, 67040 L'Aquila, Italy, germani@ing.univaq.it.

Gabriella Mavelli – Istituto di Analisi dei Sistemi ed Informatica del CNR, viale Manzoni 30 - 00185 Roma, Italy. Email: gabriella.mavelli@iasi.rm.cnr.it.

Pasquale Palumbo – Istituto di Analisi dei Sistemi ed Informatica del CNR, viale Manzoni 30 - 00185 Roma, Italy. Email: pasquale.palumbo@iasi.rm.cnr.it.

ISSN: 1128–3378

Collana dei Rapporti dell'Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", CNR

viale Manzoni 30, 00185 ROMA, Italy

tel. ++39-06-77161
fax ++39-06-7716461
email: iasi@iasi.rm.cnr.it
URL: http://www.iasi.rm.cnr.it

Abstract

The modern systems biology approach to the study of molecular cellular biology, consists in the development of computational tools to support the formulation of new hypotheses on the molecular mechanisms underlying the observed cell behavior. Recent biotechnologies are able to provide precise measures of gene expression time courses in response to a large variety of internal and environmental perturbations. In this paper, we propose a simple algorithm for the selection of the "best" regulatory network motif among a number of alternatives, using the expression time course of the genes which are the final targets of the activated signalling pathway. To this aim, we considered the Hill nonlinear ODEs model to simulate the behavior of two ubiquitous motifs: the *single input motif* and the *multi output feed-forward loop motif*. Our algorithm has been tested on simulated noisy data assuming the presence of a step-wise regulatory signal. The results clearly show that our method is potentially able to robustly discriminate between alternative motifs, thus providing a useful *in silico* identification tool for the experimenter.

1. INTRODUCTION

The cell is the basic unit of life able to display a wide range of autonomous functions in a flexible and robust way [18, 9]. The genome of a cell is a key player for the control and coordination of its functions and processes. It actively contributes to the modulation of the timing and amount of proteins in response to internal and environmental signals. Such appropriate behavior is the result of many factors acting on different hierarchical levels and time scales, including the regulation exerted by a family of proteins, called *transcription factors*, able to selectively bind to DNA regions and activate or repress the expression of downstream target genes in a coordinated fashion. Gene expression is to a large extent regulated at the level of mRNA abundance which is the result of two basic biological mechanisms: mRNA synthesis and degradation. The synthesis of new mRNA, called *transcription*, is mainly driven by transcription factors able to positively or negatively interfere with RNA polymerase, whereas mRNA degradation is mainly exerted by ribonucleases, enzymes able to catalyze the hydrolysis of RNA. It is worth noting that recent experimental observations have revealed new degradation pathways involving small RNAs, thus showing the very complex nature of the degradation process and its regulation [7, 5, 15].

One of the main goals of systems biology is to provide computational support for the formulation of new biological hypotheses on the biochemical mechanisms underlying the observed cell behavior and their experimental validation [8]. In fact, the integration of computational modeling, system analysis and quantitative experiments, has proved to be very successful in providing the field of molecular biology with new paradigms and insight [19, 2, 6]. The systems biology approach to the study of the gene regulatory system in living cells has suffered so far by the lack of time series experimental data resulting from perturbation experiments. In fact, the large majority of the available datasets, are steady state measurements or very short time series. However, recently, an experimental technique able to continuously measure transcriptional activity in *E. coli* at promoters using a Green Fluorescent Protein (GFP) as a reporter gene, has been developed by the Uri Alon group [20], thus providing high resolution time courses in living bacterial cells and opening the possibility of validating model-based computational methods on a sound and reliable experimental ground [14].

A major biological feature of gene networks is the presence of *network motifs* [1, 11], *i.e.* small subset of basic building-block regulatory "circuits" able to perform a large variety of biological functions alone or in combinations with other circuits, depending on the specific environmental and internal conditions. It is important to note that such biological circuits are inherently different from electronic ones in that, their regulatory modules function and structure, can be re-arranged in response to the specific task to be performed or to the environmental condition to be dealt with [12]. In fact, the cell must rewire the interconnection of modules in order to efficiently and robustly perform the large number of biological tasks to be accomplished during its life, as for metabolic shifts, reproductive cycles, development, growth, starvation, damage repairs and so on. The most frequently encountered network motifs, in many organisms, are the *single input module* (SIM) [11] and the *feed-forward loop* (FFL) [1]. The SIM motif (shown in Figure 1) is composed of a single (often self-repressing) transcription factor able to activate or de-repress the expression of a number of target genes. An example is the regulatory circuits driving the response induced by the presence of a damage in DNA strands (SOS system) in *E. coli*, [10]. The FFL motif is composed of a transcription factor X that regulates another transcription factor Y and both X and Y regulates a gene Z or a number of target genes Z_i (*multi output feed-forward loop motif*, MO-FFL) and it is shown in Figure 2. Examples of FFLs can be found in the regulatory circuit of the *lac* operon [13].

In this paper we address the problem of finding the "best" network motif, among a number of alternatives, able to explain the observed gene expression data resulting from a perturbation experiment on a population of cells. We will focus on a realistic situation in which the experimenter, after an environmental perturbation, *e.g.* adding a specific carbon source to the culture medium, is able to determine the genes that significantly change their expression with respect to the basal level and continously measures the resulting time course using a GFP based approach. Once data are available, we suppose that the experimenter wants to know whether the measured expression time course is consistent with the presence of a single master regulator X directly acting on its target genes Z_i (SIM motif) or with the combinatorial action in cooperation with another transcription factor Y (MO-FFL motif). The information about the most likely regulatory motif is very important, since it allows the design of new appropriate experiments for the detection of such regulators by using, for example, chromatin immuno-precipitation techniques (ChiP-on-chip).

In this paper we modeled gene regulatory circuits by means of standard nonlinear differential equations [4, 17]. The unknown biological parameters of the assumed models are estimated, from target gene expression data, using the Matlab identification tool able to robustly identify such parameters and to recognize the presence of a second unknown regulatory gene, thus discriminating between a SIM and a MO-FFL motif. Moreover, we also considered in our models the possibility that gene expression regulation might be also exerted at the post-transcriptional level by variations in transcripts turnover. In the sequel, we will consider a SIM motif with 4 nodes (X, Z_1, Z_2, Z_3)) a MO-FFL motif with 5 nodes (X, Y, Z_1, Z_2, Z_3)) and denote them as *motif A* (Fig. 1) and *motif B* (Fig. 2), respectively. Furthermore, gene X is assumed be part of a negative autoregulation loop.

The paper is organized as follows. In section II we will provide model equations of the two examined motifs and formulate the setting of the problem. Section III deals with the identification algorithm and simulations results are described and illustrated in section IV.

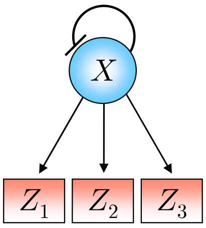

Figure 1: SIM motif with a negative autoregulation loop for gene X.

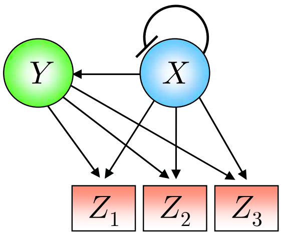

Figure 2: MO-FFL motif with a negative autoregulation loop for gene X.

2. Model equations and problem setting

In this section we will use nonlinear ordinary differential equations (ODEs) to analyze the gene network motifs and to recover *in silico* the observed time series data, *i.e.* by means of computer simulations (see [1], [4], [6], [17]). Consider the case of an N-genes transcription network, where the gene X_i is regulated by the genes X_j, $j \in A_i \subseteq \{1, \ldots, N\}$, acting as activators and by the genes X_k, $k \in R_i \subseteq \{1, \ldots, N\}$, acting as repressors. The gene X_i product concentration will be referred to as $x_i(t)$ and it is the i-th state variable of the system:

$$\begin{aligned}\frac{dx_i(t)}{dt} &= V_i^0 \prod_{j \in A_i} \left(1 + (\bar{V}_{ij} - 1) \frac{x_j^{\nu_{ij}}(t)}{x_j^{\nu_{ij}}(t) + \theta_{ij}^{\nu_{ij}}} \right) \\ &\cdot \prod_{k \in R_i} \frac{\theta_{ik}^{\nu_{ik}}}{x_k^{\nu_{ik}}(t) + \theta_{ik}^{\nu_{ik}}} - \lambda_i x_i(t); \qquad (1)\end{aligned}$$

V_i^0 is the basal transcription rate of gene X_i, $\bar{V}_{ij}$ is the V_i^0 multiplicative factor giving the maximal transcription rate exerted by X_i when the activator gene product x_j saturates the X_i activation sites, θ_{ij} (θ_{ik}) is the activation (repression) threshold value for $x_j(t)$ ($x_k(t)$), λ_i accounts for a constant degradation rate of $x_i(t)$, and ν_{ij} (ν_{ik}) is the activator (repressor) Hill coefficient. The activation (repressor) exponent ν_{ij} (ν_{ik}) is determined by the number of copies of the binding sites for $x_j(t)$ ($x_k(t)$) in the $X_i(t)$ promoter.

Remark. It has to be stressed that, in an experimental framework, data concentrations are usually acquired as ratios with respect to their basal level. Denoting such ratios as $x_{ir}(t) = x_i(t)/x_{ib}$, with x_{ib} being the basal concentration of gene X_i product, equations (1) become:

$$\begin{aligned}\frac{dx_{ir}(t)}{dt} &= V_{ir}^0 \prod_{j \in A_i} \left(1 + \left(\bar{V}_{ij} - 1\right) \frac{x_{jr}^{\nu_{ij}}(t)}{x_{jr}^{\nu_{ij}}(t) + \theta_{ij,r}^{\nu_{ij}}}\right) \\ &\quad \cdot \prod_{k \in R_i} \frac{\theta_{ik,r}^{\nu_{ik}}}{x_{kr}^{\nu_{ik}}(t) + \theta_{ik,r}^{\nu_{ik}}} - \lambda_i x_{ir}(t), \qquad (2)\end{aligned}$$

with:

$$V_{ir}^0 = \frac{V_i^0}{x_{ib}}, \qquad \theta_{ij,r} = \frac{\theta_{ij}}{x_{jb}}, \qquad \theta_{ik,r} = \frac{\theta_{ik}}{x_{kb}}. \qquad (3)$$

In this case, according to the usual experimental framework consisting of a perturbation of the basal steady state value, conditions at time zero for (2) are simply $x_{ir}(0) = 1$, $i = 1, \ldots, N$. In the following the subindex "r" will be omitted thus formally writing (2) as the absolute concentration equations (1).

As previously stated in the Introduction, two different gene motifs will be investigated (motif A and B), with the master gene X activated by an external signal $u(t)$ (*e.g.* a generic perturbation such as that produced by a nutrient added to the culture medium or by an heat shock). Then, by using (1), the following ODE models are given:

Motif A:

$$\begin{aligned}\frac{dx(t)}{dt} &= V_x^0 \frac{\theta_{xx}^{\nu_{xx}}}{x^{\nu_{xx}}(t) + \theta_{xx}^{\nu_{xx}}} - \lambda_x x(t) + \alpha u(t), && (4)\\ x(0) &= 1, && (5)\\ \frac{dz_i(t)}{dt} &= V_{z_i}^0 \left(1 + \left(\bar{V}_{z_i x} - 1\right) \frac{x^{\nu_{z_i x}}(t)}{x^{\nu_{z_i x}}(t) + \theta_{z_i x}^{\nu_{z_i x}}}\right) \\ &\quad -\lambda_{z_i} z_i(t), && (6)\\ z_i(0) &= 1, \qquad\qquad i = 1, 2, 3. && (7)\end{aligned}$$

Motif B:

$$\begin{aligned}\frac{dx(t)}{dt} &= V_x^0 \frac{\theta_{xx}^{\nu_{xx}}}{x^{\nu_{xx}}(t) + \theta_{xx}^{\nu_{xx}}} - \lambda_x x(t) + \alpha u(t), && (8)\\ x(0) &= 1, && (9)\\ \frac{dy(t)}{dt} &= V_y^0 \left(1 + \left(\bar{V}_{yx} - 1\right) \frac{x^{\nu_{yx}}(t)}{x^{\nu_{yx}}(t) + \theta_{yx}^{\nu_{yx}}}\right) \\ &\quad -\lambda_y y(t) && (10)\\ y(0) &= 1, && (11)\\ \frac{dz_i(t)}{dt} &= V_{z_i}^0 \left(1 + \left(\bar{V}_{z_i x} - 1\right) \frac{x^{\nu_{z_i x}}(t)}{x^{\nu_{z_i x}}(t) + \theta_{z_i x}^{\nu_{z_i x}}}\right) \\ &\quad \cdot \left(1 + \left(\bar{V}_{z_i y} - 1\right) \frac{y^{\nu_{z_i y}}(t)}{y^{\nu_{z_i y}}(t) + \theta_{z_i y}^{\nu_{z_i y}}}\right) - \lambda_{z_i} z_i(t), && (12)\\ z_i(0) &= 1, \qquad\qquad i = 1, 2, 3. && (13)\end{aligned}$$

For each motif A and B, two different biological hypotheses are considered: the *operon* case and the general *non-operon* case. It is well known that an operon is a group of adjacent genes (3 genes in our case) sharing the same promoter region. This case corresponds to the following conditions on the parameters of the motif A model (operon-motif A scheme):

$$\theta_{z_1x} = \theta_{z_2x} = \theta_{z_3x} = \theta_{zx}, \tag{14}$$

$$\nu_{z_1x} = \nu_{z_2x} = \nu_{z_3x} = \nu_{zx}, \tag{15}$$

$$\bar{V}_{z_1x} = \bar{V}_{z_2x} = \bar{V}_{z_3x} = \bar{V}_{zx}, \tag{16}$$

and to (14)-(16), together with the following:

$$\theta_{z_1y} = \theta_{z_2y} = \theta_{z_3y} = \theta_{zy}, \tag{17}$$

$$\nu_{z_1y} = \nu_{z_2y} = \nu_{z_3y} = \nu_{zy}, \tag{18}$$

$$\bar{V}_{z_1y} = \bar{V}_{z_2y} = \bar{V}_{z_3y} = \bar{V}_{zy}, \tag{19}$$

for the motif B model (operon-motif B scheme). In the second case no *a* priori constraint has been considered for the model parameters (non operon-motif A and non operon-motif B schemes).

3. Identification of model parameters

The aim of the paper is to provide an identification algorithm able to identify which one of the four biological schemes (namely motif A, operon/non operon case; motif B, operon/non operon case), is the one best fitting a set of experimental data regarding the expression time course of target genes Z_1, Z_2, Z_3. Moreover, the proposed algorithm also allows the estimation of the model unknown biological parameters. A piecewise constant input signal $u(t)$ is considered, assuming the presence of only two possible normalized values 1 and 0 (the effective strength of the signal is given by parameter α). From a computational point of view, some parameters do not need to be estimated, because they are related to the others by algebraic constraints. For instance, at steady state, with no input signal $u(t) \equiv 0$, eq.(4) as well as eq.(8) provide the following relationship:

$$V_x^0 \frac{\theta_{xx}^{\nu_{xx}}}{1+\theta_{xx}^{\nu_{xx}}} = \lambda_x \qquad \Longrightarrow \qquad V_x^0 = \frac{\lambda_x(1+\theta_{xx}^{\nu_{xx}})}{\theta_{xx}^{\nu_{xx}}} \tag{20}$$

that means there are only 4 independent parameters (θ_{xx}, ν_{xx}, λ_x, α) to be estimated. The same may be found for eq.s(6), Motif A:

$$\frac{V_{z_i}^0 \left(\theta_{z_ix}^{\nu_{z_ix}} + \bar{V}_{z_ix}\right)}{1+\theta_{z_ix}^{\nu_{z_ix}}} = \lambda_{z_i}, \tag{21}$$

from which an explicit relationship comes out for $V_{z_i}^0$:

$$V_{z_i}^0 = V_{z_i}^0(\theta_{z_ix}, \nu_{z_ix}, \bar{V}_{z_ix}, \lambda_{z_i}),$$

and for eq.(10), eq.s(12) of Motif B:

$$\frac{V_y^0 \left(\theta_{yx}^{\nu_{yx}} + \bar{V}_{yx}\right)}{1+\theta_{yx}^{\nu_{yx}}} = \lambda_y, \tag{22}$$

$$\frac{V_{z_i}^0 \left(\theta_{z_ix}^{\nu_{z_ix}} + \bar{V}_{z_ix}\right)\left(\theta_{z_iy}^{\nu_{z_iy}} + \bar{V}_{z_iy}\right)}{\left(1+\theta_{z_ix}^{\nu_{z_ix}}\right)\left(1+\theta_{z_iy}^{\nu_{z_iy}}\right)} = \lambda_{z_i}, \tag{23}$$

from which explicit relationships come out for V_y^0 and $V_{z_i}^0$:

$$\begin{aligned} V_y^0 &= V_y^0(\theta_{yx}, \nu_{yx}, \bar{V}_{yx}, \lambda_y) \\ V_{z_i}^0 &= V_{z_i}^0(\theta_{z_ix}, \nu_{z_ix}, \bar{V}_{z_ix}, \theta_{z_iy}, \nu_{z_iy}, \bar{V}_{z_iy}, \lambda_{z_i}). \end{aligned}$$

In summary, the following schemes emerge:

- scheme $\mathcal{S}_1$: operon-motif A. $n_1 = 10$ independent parameters:

$$\theta_{xx}, \nu_{xx}, \lambda_x, \alpha, \bar{V}_{zx}, \theta_{zx}, \nu_{zx}, \lambda_{z_1}, \lambda_{z_2}, \lambda_{z_3}$$

- scheme $\mathcal{S}_2$: non operon-motif A. $n_2 = 16$ independent parameters:

$$\begin{array}{l}\theta_{xx}, \nu_{xx}, \lambda_x, \alpha, \bar{V}_{z_1x}, \bar{V}_{z_2x}, \bar{V}_{z_3x}, \\ \theta_{z_1x}, \theta_{z_2x}, \theta_{z_3x}, \nu_{z_1x}, \nu_{z_2x}, \nu_{z_3x}, \lambda_{z_1}, \lambda_{z_2}, \lambda_{z_3},\end{array}$$

- scheme $\mathcal{S}_3$: operon-motif B. $n_3 = 17$ independent parameters:

$$\begin{array}{l}\theta_{xx}, \nu_{xx}, \lambda_x, \alpha, \bar{V}_{yx}, \theta_{yx}, \nu_{yx}, \lambda_y, \\ \bar{V}_{zx}, \theta_{zx}, \nu_{zx}, \\ \bar{V}_{zy}, \theta_{zy}, \nu_{zy}, \lambda_{z_1}, \lambda_{z_2}, \lambda_{z_3}\end{array}$$

- scheme $\mathcal{S}_4$: non operon-motif B. $n_4 = 29$ independent parameters:

$$\begin{array}{l}\theta_{xx}, \nu_{xx}, \lambda_x, \alpha, \bar{V}_{yx}, \theta_{yx}, \nu_{yx}, \lambda_y, \lambda_{z_1}, \lambda_{z_2}, \lambda_{z_3} \\ \bar{V}_{z_1x}, \bar{V}_{z_2x}, \bar{V}_{z_3x}, \theta_{z_1x}, \theta_{z_2x}, \theta_{z_3x}, \nu_{z_1x}, \nu_{z_2x}, \nu_{z_3x}, \\ \bar{V}_{z_1y}, \bar{V}_{z_2y}, \bar{V}_{z_3y}, \theta_{z_1y}, \theta_{z_2y}, \theta_{z_3y}, \nu_{z_1y}, \nu_{z_2y}, \nu_{z_3y}.\end{array}$$

Measurements are taken from Z_1, Z_2, Z_3 products, according to an experimental device based on a GFP reporter gene, so that we can assume an arbitrary small sampling time. In order to mathematically formalize the motif identification algorithm, let Ξ_i, $i = 1, \ldots, 4$, the positive orthant of $I\!R^{n_i}$, be the parameter space for the candidate scheme to be representative of the experimental data, and let $\mathcal{Z}_u = \{z_j(kT),\ j = 1, 2, 3;\ k = 1, \ldots, K\}$ the available set of measurements corresponding to the case of a given perturbation signal $u(\cdot)$. Conversely, let $\widetilde{\mathcal{Z}}_{\mathcal{S}_i} = \{\tilde{z}_j(\xi, kT),\ j = 1, 2, 3;\ k = 0, 1, \ldots, K;\ \xi \in \Xi_i\}$ data obtained by means of simulating scheme $\mathcal{S}_i$ with a given $\xi \in \Xi_i$. To each triple $(\mathcal{Z}_u, \xi, \mathcal{S}_i)$ a standard sum of square loss function $J_{\mathcal{Z}_u}(\xi, \mathcal{S}_i)$ is defined:

$$J_{\mathcal{Z}_u}(\xi, \mathcal{S}_i) = \sum_{j=1}^{3} \sum_{k=1}^{K} \omega_{jk} \big(z_j(kT) - \tilde{z}_j(\xi, kT) \big)^2 \tag{24}$$

with $\omega_{jk} > 0$, $j = 1, 2, 3$, $k = 1, \ldots, K$ a set of weights; to each pair $(\mathcal{Z}_u, \mathcal{S}_i)$ is associated the following measure:

$$\mathcal{M}_{\mathcal{Z}_u}(\mathcal{S}_i) = \min_{\xi \in \Xi_i} J_{\mathcal{Z}_u}(\xi, \mathcal{S}_i). \tag{25}$$

Then the following criterion is given:

Criterion. Scheme $\mathcal{S}_i$ is *more likely* than scheme $\mathcal{S}_r$, according to a set of experimental measurements $\mathcal{Z}_u$ related to a perturbation signal $u(\cdot)$, if the measure associated to $\mathcal{S}_i$ is less than the measure associated to $\mathcal{S}_r$, more formally:

$$\mathcal{S}_i \prec_{\mathcal{Z}_u} \mathcal{S}_r \quad \Longleftarrow \quad \mathcal{M}_{\mathcal{Z}_u}(\mathcal{S}_i) < \mathcal{M}_{\mathcal{Z}_u}(\mathcal{S}_r). \tag{26}$$

4. Simulation results

Computer simulations have been performed in order to test the effectiveness of the proposed motif identification algorithm. Simulations reported here, all refer to an experimental time interval of 500 time units and to a perturbation rectangular-wise input $u(t)$. Data are acquired since time $t = 0$; signal $u(t)$ is at its high level for $t \in [10, 350]$, otherwise it is set to zero. Measurements undergo an error of 5%. The cases of schemes $\mathcal{S}_3$, $\mathcal{S}_4$ have been simulated. The twofold purpose of motif (presence or not of gene Y in the network) and biological (operon/non operon) identification has been performed. For both the operon-motif B and non operon-motif B cases, all the schemes $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$, $\mathcal{S}_4$ have been used to

Table 1: Scheme $\mathcal{S}_3$ parameter values

$\theta_{xx} = 0.9$	$\nu_{xx} = 1.0$	$\lambda_x = 0.05$	$\alpha = 0.4913$
$\bar{V}_{yx} = 4.0361$	$\theta_{yx} = 2.0$	$\nu_{yx} = 3.0$	$\lambda_y = 0.05$
$\bar{V}_{zx} = 3.6886$	$\theta_{zx} = 2.1$	$\nu_{zx} = 2.0$	$\lambda_{z_1} = 0.08$
$\bar{V}_{zy} = 4.8$	$\theta_{zy} = 1.5$	$\nu_{zy} = 4.2$	$\lambda_{z_2} = 0.02$
$\lambda_{z_3} = 2.0$			

Table 2: Scheme $\mathcal{S}_4$ parameter values

$\theta_{xx} = 0.9$	$\nu_{xx} = 1.0$	$\lambda_x = 0.05$	$\alpha = 0.4913$
$\bar{V}_{yx} = 5.7793$	$\theta_{yx} = 2.0$	$\nu_{yx} = 3.0$	$\lambda_y = 0.05$
$\bar{V}_{z_1x} = 3.6886$	$\bar{V}_{z_2x} = 2.7743$	$\bar{V}_{z_3x} = 3.0924$	$\bar{V}_{z_1y} = 4.8$
$\theta_{z_1x} = 2.1$	$\theta_{z_2x} = 4.5$	$\theta_{z_3x} = 9.99$	$\bar{V}_{z_2y} = 5.1$
$\nu_{z_1x} = 3.0$	$\nu_{z_2x} = 3.5$	$\nu_{z_3x} = 5.1$	$\bar{V}_{z_3y} = 6.3$
$\theta_{z_1y} = 1.875$	$\theta_{z_2y} = 2.125$	$\theta_{z_3y} = 3.7375$	$\nu_{z_1y} = 4.2$
$\nu_{z_2y} = 3.1$	$\nu_{z_3y} = 5.5$	$\lambda_z = 0.08$	

estimate the scheme likelihood in order to identify the more likely network motif: the loss function (24), for $i = 1, 2, 3, 4$, has been evaluated to solve the optimization problem (25), with all the weight set to 1. The parameters values chosen for the $\mathcal{S}_3$ and $\mathcal{S}_4$ schemes are shown in tables I and II, respectively, and the simulated expression time courses of the target genes Z_1, Z_2, Z_3 are reported in Fig.3 (reference to Table I), and Fig.4 (reference to Table II). Note that, as far as the non-operon case, we have supposed that:

$$\lambda_{z_1} = \lambda_{z_2} = \lambda_{z_3} = \lambda_z, \tag{27}$$

that is, we considered an equal degradation rate for the three transcripts z_i, $i = 1, 2, 3$, whereas the other parameters are assumed to be different. Given the set of available gene expression data $\mathcal{Z}_u$ for the target genes Z_1, Z_2, Z_3 according to a sample time of 0.5min, the measures obtained for (25) are reported in Table III.

Improvements are apparent when attempting to identify the correct scheme. For instance, according to an $\mathcal{S}_3$ original scheme, by choosing a wrong topological scheme (operon motif A, $\mathcal{S}_1$), the index increases of about 10%, while by choosing a wrong biological scheme (non-operon motif B, $\mathcal{S}_4$) the index increases of more than 700%. Analogously, according to an $\mathcal{S}_4$ original scheme, by choosing a wrong topological scheme (non-operon motif A, $\mathcal{S}_2$), the index increases of about 67%, while by choosing a wrong biological scheme (operon motif B, $\mathcal{S}_3$) the index increases of about 300%. It has to be stressed that, according to the first simulation set, a good fitting is available for genes Z_i from both $\mathcal{S}_1$ and $\mathcal{S}_3$, and only a numerical criterion allows to establish which of the two scheme is actually underlining the experimental framework: the only trivial visual inspection is not as efficient as the numerical identification procedure proposed.

Note that, as it is shown in Fig.4, the thresholds in scheme $\mathcal{S}_4$ have been chosen in order to obtain a very common scheme consisting in temporal programs of expression, in which genes are activated one by one according to a define order. In this case, a *Last In First Out* order has been considered, which is very common to SIM motif. Nevertheless, the proposed algorithm correctly identified the MO-FFL scheme of $\mathcal{S}_4$.

As a by product, the proposed approach allows also to estimate the time with a good degree of precision also the parameters corresponding to genes X and Y. For instance, according to the original $\mathcal{S}_3$ scheme, Fig.5 and Fig.6 show the real/estimated time evolutions of genes products $x(t)$ and $y(t)$, respectively.

5. Conclusions

Living cells are the product of gene expression programs involving regulated transcription of thousands of genes. The regulatory network controls the transcriptional activity of genes by turning them on or

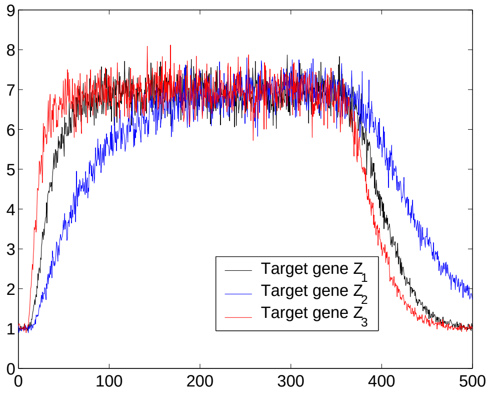

Figure 3: Simulated target genes time course: operon-motif B.

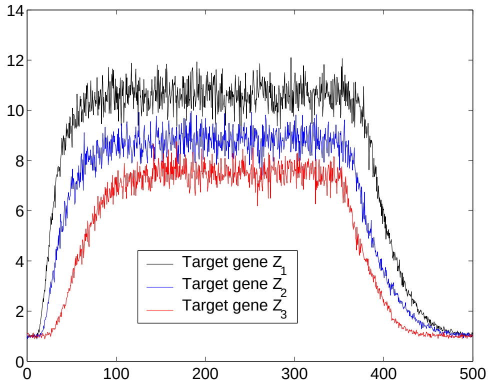

Figure 4: Simulated target genes time course: non operon-motif B.

off. Such coordination is required to precisely regulate the timing and amount of the proteins needed for an appropriate cell response to the internal or environmental perturbation. The basic regulatory structures underlying such behavior are the *single input motif* and the *feed-forward loop motif*. The first is composed by a transcription factor regulating a number of target genes, whereas the second is composed by a master transcription factor regulating another transcription factor, both regulating a group of target genes. Those two motifs are ubiquitous in many organisms, including human cells, and their behavior can be modeled using nonlinear ODEs, based on a Hill model of cooperative/competitive binding. In this paper, we proposed an algorithm able to identify the correct regulatory motif basing only on the expression data of the targets genes. The results showed in this paper on simulated data clearly demonstrate that our method may be an important computational tool for the formulation of hypotheses on the structure of the regulatory network and for the design of the experimental validation.

Table 3: Identification algorithm performances

Scheme $\mathcal{S}_3$	Scheme $\mathcal{S}_4$
$\mathcal{M}_{\mathcal{Z}_u}(\mathcal{S}_1) = 257.09$	$\mathcal{M}_{\mathcal{Z}_u}(\mathcal{S}_1) = 2169.64$
$\mathcal{M}_{\mathcal{Z}_u}(\mathcal{S}_2) = 11435.00$	$\mathcal{M}_{\mathcal{Z}_u}(\mathcal{S}_2) = 688.28$
$\mathcal{M}_{\mathcal{Z}_u}(\mathcal{S}_3) = 232.29$	$\mathcal{M}_{\mathcal{Z}_u}(\mathcal{S}_3) = 1672.27$
$\mathcal{M}_{\mathcal{Z}_u}(\mathcal{S}_4) = 1873.35$	$\mathcal{M}_{\mathcal{Z}_u}(\mathcal{S}_4) = 411.64$

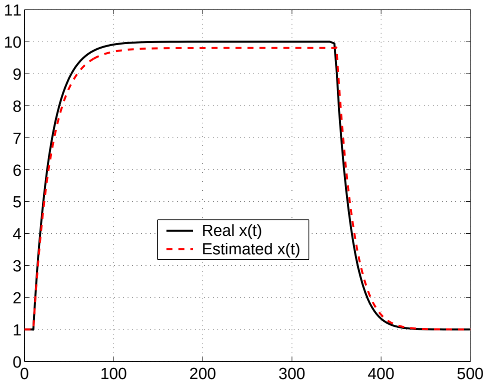

Figure 5: Real/estimated gene product $x(t)$ time courses: non operon-motif B.

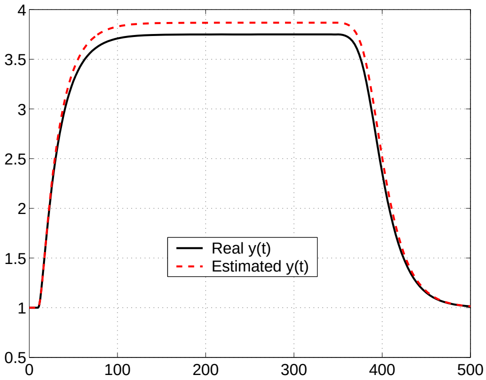

Figure 6: Real/estimated gene product $y(t)$ time courses: non operon-motif B.

Properly extending the state vector of the nonlinear system by considering its unknown parameters as state variables, the state observer real time technique for nonlinear systems, given in [3], can be used to solve the parameters identification problem. A potential application of our algorithm for real time applications is in the field of biosensors signal processing. In fact, a recent technology [16] consist of a bioreporter organism interfaced with an integrated circuit. The bioreporter is engineered to luminesce when a targeted substance is encountered, while the circuit is designed to detect the luminescence, process the signal, and communicate the results. Therefore, our approach may provide the correct reconstruction algorithm for the detection of the environmental signal driving the gene expression response measured by the luminescence.

References

[1] U. Alon, Network motifs: theory and experimental approaches, Nature Reviews Genetics 8, pp.450-461, 2007

[2] D. Angeli and E.D. Sontag. Oscillations in I/O monotone systems. IEEE Transactions on Circuits and Systems, Special Issue on Systems Biology, 55:166-176, 2008

[3] M. Dalla Mora, A. Germani, C. Manes, Design of state observers from a drift-observability property, IEEE Trans. Automatic Control, Vol.45, No.8, pp. 1536-1540, 2000

[4] H. de Jong, Modeling and simulation of genetic regulatory systems: a literature review, Journal of Computational Biology, Vol.9, No.1, pp.67-103, 2002

[5] L. Farina, A. De Santis, G. Morelli and I. Ruberti, Dynamic measure of gene co-regulation, *IET Systems Biology*, Vol.1, No.1, pp.10-17, 2007

[6] T. S. Gardner, J. J., Faith, Reverse-Engineering Transcription Control Networks, *Physics of Life Reviews,* Vol. 2, pp.65-88, 2005

[7] N.L. Garneau, J. Wilusz and C.J. Wilusz, The highways and byways of mRNA decay, Nature Molecular Cell Biology Review, Vol.8, pp.113-126, 2007

[8] H. Kitano, Computational systems biology. Nature 420, pp.206-210, 2002

[9] H. Kitano, Towards a theory of biological robustness, Molecular Systems Biology, 3:137, 2007

[10] S. Krishna, S. Maslov, K. Sneppen, UV-Induced Mutagenesis in Escherichia coli SOS Response: A Quantitative Model. PLoS Comput Biol 3(3): e41, 2007

[11] Lee *et al*, Transcriptional regulatory networks in *Saccharomyces cerevisiae*, Science, Vol. 298, No 25, pp.799-804, 2002

[12] N. M. Luscombe, M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann and M. Gerstein, Genomic analysis of regulatory network dynamics reveals large topological changes, Nature, Vol. 431, No. 16, pp.308-312, 2004

[13] S. Mangan, A. Zaslaver, U. Alon, The Coherent Feedforward Loop Serves as a Sign-sensitive Delay Element in Transcription Networks, Journal of Molecular Biology, Vol.334, No.2, pp.197-204, 2003

[14] I. Mogno, L. Farina and T. Gardner, CRP drives a non-specific expression burst in E. coli sugar catabolic operons during starvation, *submitted*

[15] M.C. Palumbo, L. Farina, A. De Santis, A. Giuliani, A. Colosimo, G. Morelli and I. Ruberti, The relevance of post-transcriptional regulation for the temporal compartmentalization of cellular cycles, *FEBS Journal*, to appear, 2008

[16] S. Ripp, S. Moser, B. Weathers, S. Caylor, B. Blalock, S. Islam, G. Sayler, Bioluminescent bioreporter integrated circuit (BBIC) sensors, Bio Micro and Nanosystems Conference, BMN '06, San Francisco, CA, pp.59-60, 2006

[17] N. Soranzo, G. Bianconi, C. Altafini, Comparing Association Network Algorithms for Reverse Engineering of Large Scale Gene Regulatory Networks: Synthetic vs Real Data, *Bioinformatics,* Vol. 23, No. 13, pp.1640-1647, 2007

[18] J. Stelling, U. Sauer, Z. Szallasi, F.J. Doyle III, J. Doyle, Robustness of cellular functions. Cell 118, pp.675-685, 2004

[19] T.M. Yi, Y. Huang, M.I. Simon and J. Doyle, Robust perfect adaptation in bacterial chemotaxis through integral feedback control, Proceedings of the National Academy of Sciences USA, Vol. 97, No. 9, pp.4649-4653, 2000

[20] A. Zaslaver, A. Bren, M. Ronen, S. Itzkovitz, I. Kikoin, S. Shavit, W. Liebermeister, M. G. Surette and U. Alon, A comprehensive library of fluorescent transcriptional reporters for Escherichia coli, Nature Methods 3, pp.623-628, 2006