52nd IEEE Conference on Decision and Control
December 10-13, 2013. Florence, Italy

Optimizing Regulation Functions in Gene Network Identification

Guilhem Richard, A. Agung Julius, and Calin Belta

Abstract— **This paper is concerned with the problem of identifying a discrete-time dynamical system model for a gene regulatory network with unknown topology using time series gene expression data. The topology of such a network can be characterized by a set of regulation hypotheses, one for each gene. In our earlier work, we formulated a convex optimization method to select the regulation hypotheses (and hence the network topology). In this paper, we further optimize the dynamics of the inferred network. Specifically, for a given topology, we minimize the ℓ_2 distance between the experimental data and the model prediction. We illustrate the performance of our algorithm by identifying models for gene networks with known topology.**

Keywords: gene network identification, monotone functions, optimization.

I. Introduction

Gene regulatory network identification is one of the main challenges in systems biology and is related to the problem of identifying how a group of genes interact based on their expression activities [1], [2]. As gene regulatory networks play a fundamental role in the regulation of biological processes, gene network identification (also termed *reverse engineering*) is a crucial step in understanding these processes.

Identification of Gene Regulatory Networks (GRNs) is a difficult problem for the following main reasons:

- The size of the network can be very large.
- The measurements are noisy.
- Although it is possible to have a large quantity of data (genome-wide) for each snapshot, the number of snapshots is typically small. This is because obtaining a large number of snapshots is highly impractical, due to logistical and cost considerations.
- The dynamics of the GRNs are highly nonlinear.

There are several families of methods for identification of GRNs from gene expression data. For a detailed overview, we refer the reader to the review paper by Bansal *et al.* [3]. Within the systems and control community, identification of GRNs is usually done by modeling networks as dynamical systems [4]. This is also the approach reported in this paper.

For certain model structures (*e.g.* linear systems), with the availability of gene expression measurements, the network can in principle be reconstructed by inverting the data. However, as the measurements are noisy, the reconstructed network tends to be populated with spurious interconnections (*i.e.* false positives). This concern motivates sparse identification that aims at getting a network model with as few connections as possible without losing the fitness to the data [1], [5], [6]. Identification of GRNs in general and sparse identification of GRNs in particular are quite active research areas. For example, de Jong *et al.* developed a method for identification of GRNs using the structure of piecewise affine dynamical systems [7], [8]. Papachristodoulou *et al.* developed a model for identification of sparse networks using Hill functions to describe the dynamics of gene-gene interactions [9]. Earlier work by one of the authors of the current paper aimed at identifying sparse networks based on genetic perturbation data, assuming that the dynamics can be described (locally) as a linear system [10], [11], [12]. Recent work by Yuan *et al.* [13], [14] investigated handling sparsity by using Akaike information criterion. A different approach was taken by Chang and Tomlin [15], where compressive sensing is used to identify a sparse linear model.

The method reported in this paper is an extension of our earlier work in [16], [17], which has been applied to modeling the regulation of the Toll-like receptor signaling pathway [18], [19]. In short, given a set of expression activity data, we consider the question whether the data could have come from a given network topology, where the interaction dynamics can be represented as **continuous nonnegative monotonic** functions (in short, CNM functions). Requiring monotonicity is in line with the fact that most, if not all, known mathematical models for gene regulation use monotonic functions. In [16], [17], we proved that this question is equivalent to the feasibility of a Linear Program involving the expression activity data. Our method is essentially based on model invalidation, rather than model identification. A nice feature of our method is that the regulator sets of the genes in the network can be computed in parallel.

Using monotonicity as criterion for invalidating regulation hypothesis has also been pursued by others in the field. For example, the work by Porreca *et al.* [20] proposed a two-staged process in identifying a continuous-time differential equation model for GRNs. In the first stage, network topologies that are inconsistent with the data are rejected. This first stage is also based on the monotonicity argument. In a more recent work, Angeli and Sontag [21] exploited the monotonicity argument to reject regulation hypotheses based on the sign pattern of the expression data's temporal gradient.

The focus is this paper is the optimal identification of the network model after its topology is found using the approach reported in [16], [17]. We propose two procedures

This work was partially supported at Boston University by the NSF under grant EF 1137900 and at RPI by grant EF 1137906.

G. Richard is with the Program in Bioinformatics, Boston University, Boston, MA 02215, Email: grichard@bu.edu. C. Belta is with the Department of Mechanical Engineering, the Division of Systems Engineering, and the Program in Bioinformatics, Boston University, Boston, MA 02215, Email:cbelta@bu.edu. A. Julius is with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, Email: agung@ecse.rpi.edu.

978-1-4673-5717-3/13/$31.00 ©2013 IEEE

that aim at optimizing the regulation functions obtained after the topology identification. Specifically, we seek to minimize the ℓ_2 distance between experimental and predicted time-series expression data. The first method refines the domains of the regulation functions, while the second technique uses a gradient descent method to explore the space of regulation functions. Both methods diminish the error of the network model and allow for better prediction of gene dynamics.

II. Notations and Preliminary Results

In the following, we introduce some mathematical notations that we shall use in the subsequent discussion. These notations are similar to those in [17], but repeated here to make this paper self-contained.

Assume that the GRN we are working with consists of G genes. We propose a discrete-time model for the dynamics of the gene expression in the form:

$$x_i[k+1] - x_i[k] = -\lambda_i x_i[k] + f_i(x_1^i[k], \cdots, x_{K_i}^i[k]), \quad (1)$$

for all $i \in \{1, \cdots, G\}$, where x_i denotes the expression activity of Gene i, $\lambda_i \geq 0$ is the decay parameter of Gene i (degradation component), f_i is the regulation function of Gene i (production component), and $(x_1^i, \cdots, x_{K_i}^i)$ are the expression activities of the regulators of Gene i. To simplify the discussion, we abuse the notation and define the regulator set of Gene i as $G_i^R \triangleq \{x_1^i, \cdots, x_{K_i}^i\}$. We assume that λ_i, the regulator set G_i^R, and function f_i are unknown, and have to be identified from the experimental data.

Suppose that we are provided with a sequence of gene expression activities for $(N+1)$ equally spaced time points. Denote the expression data of Gene i at time j as $x_{i,j}, 1 \leq i \leq G, 0 \leq j \leq N$. The (time) differential expression activity $q_{i,j}$, $1 \leq i \leq G$, $0 \leq j \leq N-1$, is defined as

$$q_{i,j} \triangleq x_{i,j+1} - x_{i,j}. \quad (2)$$

Therefore, for the data to fit the model, the following relation must hold:

$$q_{i,j} = -\lambda_i x_{i,j} + f_i(x_{1,j}^i, x_{2,j}^i, \ldots, x_{K_i,j}^i). \quad (3)$$

As in [16], [17], the following assumption for the regulatory function f_i is adopted.

CNM Assumption: The function $f_i(x_1^i, x_2^i, \ldots, x_{K_i}^i)$ is continuous, nonnegative and monotonic (CNM) in each $x_1^i, \ldots, x_{K_i}^i \in G_i^R$.

If f_i is monotonically increasing in x_k^i, then the k-th regulator of Gene i is considered an activator. Conversely, when f_i is monotonically decreasing in x_k^i, the regulator is considered a repressor of Gene i. Thus, the set of regulators G_i^R can be split into two disjoint sets

$$G_i^R = G_i^{R+} \cup G_i^{R-}, \quad (4)$$

where G_i^{R+} and G_i^{R-} are the sets of activators and repressors of Gene i, respectively.

Regulation Hypothesis: A regulation hypothesis $\mathcal{R}_i$ for Gene i states that $G_i^R := G_i^{R+} \cup G_i^{R-}$ is the set of regulators of Gene i.

Given a regulation hypothesis $\mathcal{R}_i = (G_i^{R+}, G_i^{R-})$, we denote:

- the vector of expression activities of all activator genes at time j as $x_{\mathbf{a},j}^i$,
- the vector of expression activities of all repressor genes at time j as $x_{\mathbf{r},j}^i$,
- the vector of expression activities of regulator genes at time j as $x_{\mathbf{R},j}^i$,

The regulation hypothesis $\mathcal{R}_i$ is equivalent to the existence of some $\lambda_i \geq 0$, and a CNM function $f_i(\diamond, \circ)$ that is monotonically increasing in the variables in $\diamond$ and monotonically decreasing in the variables in $\circ$, such that

$$q_{i,j} = -\lambda_i x_{i,j} + f_i(x_{\mathbf{a},j}^i, x_{\mathbf{r},j}^i), \ \forall j \in \{0, \ldots, N-1\}. \quad (5)$$

In [17], we showed that the validity of the regulation hypothesis $\mathcal{R}_i$ is equivalent to the feasibility of a Linear Program involving $x_{\mathbf{R},j}^i$ for $j \in \{0, \ldots, N-1\}$ and $x_{i,j}$ for $j \in \{1, \ldots, N\}$.

Definition 1: Given a regulation hypothesis $\mathcal{R}_i = (G_i^{R+}, G_i^{R-})$, we define the partial ordering $\preceq_{\mathcal{R}_i} \subset \mathbb{R}^{|G_i^R|} \times \mathbb{R}^{|G_i^R|}$ as

$$x_{\mathbf{R},j}^i \preceq_{\mathcal{R}_i} x_{\mathbf{R},n}^i :\Leftrightarrow \begin{cases} x_{\mathbf{a},j}^i \leq x_{\mathbf{a},n}^i \\ x_{\mathbf{r},j}^i \geq x_{\mathbf{r},n}^i \end{cases} \quad (6)$$

$\mathcal{S}_{\mathcal{R}_i} := \{(j,n) \in \{0, \ldots, N-1\}^2 \mid x_{\mathbf{R},j}^i \preceq_{\mathcal{R}_i} x_{\mathbf{R},n}^i\}$.

In addition to formulating the necessary and sufficient conditions for the validity of the regulation hypothesis, we also formulated a Linear Quadratic optimization problem to quantify how far the data are from satisfying the regulation hypothesis $\mathcal{R}_i$. The Frobenius norm of ε_i (denoted as $\|\varepsilon_i\|_F$ below) quantifies this distance:

$$\begin{aligned} & \min \|\varepsilon_i\|_F \ \text{ subject to} && (7) \\ & q_{i,j} = -\lambda_i x_{i,j} + \hat{q}_{i,j} + \varepsilon_{i,j}, \ \forall j \in \{0, \ldots, N-1\}, \\ & \lambda_i \geq 0, \\ & \hat{q}_{i,j} \geq 0, \ \forall j \in \{0, \ldots, N-1\}, \\ & \hat{q}_{i,j} \geq \hat{q}_{i,n}, \ \forall (j,n) \in \mathcal{S}_{\mathcal{R}_i}, \end{aligned}$$

with λ_i, $\varepsilon_{i,j}$, and $\hat{q}_{i,j}$, $j \in \{0, \ldots, N-1\}$ as the optimization variables.

The following statement, which follows immediately from the results in [17], summarizes the relationship between $\|\varepsilon_i\|_F$ and the validity of the regulation hypothesis $\mathcal{R}_i$.

Theorem 1: $\|\varepsilon_i\|_F = 0$ if and only if the regulation hypothesis $\mathcal{R}_i$ is valid.

Because of measurement noise and/or unmodelled external variables in the network, in practice, $\|\varepsilon_i\|_F > 0$. A regulation hypothesis $\mathcal{R}_i$ is accepted if $\|\varepsilon_i\|_F$ is small enough, or if it is the smallest among competing hypotheses.

III. Problem Formulation

Once a regulation hypothesis $\mathcal{R}_i$ is accepted with $\|\varepsilon_i\|_F$ small enough, we showed in [17] that the regulation function f_i can be constructed by interpolating the solution $\hat{q}_{i,j}$ from the optimization problem given in (7). There is not a unique way to do this, and in [17] we presented an multi-affine interpolation technique for this purpose.

In this paper, we want to exploit the non-uniqueness of the interpolation. That is, we formulate a cost criterion that aims to fit the model prediction and the experimental time-series data, and optimize this cost criterion while we find the regulation function. We propose two methods that aim at fine tuning an initial model of the regulation function. Both methodologies diminish our cost criterion and yield an optimized model that predicts more accurate gene dynamics.

Suppose that we form a candidate network by proposing the regulation hypotheses $\mathcal{R}_1, \ldots, \mathcal{R}_G$. That is, we define the (graph of the) network by specifying the regulators for each gene. Following the reasoning for individual genes above, this set of hypotheses is valid if and only if there exist:

(C1) A regulation function $F = (f_1, f_2, \ldots, f_G)$ that is continuous, non-negative, and monotonic in the sense that is required by the regulation hypothesis. That is if the regulation hypotheses state that the k-th regulator is an activator of Gene i, then f_i is monotonically increasing w.r.t. x_k^i, etc.

(C2) Some non-negative $\lambda_1, \ldots, \lambda_G$, such that for all $i \in \{1, \ldots, G\}$ and $j \in \{0, \ldots, N-1\}$

$$q_{i,j} \triangleq x_{i,j+1} - x_{i,j} = -\lambda_i x_{i,j} + f_i\left(x_{1,j}, \ldots, x_{G,j}\right). \quad (8)$$

If F and $\lambda_1, \ldots, \lambda_G$, are given, we can define an error quantity that measures the ℓ_2 distance between the experimental gene expression time series data and the gene expression time series generated by the model as follows:

$$\varepsilon_\ell^2(F, \lambda_1, \ldots, \lambda_G) \triangleq \sum_{i=1}^{G} \sum_{j=0}^{N+1} \left(x_{i,j} - \hat{x}_{i,j}\right)^2, \quad (9)$$

where the model prediction $\hat{x}_{i,j}$ is defined recursively as follows.

$$\hat{x}_{i,0} = x_{i,0}, \quad (10a)$$

$$\hat{x}_{i,j+1} = \hat{x}_{i,j} - \lambda_i \hat{x}_{i,j} + f_i\left(\hat{x}_{1,j}, \ldots, \hat{x}_{G,j}\right), \quad (10b)$$

for all $i \in \{1, \ldots, G\}$, and $j \in \{0, \ldots, N-1\}$.

The problem of finding the regulatory functions that best fit the time series data in the ℓ_2 sense, given the regulation hypotheses $\mathcal{R}_1, \ldots, \mathcal{R}_G$, can then be formulated as follows:

$$\begin{aligned} &\min \varepsilon_\ell^2(F, \lambda_1, \ldots, \lambda_G) \text{ subject to} \\ &F \text{ conforms with } \mathcal{R}_1, \ldots, \mathcal{R}_G, \\ &\lambda_1, \ldots, \lambda_G \geq 0, \end{aligned} \quad (11)$$

with $F, \lambda_1, \ldots, \lambda_G$ as the optimization variables. By "conforming with $\mathcal{R}_1, \ldots, \mathcal{R}_G$" we mean that F is continuous, nonnegative, and has the monotonicity property as dictated by $\mathcal{R}_1, \ldots, \mathcal{R}_G$.

Remark 2: Notice that the optimization "variable" F assumes its value in a function space. Also, we can observe that the feasible set in (11) is a cone, which is a convex set.

Remark 3: Note that (11) differs from (7) in the sense that (11) uses the model **recursively** to propagate the prediction from the initial state. If we were to replace (10b) with

$$\hat{x}_{i,j+1} = x_{i,j} - \lambda_i x_{i,j} + f_i\left(x_{1,j}, \ldots, x_{G,j}\right), \quad (12)$$

then (11) would be equivalent to (7).

The following analog of Theorem 1 can be stated.

Theorem 4: $\varepsilon_\ell^2(F, \lambda_1, \ldots, \lambda_G) = 0$ if and only if the regulation hypotheses $\mathcal{R}_1, \ldots, \mathcal{R}_G$ are valid.

Although the feasible set in (11) is convex, the cost function is generally not, making (11) a non-convex optimization problem in general. In the next section we present techniques that we developed to approximate the solution of (11).

IV. Optimization of Regulatory Functions

In order to find the best regulation function F that minimizes $\varepsilon_\ell^2(F, \lambda_1, \ldots, \lambda_G)$ in (11), we optimized the generation of the individual CNM functions f_i that compose F. If Gene i has K_i regulators, then its regulation function is modeled as a piecewise multi-affine function $f_i : \mathbb{R}^{K_i} \to \mathbb{R}$. We define a grid $\mathcal{G}_i$ in the $\mathbb{R}^{K_i}$ space, whose vertices map to the boundaries of the domains of the affine pieces of f_i (Fig. 1). We assign a *regulation value* to each vertex, corresponding to the values of the regulation function at these specific locations in $\mathbb{R}^{K_i}$. Optimizing f_i translates to finding the best sets of vertices and regulation values such that f_i accurately predicts the experimental data. Our optimization procedures start from an initial regulation function obtained by solving problem (7) for the chosen set of regulatory genes. The vertices of $\mathcal{G}_i$ are given by the elements of $\{x_{k,j}^i\}, k \in \{1, \ldots, K_i\}$, and $j \in \{0, \ldots, N-1\}$ (Fig. 1). From the N^{K_i} vertices in $\mathcal{G}_i$, N are defined by the experimental time points. We denote this set as $\mathcal{V}_e$. The value associated with each vertex in $\mathcal{V}_e$ is given by the optimization variables $\hat{q}_{i,j}$ from problem (7), such that:

$$f_i(x_{1,j}^i, x_{2,j}^i, \ldots, x_{K_i,j}^i) \triangleq \hat{q}_{i,j}, \quad (13)$$

where $j \in \{0, \ldots, N-1\}$. We denote the $N^{K_i} - N$ remaining vertices of $\mathcal{G}_i$ as $\mathcal{V}_f$. The value assigned to each vertex in $\mathcal{V}_f$ is constrained such that f_i must remain continuous, non-negative, and monotonic. The non-uniqueness of this assignment gives rise to a family of regulation functions that conform with $\mathcal{R}_i$. Our goal is to find a CNM function f_i that satisfies the regulation hypothesis $\mathcal{R}_i$ and that minimizes $\varepsilon_\ell^2(F, \lambda_1, \ldots, \lambda_G)$. We propose two optimization procedures that aim at identifying such a function.

The first method focuses on increasing the number of vertices by refining the grid $\mathcal{G}_i$. New vertices are identified while simulating expression time-series from the gene model. The second method concentrates on the regulation values associated with the vertices in $\mathcal{V}_f$. Values are assigned by performing a stochastic gradient descent method.

A. Domain Extension

The regulation values associated with the vertices in $\mathcal{G}_i$ are used to estimate the value of the regulation function for the remaining points in $\mathbb{R}^{K_i}$. A simple improvement to the model consists in increasing the number of vertices to obtain a finer grid. Refining $\mathcal{G}_i$ allows f_i to be more sensitive to variations in the expression of the regulatory genes, but comes at the expense of increasing the state space of the gene model.

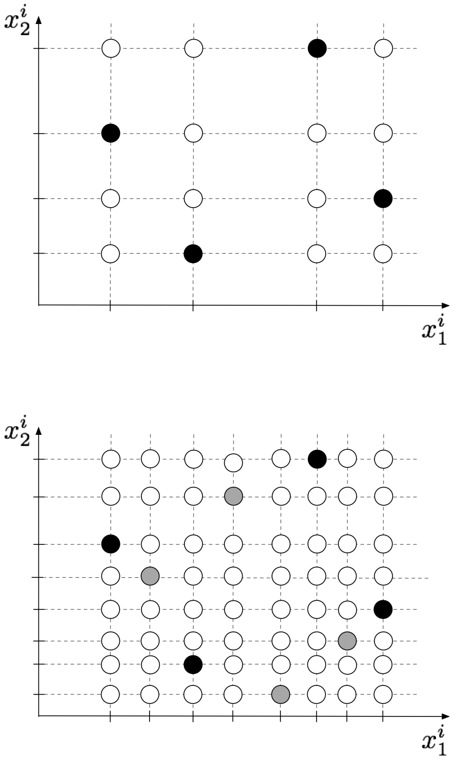

Fig. 1: Schematic representations of the grid $\mathcal{G}_i$. For simplicity, we show the case where Gene i has two regulators genes (x_1^i and x_2^i) and $N+1=5$. The top figure represents the grid $\mathcal{G}_i$ before any optimization procedure. Vertices in $\mathcal{V}_e$ and $\mathcal{V}_f$ are shown with black and white circles, respectively. The bottom figure shows the same grid after the Domain Extension procedure. New vertices $\mathcal{V}_r$, shown in gray, are identified by following Alg. 1. This procedure allows to refine the grid $\mathcal{G}_i$ and to define a new CNM function f_i that is more sensitive to the expression levels of x_1^i and x_2^i.

An initial gene model F is constructed from the list of vertices $\mathcal{V}_e$ defined by each regulation hypothesis. Predicting expression time-series with this model will give rise to simulated expression levels $\hat{x}_{i,j}$ that do not exactly match experimental values, since $\|\varepsilon_i\|_F > 0$ in practice. Better predictions can be obtained by identifying the value of the regulation function that minimizes the distance between simulated and experimental expressions at each iteration j of the prediction. This translates to defining a new vertex $\mathcal{V}_r$ on the grid $\mathcal{G}_i$, such that its coordinate is defined by $\hat{x}_{1,j}^i, \ldots, \hat{x}_{K_i,j}^i$. The regulation value associated with this vertex is found by solving the following problem:

$$\begin{aligned} &\min(x_{i,j+1} - \hat{x}_{i,j+1})^2 \text{ subject to} \\ &f_i \text{ conforms with } \mathcal{R}_i, \\ &\hat{x}_{i,j+1} = \hat{x}_{i,j} - \lambda_i \hat{x}_{i,j} + f_i(\hat{x}_{1,j}^i, \ldots, \hat{x}_{K_i,j}^i), \end{aligned} \tag{14}$$

with $f_i(\hat{x}_{1,j}^i, \ldots, \hat{x}_{K_i,j}^i)$ as the optimization variable. Note that by "f_i conforms with $\mathcal{R}_i$" we mean that the choice of $f_i(\hat{x}_{1,j}^i, \ldots, \hat{x}_{K_i,j}^i)$ must not violate the properties of regulation functions. This choice is thus constrained by the previous vertices $\mathcal{V}_r$ and the ones in $\mathcal{V}_e$.

Algorithm 1 DOMAIN EXTENSION

```
Input: Gene network with $G$ genes, regulation function
    $F$ composed of individual functions $f_i, i \in \{1, \ldots G\}$.
    Gene expression time series with $N+1$ time points.
Output: Gene network with updated regulation function $F$
for $j = 0$ to $N-1$ do
    for all gene $i \in \{1, \ldots G\}$ do
        if $j = 0$ then
            $\hat{x}_{i,0} = x_{i,0}$
        else
            Solve problem (14) to find the best value for
            $f_i(\hat{x}_{1,j}^i, \ldots, \hat{x}_{K_i,j}^i)$
            Refine grid $\mathcal{G}_i$ with the solution of problem (14)
            $\hat{x}_{i,j+1} = \hat{x}_{i,j} - \lambda_i \hat{x}_{i,j} + f_i(\hat{x}_{1,j}^i, \ldots, \hat{x}_{K_i,j}^i)$
        end if
    end for
end for
for all gene $i \in \{1, \ldots G\}$ do
    Assign new values to the vertices in $\mathcal{V}_f$
    Generate a new CNM function $f_i$ that satisfies the
    properties of $\mathcal{G}_i$
end for
return $F$
```

Our refinement method is summarized in Alg. 1. Briefly, we start the model predictions by initializing gene expression levels with the experimental values. For each gene i, we solve problem (14) at each time step j. The solution defines a new vertex $\mathcal{V}_r$ and a regulation value used to refine $\mathcal{G}_i$. We then update gene expressions and proceed to the next iteration. We assign at the end of the procedure values to each vertices $\mathcal{V}_f$ and construct a new CNM function f_i that satisfies the regulation hypothesis $\mathcal{R}_i$.

B. Gradient Descent

Our second optimization method focuses on the regulation values associated with the vertices in $\mathcal{V}_f$. Our goal is to assign regulation values that will minimize the error of the model $\varepsilon_\ell^2(F, \lambda_1, \ldots, \lambda_G)$. This is done by following a gradient descent procedure.

Gradient descent methods are used to find the local minima of a function by following the negative of its gradient. Since we cannot easily estimate the gradient of F, we randomly explore the space of all piecewise multi-affine functions that satisfy the regulation hypotheses $\mathcal{R}_1, \ldots, \mathcal{R}_G$. Random steps are taken in this space if they decrease $\varepsilon_\ell^2(F, \lambda_1, \ldots, \lambda_G)$.

Our gradient descent method is described in Alg. 2. Briefly, each iteration of the algorithm starts by creating a temporary regulation function F' from the original function F. Each individual regulation function $f_i'(\cdot)$ of F' is modified by changing the regulation value of a random number of vertices in $\mathcal{V}_f$. The altered function F' is saved if these adjustments reduce $\varepsilon_\ell^2(F, \lambda_1, \ldots, \lambda_G)$. Additionally, the number of vertices whose value is changed decreases with each iteration. This allows for "big" jumps at the

Algorithm 2 Gradient Descent Optimization

```
Input: Gene network with regulation function $F$ composed of individual functions $f_i, i \in \{1, \dots G\}$. Desired error $\varepsilon_\ell^{2,\max}$. Gene expression time series with $N+1$ time points.
Output: Gene network with optimized regulation function $F$
while $\varepsilon_\ell^2(F, \lambda_1, \dots, \lambda_G) > \varepsilon_\ell^{2,\max}$ do
    $F' = F$
    for all $f'_i$ in $F'$, $i \in \{1, \dots G\}$ do
        Modify the value of random number of vertices in $\mathcal{V}_f$. $f'$ must conform with $\mathcal{R}_i$
    end for
    if $\varepsilon_\ell^2(F, \lambda_1, \dots, \lambda_G) > \varepsilon_\ell^2(F', \lambda_1, \dots, \lambda_G)$ then
        $F = F'$
    end if
end while
return $F$
```

beginning of the optimization and more refined steps towards the end of the procedure. The algorithm terminates once the error reaches a specified threshold or converges.

V. Application

We applied our optimization techniques to several models of gene networks with known topology. The topologies were retrieved from the Transcriptional Regulatory Element Database (http://rulai.cshl.edu/TRED) [22], which contains gene networks for several families of transcription factors. We selected the AP1, NFκB, STAT, and p53 transcription factor families in mouse from this database. These transcription factors have all been associated with pathways of the immune system. The gene expression time series used throughout our application comes from a study by Amit *et al.* [23]. This work followed gene expression levels of mouse immune cells for five different conditions. Each condition corresponded to the triggering of immune signaling pathways by different agonists: PAM3CSK4 (PAM), polyinosine-polycytidylic acid [poly(I:C)], lipopolysaccharide (LPS), gardiquimod, and CpG. Each condition contained two replicates measured over nine time points. We used the first replicate of the LPS condition for identifying the topology of the different networks along with initial regulation functions F. The second replicate of this condition was used throughout the optimization procedures. This set of expression can be seen as identical to the first one with the addition of some noise.

We first re-identify the topology of the different networks starting from the list of genes present in each of them. We solved problem (7) for each gene using the first replicate of the LPS condition and allowed each gene to be regulated by up to four regulatory genes. The error $\|\varepsilon_i\|_F$ of the accepted regulation hypotheses did not exceed 10^{-9}. We obtained an initial regulation function F as a by-product of this topology identification procedure. The regulation hypotheses we identified were slightly different from the topologies present in the TRED database. This discrepancy may be due to the limit of regulatory genes we set. Genes in the AP1, NFκB, STAT, and p53 networks had up to eight regulators listed in the database. Increasing the maximum number of regulator genes is however computationally prohibitive, as each possible combination of regulatory genes is tested.

TABLE I: Error of the gene models after the different optimization procedures. DE: Domain Extension. GD: Gradient Descent. The Gradient Descent was performed 10 times (averages μ and standard deviations σ are shown in the table).

Networks	$\varepsilon_\ell^2(F, \lambda_1, \dots, \lambda_G)$ after optimization		
	None	DE	GD (μ / σ)
AP1	108.39	54.55	84.59 / 3.35
NFκB	96.65	76.22	81.75 / 1.99
p53	106.49	61.15	63.82 / 4.76
STAT	202.94	89.96	165.49 / 4.86

We proceeded to optimize F by following the techniques described in Sec. IV. We performed each method, *i.e.* Domain Extension and Gradient Descent, separately using the second replicate of the LPS condition. We compared each model (before and after optimization) to the experimental gene levels found in the second replicate. Each of the optimization procedure decreased the error $\varepsilon_\ell^2(F, \lambda_1, \dots, \lambda_G)$ of the different models (Table I). We performed our Gradient Descent method 10 times to determine if the final errors were consistent. None of these optimizations was able to decrease the error bellow the specified threshold ($\varepsilon_\ell^{2,\max} = 5$). Each procedure ended after the error did not change for 200 iterations (after ~ 400 iterations on average).

We present the gene expression time series of three genes from the p53 network (Ing4, Foxo3a, and Hnrpab) in Fig. 2. Each plot displays the expression levels found in the second replicate of the LPS condition and the ones obtained before and after each optimization technique. The models obtained before optimization performed poorly in predicting correct expression levels. The Domain Extension method was able to correct the regulation functions of Foxo3a and Hnrpab such that the gene dynamics matched between predicted and experimental time series. This method was however unable to correct the trajectory of Ing4. The dynamics of Ing was best corrected using the Gradient Descent optimization.

The Domain Extension method was better at decreasing the error of each model than the Gradient Descent, since this method estimates the best modifications to apply to f_i to decrease the error. In comparison, the Gradient Descent method searches the entire space of piecewise multi-affine functions that conform with all the regulation hypotheses. As the number of regulatory genes increases, the bigger this space becomes. However, the latter method has the advantage of not modifying the size of the original domain X^i. Assigning a value for the vertices in $\mathcal{V}_f$ for each f_i becomes computationally very expensive as the number of regulator genes increases.

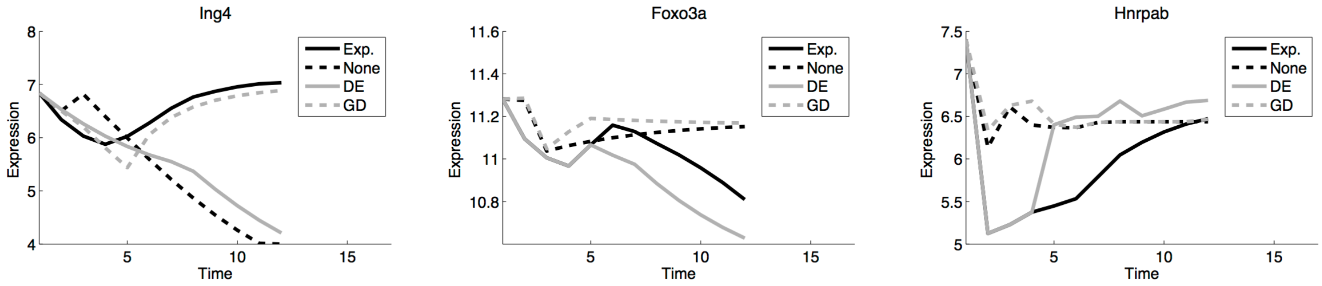

Fig. 2: Validation of predicted gene expression time series. The expression levels of three genes from the p53 network are displayed: Ing4, Foxo3a, and Hnrpab. Experimental time series (Exp.) are shown, along with the predicted expressions from the models obtained before optimization (None) and after the Domain Extension (DE) and Gradient Descent (GD) optimization techniques.

VI. Conclusion

In this paper, we presented methods that allow to generate an optimal gene network model after identifying the network topology. Our optimization procedures are based on identifying the best set of regulation functions, such that the ℓ_2 distance between predicted and experimental gene levels is minimized. Our first procedure refines the domains of the regulation functions and allows them to be more sensitive to the expression levels of regulatory genes. The second method explores the space of regulation functions and update the function parameters to decrease the error of the predicted gene dynamics. We applied both methods on networks with know topology and showed that the error of the models decreased after each optimization.

VII. Acknowledgments

The authors would like to thank Nicholas Cooper (Rensselaer Polytechnic Institute) for valuable discussions.

References

[1] R. Bonneau, D. J. Reiss, P. Shannon, M. Facciotti, L. Hood, N. Baliga, and V. Thorsson, "The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets *de novo*," *Genome Biology*, vol. 7, p. R36, 2006.

[2] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner, "Large-scale mapping and validation of *escherichia coli* transcriptional regulation from a compendium of expression profiles," *PLoS Biology*, vol. 5, no. 1, p. e8, 2007.

[3] M. Bansal, V. Belcastro, A. Ambesi-Impiombato, and D. di Bernardo, "How to infer gene networks from expression profiles," *Molecular Systems Biology*, vol. 3, no. 10.1038/msb4100120, 2007.

[4] E. Sontag, A. Kiyatkin, and B. N. Kholodenko, "Inferring dynamic architecture of cellular networks using time series of gene expression, protein and metabolite data," *Bioinformatics*, vol. 20, no. 12, pp. 1877–1886, 2004.

[5] M. K. S. Yeung, J. Tegner, and J. J. Collins, "Reverse engineering gene networks using singular value decomposition and robust regression," *Proc. of the National Academy of Science*, vol. 99, no. 9, pp. 6163–6168, 2002.

[6] T. S. Gardner, D. di Bernardo, D. Lorenz, and J. J. Collins, "Inferring genetic networks and identifying compound mode of action via expression profiling," *Science*, vol. 301, pp. 102–105, 2003.

[7] S. Drulhe, G. Ferrari-Trecate, and H. de Jong, "The switching threshold reconstruction problem for piecewise affine models of genetic regulatory networks," *IEEE Trans. Automatic Control*, vol. 53(1), pp. 153–165, 2008.

[8] R. Porreca, S. Drulhe, H. de Jong, and G. Ferrari-Trecate, "Structural identification of piecewise-linear models of genetic regulatory networks," *Journal of Computational Biology*, vol. 15, pp. 1365–1380, 2008.

[9] E. August and A. Papachristodoulou, "Efficient, sparse biological network determination," *BMC Systems Biology*, vol. 3, no. 25, 2009.

[10] M. M. Zavlanos, A. A. Julius, S. Boyd, and G. J. Pappas, "Identification of stable genetic networks using convex programming," in *Proc. American Control Conference*, Seattle, USA, 2008.

[11] A. A. Julius, M. M. Zavlanos, S. P. Boyd, and G. J. Pappas, "Genetic network identification using convex programming," *IET Systems Biology*, vol. 3, no. 3, pp. 155–166, 2009.

[12] M. M. Zavlanos, A. A. Julius, S. P. Boyd, and G. J. Pappas, "Inferring stable genetic networks from steady-state data," *Automatica*, vol. 47, no. 6, pp. 1113–1122, 2011.

[13] Y. Yuan, G. B. Stan, S. Warnick, and J. Goncalves, "Robust dynamical network reconstruction," in *Proc. IEEE Conf. Decision and Control*, Atlanta, GA., 2010.

[14] Y. Yuan, G. Stan, S. Warnick, and J. Goncalves, "Robust dynamical network structure reconstruction," *Automatica*, vol. 47, no. 6, pp. 1230–1235, 2011.

[15] Y. H. Chang and C. J. Tomlin, "Data-driven graph reconstruction using compressive sensing," in *Proc. IEEE Conf. Decision and Control*, Maui, Hawaii, 2012.

[16] A. A. Julius and C. Belta, "Genetic regulatory network identification using monotone functions decomposition," in *Proc. IFAC World Congress*, Milano, Italy., 2011.

[17] N. Cooper, A. A. Julius, and C. Belta, "Genetic regulatory network identification using multivariate monotone functions," in *Proc. IEEE Conf. Decision and Control*, Orlando, FL., 2011, pp. 2208–2213.

[18] G. Richard, H. Chang, I. Cizelj, C. Belta, A. A. Julius, and S. Amar, "Integration of large-scale metabolic, signaling, and gene regulatory networks with application to infection responses," in *Proc. IEEE Conf. Decision and Control*, Orlando, FL., 2011, pp. 2227–2232.

[19] H. Chang, G. Richard, A. A. Julius, C. Belta, and S. Amar, "An application of monotone functions decomposition to the reconstruction of gene regulatory networks," in *Proc. IEEE Int. Conf. Engineering in Medicine and Biology Society*, Boston, MA, 2011, pp. 2430–2433.

[20] R. Porreca, E. Cinquemani, J. Lygeros, and G. Ferrari-Trecate, "Identification of genetic network dynamics with unate structure," *Bioinformatics*, vol. 26, no. 9, pp. 123–1245, 2010.

[21] D. Angeli and E. Sontag, "Remarks on the invalidation of biological models using monotone systems theory," in *Proc. IEEE Conf. Decision and Control*, Maui, Hawaii, 2012.

[22] C. Jiang, Z. Xuan, F. Zhao, and M. Q. Zhang, "TRED: a transcriptional regulatory element database, new entries and other development," *Nucleic Acids Res*, vol. 35, no. Database issue, pp. D137–40, Jan 2007.

[23] I. Amit, M. Garber, N. Chevrier, A. P. Leite, Y. Donner, T. Eisenhaure, M. Guttman, J. K. Grenier, W. Li, O. Zuk, L. A. Schubert, B. Birditt, T. Shay, A. Goren, X. Zhang, Z. Smith, R. Deering, R. C. McDonald, M. N. Cabili, B. E. Bernstein, J. L. Rinn, A. Meissner, D. E. Root, N. Hacohen, and A. Regev, "Unbiased reconstruction of a mammalian transcriptional network mediating pathogen responses," *Science*, vol. 326, no. 5950, pp. 257–63, Oct 2009.