2011 50th IEEE Conference on Decision and Control and
European Control Conference (CDC-ECC)
Orlando, FL, USA, December 12-15, 2011

Genetic Regulatory Network Identification Using Multivariate Monotone Functions

Nicholas Cooper*, Calin Belta†, A. Agung Julius*

Abstract— **We present a method for identification of gene regulatory network topology using a time series of gene expression data. The underlying assumption is that the regulatory effects of a set of regulators to a gene can be described by a multivariate function. The multivariate function is constrained to be continuous, nonnegative and monotonic in each variable. We present necessary and sufficient conditions for the validity of the regulation hypothesis. Checking these conditions can be expressed as a Linear Programming feasibility problem. This paper builds on our previous work, where the regulation is described by a summation of multiple regulator functions, one function for each gene in the regulator set. Our procedure is two phased; the first identifies the correct set of regulators, the second uses the data and the regulator set to generate an appropriate regulator function. This paper focuses on the identification of the correct regulator set. As demonstration, we run our main algorithm on some experimental data from a synthetic gene network in yeast. We are able to show that the correct set of regulators is picked by the algorithm.**

I. Introduction

One of the main challenges in systems biology is *to identify the interaction topology among a set of genes based on their expression activities data*. Gene expression levels are typically measured as transcript concentrations with DNA microarray (c.f. [1], [2]). Due to the nature of the measurements, the data are typically organized as genome wide snapshots of gene expression activities.

Identification of Gene Regulatory Networks (GRNs) is a difficult problem because of several 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 fairly 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 profiles. They are based on clustering (e.g. [3]), information theoretic networks (e.g. [4]), Bayesian networks (e.g. [5]), and dynamical systems (e.g. [6], [7], [8]). The method that we present in this paper falls within the last category, and it is a generalization of our earlier work in [9].

Nicholas Cooper and A. Agung Julius are with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, Email: coopen3@rpi.edu, agung@ecse.rpi.edu. Calin Belta is with the Department of Mechanical Engineering, Boston University, Boston, MA 02215. Email: cbelta@bu.edu

With the availability of gene expression measurements, the network can in principle be reconstructed by inverting the data (c.f. [10]). However, as the measurements are noisy, the reconstructed network tends to be populated with spurious interconnections (i.e. false positives). This concern gives rise to sparse identification or parsimonious identification that aims at getting a network model with as few connections as possible without losing the fitness to the data (c.f. [6], [7], [1]).

Within the systems and control community, 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 (c.f. [11], [12]). Papachristodoulou *et al* developed a model for identification of sparse networks using Hill functions to describe the dynamics of gene-gene interaction (c.f. [13]). Earlier work by one of the authors of the current paper also aimed at identifying sparse networks based on genetic perturbation data, assuming that the dynamics can be described (locally) as a linear system (c.f. [14], [15]). A recent work by Yuan *et al* [16] investigated handling sparsity by using Akaike information criterion.

The method presented in this paper is a generalization of our earlier work in [9], which proposed a different approach from those in the above references. Although we still use the dynamical system formulation, there is no assumption made about the type of the functions used in the regulation (e.g. linear, polynomial, Hill functions, etc.). The only imposed assumption is that the interaction dynamics can be represented as **continuous nonnegative monotonic** functions (in short, CNM functions). The network structure is built by identifying the set of regulators for each gene. The regulators of a gene X are the genes that directly[1] regulate the expression activity of X. Our method is essentially based on model invalidation, rather than model identification. Along with a validation of a regulator set, this method can give insight into how strong the regulator set fits the data.

A recent work by Porreca et al, published earlier in [17], 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 very similar to our approach, in the sense that it separates the issues of network topology and the functional/parametric representation of the dynamics.

[1]Note that by *directly* we mean without going through other genes in the network under study. Therefore, this is not necessarily a statement about the binding of transcription factors to certain promoters.

However, our approach differs from the one in [17] in that, (i) we propose a discrete-time model structure that is directly derived from the time series data, (ii) we prove that the conditions that we use to reject some network topologies are both necessary and sufficient (see Theorem 4), while in [17], no such result is derived, (iii) we present some theoretical analysis on irrefutable models and data.

The main result of this paper is as follows. For any proposed set of regulators, we derive a necessary and sufficient condition for the data to be compatible with the regulation hypothesis. The major difference between this paper and previous work [9] is that in the current paper, the regulator functions can be any general multivariate CNM functions. In [9], we restrict our attention to regulator functions that can be written as sums of CNM functions (one for each variable). The significance of this generalization can be demonstrated by examining the regulator functions found in [18]. The function represents the regulation of the lacZYA operon of *Escherichia coli*, dependant on cAMP and isopropylβ-D-thiogalactoside (IPTG). The regulator function from [19],

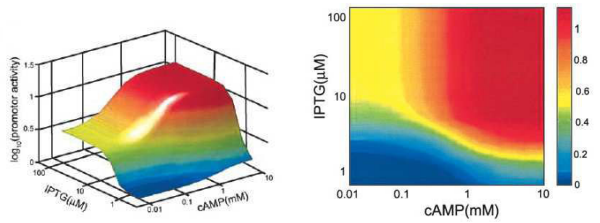

Fig. 1: Figure taken from [18], for the wild type regulator function

[18] is of the form

$$f(x,y) = V_1 \frac{1 + V_2\mathcal{A}(x) + V_3\mathcal{R}(y)}{1 + V_4\mathcal{A}(x) + V_5\mathcal{R}(y)} \tag{1}$$

where V_i are constant parameters, $\mathcal{R}(y)$ and $\mathcal{A}(x)$ are Hill functions in cAMP and IPTG respectively. As Hill functions, $\mathcal{R}(y)$ and $\mathcal{A}(x)$ are CNM functions. However, we can see that the function $f(x,y)$ cannot be expressed as the sum of CNM functions in x and y. By taking the log of $f(x,y)$, we can also see that $f(x,y)$ cannot be expressed as the product of CNM functions in x and y either. Yet as can be seen in Figure 1 the function is monotonic in both variables. Therefore the method presented in this paper can allow for a multivariate function similar to f, whereas the previous method in [9] could not.

The conditions for the validity of a regulation hypothesis in this paper is, as in previous methods, formulated as a Linear Programming feasibility problem (in the implementation, we apply a quadratic cost on the slack variables), which is computationally tractable. By verifying the compatibility of the regulation hypothesis with the data, we can invalidate a proposed model structure. Since the method isolates one gene (and its regulators) from the rest of the network, it allows for each individual gene's regulator set to be determined in parallel. Another departure from the previous work is that determining a regulator set does not give the entire regulator function(s). It is shown in Section IV that the method assigns values in a grid for the regulator function, and further work is required to 'fill in' the rest of the function. We also discuss irrefutability results for certain network topologies and time series data. This result can be used in, e.g. determining whether the data is rich enough to separate different model structures.

II. Mathematical Models for Gene-Gene Interaction

Assume that the GRN we are working with consists of G genes, and that there is a sequence of expression activity for $(N+1)$ time points for every gene. Denote the expression data for Gene i at time j as $x_{i,j}, 1 \le i \le G, 0 \le j \le N$. The (time) differential expression activity $q_{i,j}$, $1 \le i \le G$, $0 \le j \le N-1$, is defined as

$$q_{i,j} \triangleq x_{i,j+1} - x_{i,j} \tag{2}$$

Remark 1: We implicitly assume that $\Delta t = t_{j+1} - t_j$ is the same $\forall j$. If this is not true, then some interpolation technique can be used, for example by normalizing $q(i)$ to $q(i)' = q(i)\frac{t_{i+1}-t_i}{\Delta_t}$, where Δ_t is fixed and $q(i)$ is from (2).

The interaction between genes can be captured by the following model:

$$q_{i,j} = -\lambda_i x_{i,j} + f_i(x_{k_1,j}, x_{k_2,j}, \ldots, x_{k_K,j}) \tag{3}$$

where $\lambda_i \ge 0$ is the decay parameter of Gene i, $k_1, \cdots, k_K \in G_i^R \subseteq \{1, \ldots, G\}$ are the K regulators of Gene i, and f_i is the function that describes the regulation of Gene i based on the expression activities of all the genes in G_i^R. Both the λ_i and the f_i are unknown, and the objective of this paper is to identify them based on the experimental data.

The following assumption for the regulatory function $f_i(\cdot)$ is adopted.

CNM Assumption: The function $f_i(x_{k_1}, x_{k_2}, \ldots, x_{k_K})$ is continuous, nonnegative and monotonic (CNM) in each $k_1, \cdots, k_K \in G_i^R$.

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

$$G_i^R = G_i^{R+} \cup G_i^{R-} \tag{4}$$

where G_i^{R+} and G_i^{R-} are the sets of activators and repressors of Gene i, respectively. The number of regulators of a gene is called the *in-degree* of the gene, a term taken from the graph-theoretic interpretation of the GRN (see Figure 2).

Remark 2: The assumption that f is continuous, nonnegative and monotone is very general. Virtually all phenomenological regulation models that have been proposed to represent gene-gene interaction (Hill functions, sigmoid functions, linear functions, piecewise affine functions, etc.,

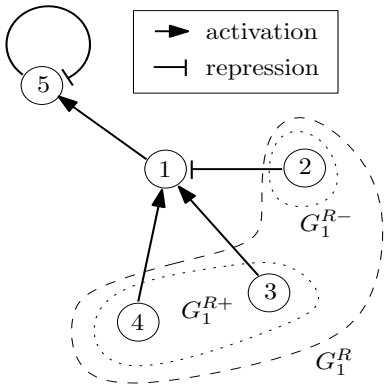

Fig. 2: (From [9]) A simple GRN. The set of regulators of Gene 1, G_1^R, is {2,3,4}. The set of activators of Gene 1, G_1^{R+}, is {3,4}, while the set of repressors, G_1^{R-} is {2}. The in-degree of Gene 1 is 3. Notice that self-loop is allowed. For example, in this network, Gene 5 is a repressor of itself.

see references in Section I) are captured in this broad class of functions. On the other hand, we can argue that it is not possible to generalize this assumption further. For example, without the monotonicity assumption, the notions of activation and repression do not make any sense. The nonnegativity assumption is adopted because virtually all known gene-gene interaction models have this property. The nonnegativity assumption can be replaced, if the λ term in (2) is dropped. With this assumption of negative autoregulation of Gene i the decay can be captured without the use of λ.

Remark 3: Notice that we do not preclude the possibility of any gene acting as activator and repressor for two different genes.

III. Sparsity Consideration

The sparsity of a GRN will be characterized by the in-degree of genes (nodes) of the network. The physical interpretation is that the sparsity of the network corresponds to the number of regulators for a gene in the network. The total regulatory network interconnection is the sum of the in-degrees of all the genes. The characterization of sparsity in this paper is in line with prior work [14], where the sum of in-degree's would be equal to the cardinality of the A matrix (5) from [14].

The sparsest regulator set for any gene in a GRN is probably not unique. For instance, suppose that in a GRN with $X, Y, Z \in G$, q_X can be expressed as a function of the expression activities of Y or Z. The regulator sets of $\{Y\}$ and $\{Z\}$ are the smallest regulator set of Gene X. We consider minimality of a set of regulators in the sense of partial ordering of sets generated by set inclusion. A set of regulators G_i^R is considered smaller than another set $G_i^{R'}$ if $G_i^R \subset G_i^{R'}$. For each gene the minimal sets of regulators can be identified by performing a breadth-first search on the lattice structure of the power set of all genes in the GRN $\{1, 2, \ldots, G\}$, as shown in Figure 3.

The lattice structure shows the set of regulators, but there is also the added complication that each regulator can be

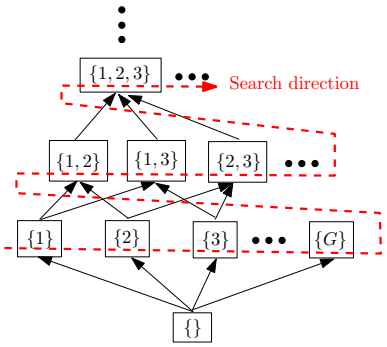

Fig. 3: (From [9]) The lattice structure of the power set of $\{1, \ldots, G\}$. The arrows indicate set inclusion, and the direction of the breadth-first search is given by the dashed red line.

either an activator or a repressor. This yields added complexity, e.g. the gene sets $\{1\}, \{1, 2\}$ have two $(\{1+, 1-\})$, or four $\big(\{\{1+, 2+\}, \{1+, 2-\}, \{1-, 2+\}, \{1-, 2-\}, \}\big)$ regulation hypotheses. A simple way of verifying all regulation hypotheses is to verify regulator sets in order of the breadth-first search. Although in the worst case we need to verify all sets, in practice this is not the case, because:

- Most GRN's found in nature have a small in-degree for all genes.
- If a subset $G_i^R \subset \{1, \ldots, G\}$ is verified as a set of regulators, all larger sets $G_i^{R'}, G_i^R \subset G_i^{R'}$ are also verifiable regulator sets. This is because a CNM function in X is also a CNM function in X and Y, where X and Y are two disjoint sets of variables.

IV. Main Result

Given the expression activity data for each Gene i at time j is given by $x_{i,j}$, $1 \leq i \leq G$, $0 \leq j \leq N$ the temporal differential expression $q_{i,j}$ is computed using (2) for $1 \leq i \leq G$, $0 \leq j \leq N-1$.

Regulation Hypothesis: A regulation hypothesis $\mathcal{R}$ is parameterized by (an ordered pair of) two disjoint subsets $G_k^{R+}, G_k^{R-} \subseteq \{1, \ldots, G\}$, which are the sets of activators and repressors of Gene k, respectively. Thus, $G_k^R := G_k^{R+} \cup G_k^{R-}$ is the set of regulators of Gene k.

Notation 1: Given a regulation hypothesis $\mathcal{R} = (G_k^{R+}, G_k^{R-})$, we denote:

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

The regulation hypothesis $\mathcal{R}$ is equivalent to the statement that the temporal differential expression satisfies the following relation:

$$q_{k,j} = -\lambda_k x_{k,j} + f(x_{\mathbf{a},j}, x_{\mathbf{r},j}), \ \forall j \in \{0, \cdots, N-1\} \quad (5)$$

for some $\lambda_k \geq 0$, and a CNM function $f(x_{\mathbf{a},j}, x_{\mathbf{r},j})$ that is monotonically increasing in the variables in $x_{\mathbf{a},j}$ and monotonically decreasing in the variables in $x_{\mathbf{r},j}$.

Before we proceed to the main result of this paper, let us define the following

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

$$x_{\mathbf{r},i} \preceq_{\mathcal{R}} x_{\mathbf{r},j} :\Leftrightarrow \begin{cases} x_{\mathbf{a},i} \leq x_{\mathbf{a},j}, \\ x_{\mathbf{r},i} \geq x_{\mathbf{r},j}. \end{cases} \tag{6}$$

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

The following theorem is used to (in)validate the regulation hypothesis above, based on the experimental data.

Theorem 4: The regulation hypothesis $\mathcal{R}$ is true, i.e. there exist λ_k and a CNM function $f(\cdot)$ such that (5) is satisfied if and only if there exist λ_k and $\hat{q}_{k,j},\ j \in \{0,\ldots,N-1\}$ such that the following linear constraints are satisfied:

$$q_{k,j} = -\lambda_k x_{k,j} + \hat{q}_{k,j},\ \forall j \in \{0,\ldots,N-1\}, \tag{7}$$

$$\lambda_k \geq 0, \tag{8}$$

$$\hat{q}_{k,j} \geq 0,\ \forall j \in \{0,\ldots,N-1\}, \tag{9}$$

$$\hat{q}_{k,j} \geq \hat{q}_{k,i},\ \forall (i,j) \in \mathcal{S}_{\mathcal{R}}. \tag{10}$$

Note that $x_{k,j}$ and $q_{k,j},\ j \in \{0,\ldots,N-1\}$, are obtained from experimental data.

Proof: Omitted because of space limitation. ■

Remark 5 (Complexity): The validation of the regulation hypothesis, as described in Theorem 4, amounts to solving a Linear Programming (LP) feasibility problem with $(N+1)$ variables, N equality constraints, and $(N+1+|\mathcal{S}_{\mathcal{R}}|)$ inequality constraints. Here, $|\mathcal{S}_{\mathcal{R}}|$ denotes the cardinality of the set $\mathcal{S}_{\mathcal{R}}$.

The GRN identification algorithm provided in this paper is summarized in Algorithm 1. Note that this algorithm results in the minimum regulator set for a gene, and does not explicitly compute the multivariate CNM regulator function itself.

Algorithm 1 Computation of the sparsest gene network based on gene expression data time-series

Require: Time-series data of gene expression activities $x_{i,j}$.

Compute the time differential expression data $q_{i,j}$.
for all $k \in \mathcal{G} \triangleq \{1,\ldots,G\}$ **do**
  Compute the lattice of the subsets of $\mathcal{G}$ as in Figure 3. Label every subset with 'unverified'.
  repeat
    Take a subset of $G_k^R \subset \mathcal{G}$ with label 'unverified'.
    for all possible activator-repressor partitioning of G_k^R **do**
      Use Theorem 4 to verify whether G_k^R is a set of regulators for Gene k.
    end for
    if G_k^R is a set of regulators for Gene k **then**
      Label G_k^R and all of its upper bounds in the lattice with 'pass'.
    else
      Label G_k^R with 'fail'.
    end if
  until all subsets are labeled with 'pass' or 'fail'.
  The possible sets of regulators for Gene k are all subsets labelled with 'pass'. The minimal sets of regulators are the minimal elements (in the sense of set inclusion) of this set.
end for

V. Implementation and Example

In the implementation, the experimental data are noisy. We therefore modify the algorithm as follows. We introduce slack variables ϵ, and instead of a feasibility problem as described in Theorem 4, we solve a minimization problem. The objective is to minimize the Frobenius norm $\|\epsilon\|_F$, leading to a Linear Quadratic (LQ) programming problem, as shown below:

$$\min \|\epsilon\|_F \text{ subject to} \tag{11}$$

$$q_{k,j} = -\lambda_k x_{k,j} + (\hat{q}_{k,j} + \varepsilon_{k,j}),\ \forall j \in \{0,\ldots,N-1\}, \tag{12}$$

$$\lambda_k \geq 0, \tag{13}$$

$$(\hat{q}_{k,j} + \varepsilon_{k,j}) \geq 0,\ \forall j \in \{0,\ldots,N-1\}, \tag{14}$$

$$(\hat{q}_{k,j} + \varepsilon_{k,j}) \geq (\hat{q}_{k,i} + \varepsilon_{k,i}),\ \forall (i,j) \in \mathcal{S}_{\mathcal{R}}, \tag{15}$$

with λ_k, $\epsilon_{k,j}$, and $\hat{q}_{k,j},\ j \in \{0,\ldots,N-1\}$ as the optimization variables. The optimal cost function can be interpreted as the least amount of perturbation of the time differential expression data required to 'fit' the data with the regulation hypothesis. The optimal cost function is zero if and only if the LP problem of 4 is feasible. Due to the inherent noise in the experimental data, we would allow for a small nonzero cost in accepting the regulation hypothesis.

Remark 6 (Complexity): The numerical implementation above involves solving an LQ problem with $(2N+1)$ variables, a quadratic cost involving N variables, N equality constraints, and $(N+1+|\mathcal{S}_{\mathcal{R}}|)$ inequality constraints.

We apply our algorithm on the experimental data of the IRMA gene network [20]. The IRMA network consists of five genes, with topology shown in Figure 4. It is a synthetic network constructed in *Saccharomyces cerevisiae*. Data from the IRMA network is generated from two different perturbation experiments, the 'switch-on' and 'switch-off' experiments. In the 'switch-on' experiments the cells are shifted from glucose to galactose, the 'switch-off' experiment the cells are shifted opposite, from galactose to glucose. The expression profiles of each gene were obtained from quantitative real time PCR (RT-PCR) analysis. The 'switch-on' experiments were sampled every 20 minutes for a period of 5 hours, and the 'switch-off' experiments were sampled every 10 minutes for a period of 3 hours. The expression data was expressed by the Δct method, where the mean of the times series is subtracted from each data point. We examine the data from these two experiments separately, due to their characteristic differences.

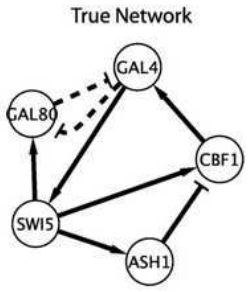

Fig. 4: Network Topology of the IRMA Network, from [20]

We test our algorithm in the identification of the regulator set of CBF1, one of the genes in the IRMA network. CBF1 is picked because it is the only gene in the network with multiple regulators. The breadth-fist search is carried out for the first two levels (all single and pairwise regulator sets) for both the 'switch-on' and 'switch-off' experiments. The 'on' experiment has five replicates, the 'off' experiment has four replicates. The minimization algorithm is run for every replicate, and the average ϵ is taken, for the switch off and switch on experiments. The regulator sets with the six smallest errors for each experiment type are shown in Table I

switch-on		switch-off	
Reg. Set	$\|\varepsilon\|_F$	Reg. Set	$\|\varepsilon\|_F$
(SWI5+,ASH1-)	0.0071	(SWI5-,ASH1+)	0.0342
(SWI5+,GAL4-)	0.0074	(ASH1+,GAL4-)	0.0433
(SWI5+,GAL4+)	0.0078	(SWI5-,GAL80+)	0.0508
(ASH1+,GAL4+)	0.0081	(GAL4-,GAL80+)	0.0519
(SWI5+,GAL80+)	0.0093	(ASH1-,GAL80+)	0.0523
(SWI5+,GAL80-)	0.0103	(ASH1+,GAL80+)	0.0523

TABLE I: Results from the identification of the IRMA network. The correct regulator set is (SWI5+, ASH1-)

The notation used in Table I for the regulator sets can be explained as follows. Positive (+) sign denotes activator gene, while negative (-) sign denotes repressor gene. For example, the set (SWI5+, ASH1-) is the regulator set with gene SWI5 acting as an activator and gene ASH1 acting as a repressor. We can observe that the actual regulator set (SWI5+, ASH1-) came out as the best regulation hypothesis (corresponding to the smallest cost function $\|\varepsilon\|_F$) when the switch-on data set is used. The switch-off data set leads to worse fit (corresponding to higher cost function, even in the best hypothesis) and incorrect predictions.

Note that CBF1 does not appear in any of the regulator sets, as it has been excluded from the computation from self regulation. The reason for avoiding autoregulation as an activator is given in the next section. We also avoid having autoregulation as a repressor because a part of this type of regulation can be absorbed into the λ_k term. Alternatively the λ_k term could be dropped and the self repression added; however this would require that the nonnegativity constraint in the CNM function is dropped.

VI. Refutability of Regulation Hypothesis

As discussed earlier, the method described is designed to reject different regulation hypotheses for each gene in the network. In this section we discuss irrefutable regulation hypothesis, i.e. the conditions where: (i) the regulation hypothesis is such that it cannot be refuted, regardless of the data, or (ii) the expression activity patterns of some genes are such that they cannot be refuted as regulators of any target gene regardless of the target gene's expression data. The results that are presented here are generalization of similar results presented in [9].

A. Autoactivation

Any regulation hypothesis involving a gene acting as its own activator, in practice, cannot be refuted. To prove that statement, we proceed with the following lemma.

Lemma 7: [9] Assume we have a data set consisting of gene expression data $x_n,\ n \in \{0,\dots,N\}$ and the corresponding time-differential expression data $q_n,\ n \in \{0,\dots,N-1\}$ satisfying $x_n \geq 0$, and $x_i \neq x_j$ if $i \neq j$. There exist $\lambda \geq 0$ and $f(\cdot)$ a CNM increasing function such that

$$q_n = -\lambda x_n + f(x_n), n \in \{0,\dots,N-1\}. \quad (16)$$

The assumptions that we impose in this lemma, i.e. the uniqueness and nonnegativity of the expression data are always satisfied in practice. In particular, because of the presence of measurement noise, the probability of having two measurements with exactly the same number is zero.

Lemma 7 states that the regulation hypothesis consisting of one regulator gene, namely a gene activating itself cannot be refuted in practice. Using the last bullet in Section III, we can extend this result to cover any regulation hypothesis involving autoactivation.

B. Empty Constraint Set $\mathcal{S}_\mathcal{R}$

The constraint set $\mathcal{S}_\mathcal{R}$ in Theorem 4 plays a crucial role in (in)validating a regulation hypothesis.

Lemma 8: If the constraint set $\mathcal{S}_\mathcal{R}$ in Theorem 4 is empty, then the regulation hypothesis is true (i.e. regardless of the time differential expression data).

Proof: In this case, the constraint in (10) is satisfied by default. Thus, we only need to show the existence of λ_k and $\hat{q}_{k,j},\ j \in \{0,\dots,N-1\}$ such that the constraints (7) - (9) are met. For this, we can choose any $\lambda_k \geq 0$ and pick

$$\hat{q}_{k,j} = q_{k,j} + \lambda_k x_{k,j}. \quad (17)$$

■

An interesting consequence of this result is that when the time series of the expression activities of two genes are monotonic, they can be used as universal regulators. That is, they cannot be refuted as regulators of any target gene regardless of the target gene's expression data. This is because we can always construct a regulation hypothesis $\mathcal{R}$ involving these two genes, corresponding to an empty constraint set $\mathcal{S}_\mathcal{R}$. To demonstrate this assertion, consider

two genes A and B, whose expression activity data $x_{A,n}$ and $x_{B,n}$, $n \in \{0, \ldots, N-1\}$ are monotonically increasing. Then, both regulation hypotheses $(A+, B-)$ and $(A-, B+)$ result in an empty constraint set $\mathcal{S}_{\mathcal{R}}$.

VII. Conclusion

We present some theoretical results that are applicable in the identification of genetic regulatory network (GRN) based on gene expression time series. Our goal is in obtaining a discrete-time (nonlinear) dynamical system model of the GRN. The core of our contribution lies in separating the issue of identifying the network topology from identifying the regulation functions.

The basis of our approach in identifying the topology of the GRN lies in the formulation of the regulatory relation between genes in an axiomatic manner. This is spelled out in the CNM assumption in Section II. We also argue that the CNM assumption that characterizes the regulatory relation is very general and satisfied by virtually all existing dynamical system models of GRN. Further, we show that for any given regulation hypothesis and data set, the validity of the CNM assumption can be checked as a Linear Programming (LP) feasibility problem. To account for noisy data, we also formulate a Linear Quadratic (LQ) optimization problem that measures how far the data are from satisfying the regulation hypothesis. The computational implementation of our results is thus based on convex optimization. Another nice feature of our approach is that the determination of the regulators of all the genes in the network can be performed in a completely parallel fashion.

Once the network topology is obtained, in principle, the regulation functions can be determined by (i) interpolation of the data points, or (ii) fitting the data points with functions from a chosen class. In particular, the CNM assumption guarantees the existence of CNM regulation functions that interpolate the data. Because of space limitation, we do not demonstrate the interpolation step in this paper.

Acknowledgment

This work was supported in part by an R01 grant from the National Institute of Health (NIH) number DE15989. The authors would like to thank Dr. Salomon Amar, Niraj Trivedi, and Guilhem Richard (all at Boston University) for the fruitful discussions leading to the development of the results presented in this paper.

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. Eisen, P. Spellman, P. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc Natl Acad Sci USA*, vol. 95, pp. 14 863–14 868, 1998.

[4] K. Basso, A. Margolin, G. Stolovittzky, and U. Klein, "Reverse engineering of regulatory networks in human b cells," *Nat Genet.*, vol. 37, pp. 382–390, 2005.

[5] F. Geier, J. Timmer, and C. Fleck, "Reconstructing gene-regulatory networks from time-series, knock-out data and prior knowledge," *BMC Systems Biology*, vol. 1, no. 11, February 2007, online publication.

[6] 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.

[7] 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.

[8] J. Tegner, M. K. S. Yeung, J. Hasty, and J. J. Collins, "Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling," *Proc. of the National Academy of Science*, vol. 100, no. 10, pp. 5944–5949, 2003.

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

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

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

[17] 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.

[18] Y. Setty, A. E. Mayo, M. G. Surette, and U. Alon, "Detailed map of a cis-regulatory input function," *Proc. National Academy of Science*, vol. 100, pp. 7702–7707, 2003.

[19] U. Alon, *An Introduction to Systems Biology*. Chapman and Hall, 2007.

[20] I. Cantone et al, "A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches," *Cell*, vol. 137, pp. 172–181, 2009.