12-9-2008

# Determining interconnections in biochemical networks using linear programming

Elias August
*University of Oxford*

Antonius Papachristodoulou
*University of Oxford*

Ben Recht
*California Institute of Technology*

Mark Roberts
*University of Oxford*

Ali Jadbabaie
*University of Pennsylvania*, jadbabai@seas.upenn.edu

# Determining interconnections in biochemical networks using linear programming

**Abstract**
We present a methodology for efficient, robust determination of the interaction topology of networked dynamical systems using time series data collected from experiments, under the assumption that these networks are sparse, i.e., have much less edges than the full graph with the same vertex set. To achieve this, we minimize the 1-norm of the decision variables while keeping the data in close Euler fit, thus putting more emphasis on determining the interconnection pattern rather than the closeness of fit. First, we consider a networked system in which the interconnection strength enters in an affine way in the system dynamics. We demonstrate the ability of our method to identify a network structure through numerical examples. Second, we extend our approach to the case of gene regulatory networks, in which the system dynamics are much more complicated.

**Comments**
August, E.; Papachristodoulou, A.; Recht, B.; Roberts, M.; Jadbabaie, A., "Determining interconnections in biochemical networks using linear programming," 47th IEEE Conference on Decision and Control, 2008. (CDC 2008), pp.3311-3316, 9-11 Dec. 2008 URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=4739286&isnumber=4738560

Proceedings of the
47th IEEE Conference on Decision and Control
Cancun, Mexico, Dec. 9-11, 2008



# Determining Interconnections in Biochemical Networks Using Linear Programming

Elias August, Antonis Papachristodoulou, Ben Recht, Mark Roberts and Ali Jadbabaie

*Abstract*— **We present a methodology for efficient, robust determination of the interaction topology of networked dynamical systems using time series data collected from experiments, under the assumption that these networks are sparse, i.e., have much less edges than the full graph with the same vertex set. To achieve this, we minimize the 1-norm of the decision variables while keeping the data in close Euler fit, thus putting more emphasis on determining the interconnection pattern rather than the closeness of fit. First, we consider a networked system in which the interconnection strength enters in an affine way in the system dynamics. We demonstrate the ability of our method to identify a network structure through numerical examples. Second, we extend our approach to the case of gene regulatory networks, in which the system dynamics are much more complicated.**

## I. INTRODUCTION

Determining the interaction topology in large-scale dynamical systems has been an active area of research for some time now. Most available results in the case of high-throughput experimental data concern information about the behavior of the system after small perturbations from the steady-state. In this case, several approaches have been considered [1]–[3]. However the problem of determining the network structure in the case where time-series data are available is much harder and we address this case in this paper.

A particular example of an area of research in which the above problem is of fundamental importance is molecular biology. One aims to robustly determine the interaction topology of biochemical networks using time series data collected from experiments. On one hand, such data are often abundant due to the development of high-throughput, effective experimental techniques. At the same time, a high computational effort is required to extract information about the network structure; moreover these data are often noisy and do not contain rich information. In particular, determining the pathways in biochemical reaction networks and gene regulatory networks from time series data has been an active area of research for over a decade. A recent review of available techniques can be found in [4] or [5], but earlier articles, such as [6], also list several approaches to this network identification problem.

Apart from these, in [7], necessary and sufficient conditions are presented for the ability to reconstruct the network structure of linear dynamical systems from input-output data only. A class of techniques that fall under the rubric of 'stationary state Jacobian Matrix Elements' analyzes the system behavior when it is perturbed locally from steady-state and look at whether the concentration of one species is increased or decreased when another species concentration is increased. In [8] and [9], Kholodenko et al have built on this approach and determined the functional interactions in cellular signaling and gene networks by taking into account the 'modular' structure of the networks in question. Alternatively, inferences about the topology of the network can be made by introducing pulse changes in concentration of a chemical species in the network, and observing the networks response, concluding causal chemical connectivities [10]. In [11], an approach was presented to apply linear programming to minimize the $L_1$-norm such as to obtain the sparsest interaction structure in the case of chemical reaction networks. In [3], a linear dynamical system was considered to represent a gene regulatory networks, and an approach proposed to minimize the $L_1$-norm in order to obtain the sparsest network structure form genetic perturbation experiments.

A variety of data-driven approaches attempt to extract structure from existing experimental data without the ability to tailor experiments to the modeling task. For example, researchers have used time series measurements of concentrations to construct correlation functions of concentrations [12]. An approach using Artificial Neural Networks [13] tries to 'learn' patterns from the complicated and noisy data and to detect trends in the chemical reaction pathways. Related to this is a genetic algorithm approach to study the evolutionary development of a reaction mechanism [14]. In [1], the Singular Value Decomposition was used to obtain a family of candidate networks. Since the optimal networks were typically much more dense than would be realistically expected, the sparsest network in the family was identified using robust regression. In [15], the Sparse Vector Autoregressive method was applied to estimate gene regulatory networks for cases when gene interactions are sparse and experimental data are rare.

This paper contains two results. The first part of our study

This work was financially supported by EPRSC project E05708X.

E. August and A. Papachristodoulou are with the Department of Engineering Science, University of Oxford, Parks Road, Oxford OX1 3PJ, UK. {elias.august,antonis}@eng.ox.ac.uk

B. Recht is with the Centre for the Mathematics of Information, California Institute of Technology, 1200 E. California Blvd, MC 136-93, Pasadena CA 91125, USA. brecht@caltech.edu

M. Roberts is with the Department of Engineering Science and the Department of Biochemistry, University of Oxford, Parks Road, Oxford OX1 3PJ, UK. mark.roberts@bioch.ox.ac.uk

A. Jadbabaie is with the Department of Electrical and Systems Engineering, University of Pennsylvania, 365 GRW Moore Bldg, 200 South 33rd Street, Philadelphia, PA 19104, USA. jadbabai@seas.upenn.edu

focuses on dynamical systems of the following form

$$\dot{x} = Af(x),\ x \in \mathbb{R}^n,\ A \in \mathbb{R}^{n\times m}, \tag{1}$$

where the unknown matrix is $A$ and functions $f$ (which need to satisfy appropriate smoothness conditions to ensure local existence and uniqueness of solutions) are known. This makes (1) linear in the unknown parameters, which is not a significant assumption as many such modelling frameworks are widely used in practice. For example, *chemical reaction networks* with *mass action kinetics* (see references [16] and [17]), are systems that have such a structure. Our main objective in such a procedure is to identify the interconnection topology that is encapsulated in $A$, given experimental time-series data. In the particular case of biochemical reaction networks, we seek to identify the chemical pathways and mechanisms, that is, how the chemical complexes interact within the chemical network. This was the topic of an earlier paper [11] where it was argued that identifying the interconnection topology in biological and biochemical systems is of greater importance than extracting the parameters (the rates of the various reactions) that best fit the particular time series data. There are several reasons for this: first, the parameters are often collected under noisy experimental conditions and are sensitive to laboratory conditions such as temperature and the environment. Second, as is often the case with large networks, persistence of observed phenomena is robust to a large range of most parameter values and therefore identifying these parameters exactly is not of great interest. Indeed, chemical reaction networks often have the same functionality in the neighborhood of most of the nominal reaction rates. But most importantly, networks are rarely robust to the random rewiring of the underlying interconnection structure and hence determining the network structure is much more important than determining the kinetic parameters that fit the particular data. An important property of the network given by $A$ is sparseness, i.e., it has much less edges than the full graph with the same vertex set. In this paper we first extend the results in [11] to general and large-scale networks; moreover, we put more emphasis on the case when data from measurements is rare. As highlighted in the paper cited, the importance here is that a linear program can be solved efficiently while searching for the sparsest network that fits data is a combinatorial problem.

In the second part of the paper, we draw our attention to models of gene regulatory networks. A gene encodes the information necessary to produce a specific protein. The process, in which the information is used to synthesize a protein, is highly controlled and this control allows the cell to vary the level of a particular protein in the cell depending on the cell's need for this protein. The first step of synthesizing a protein from a gene is RNA polymerase transcribing gene information from DNA to mRNA (see Figure 1a). This mRNA is then translated into synthesised proteins by ribosomes. Control can occur at a number of stages including the synthesis of mRNA, subsequent processing of the mRNA, control of the ribosome and control of mRNA stability. Some proteins, called transcription factors, are responsible for the regulation of the initiation of transcription. A transcription factor binds to the DNA, at the promoter site, in order to either inhibit or activate (or alternatively increase the rate of) the transcription of mRNA that is responsible for the synthesis of a specific protein (see Figure 1b). (Note that self regulation is also possible.) The collection of DNA segments which interact with each other in the manner described is called the gene regulatory network.

The three main information levels that need to be identified to understand the dynamics and behavior of a gene regulatory network are:

1) The network of connections in form of a *directed graph*;
2) Whether an edge from node $i$ to node $j$ means that transcription factor $i$ is activating or repressing $j$;
3) What are the activation/repression rates for the transcription factors.

Time-series obtained from DNA microarrays [18], [19] are extremely helpful to obtain the structure of a gene regulatory network. This is because DNA microarrays allow observation of the presence of specific mRNA within the cell and in particular, time-series data allow measuremens on how these change over time after a perturbation, or when following the cell cycle. We provide an approach using Linear Programming to obtain the gene regulatory network structure from DNA microarray time-series data.

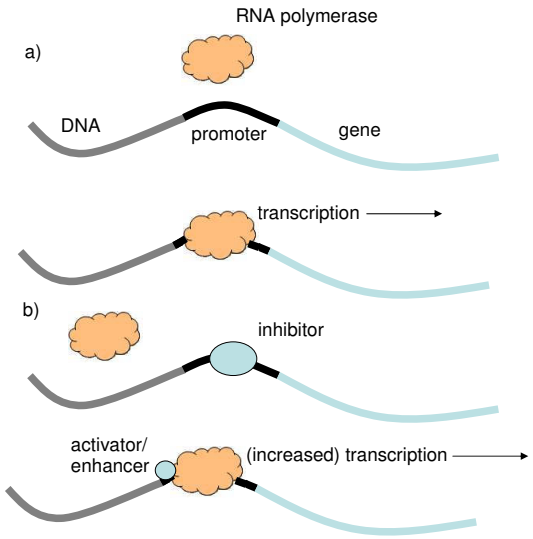


Fig. 1. Diagram showing the process of transcription. 1a) The RNA polymerase binds to the promoter sequence on the DNA and transcribes a gene. 1b) Transcription can be controlled by inhibitors or activators acting at the promoter sequence.

The paper is organized as follows. In Section II, we describe an algorithm to obtain the network structure (matrix $A$ in (1)) of a dynamical system with affine and sparse interconnections. Considering a linear dynamical system, we provide an example utilising our method and evaluating it. We then consider the more complicated case of a gene regulatory system in Section III, where the dynamics are not

affine in the unknown parameters, show how to approach this case, and provide examples. Finally, we conclude the paper and suggest future work in Section IV.

### A. Notation

| | |
|---|---|
| $\mathbb{R},\ \mathbb{R}^n,\ \mathbb{R}^{m\times n}$ | real numbers, real vector of length $n$, $m \times n$ real matrices |
| $A_{ij},\ A \in \mathbb{R}^{m\times n}$ | $(i,j)$th entry of matrix $A$ |
| $\mathrm{vec}(A)$ | is a vector which contains the columns of $A$ stacked one below each other |
| $A \circ B,\ A, B \in \mathbb{R}^{m\times n}$ | Hadamard product: $\begin{bmatrix} A_{11}B_{11} & A_{12}B_{12} & \cdots & A_{1n}B_{1n} \\ A_{21}B_{21} & A_{22}B_{22} & \cdots & A_{2n}B_{2n} \\ \vdots & \vdots & & \vdots \\ A_{m1}B_{m1} & A_{m2}B_{m2} & \cdots & A_{mn}B_{mn} \end{bmatrix}$ |

## II. Determining affine and sparse interconnections in dynamical systems

Consider a dynamical system of the following form:

$$\frac{dx}{dt} \triangleq \dot{x} = Af(x),\ x \in \mathbb{R}^n,\ A \in \mathbb{R}^{n\times m}, \tag{2}$$

where $f(\cdot) \in \mathbb{R}^m$ is a vector of known functions, which satisfy appropriate smoothness conditions to ensure local existence and uniqueness of solutions. Let neither the value of the entries nor the structure of matrix $A$ be known. What we wish to find is the structure and entries in matrix $A$, given experimental data.

For this purpose, the following discrete-time system was considered in [11]:

$$x(t_{k+1}) = x(t_k) + (t_{k+1} - t_k)Af(x(t_k)), \tag{3}$$

which is the Euler discretization of (2).

Now, the measurements, which we denote by $\hat{x}$, can be used to fit the unknown entries to $A$ such as to minimize the error between the data and the model predictions, which are given by (3). It is popular to solve the minimization problem which results in the least 2-norm on the error (least squares) between $x_i(t_{k+1})$ and $\hat{x}_i(t_{k+1})$. We can write such an error metric as:

$$\min \|Ma - b\|_2 \tag{4}$$

where $a \in \mathbb{R}^{nm}$ is a vector containing $A_{ij}$, which we treat as decision variables, and $M \in \mathbb{R}^{(\{p-1\}\times n)\times nm}$ and $b \in \mathbb{R}^{(\{p-1\}\times n)}$ are defined by 'stacking' all such conditions obtained by manipulating the data as per (3). Here $p$ corresponds to the number of measurements. This problem has the following analytical solution:

$$a^* = M^\dagger b \triangleq (M^{\mathrm{T}}M)^{-1}M^{\mathrm{T}}b. \tag{5}$$

However, the solution puts emphasis on minimizing the error between data and model prediction and not on the structure of $A$. Both converge as the number of measurements increase and the time interval between measurements approaches zero; in other words, as the amount of data increases. Note that if data points are rare (for example, when running experiments is very costly), that is $p \leq m$, and there aren't any constraints on matrix $A$ then the error between the data and the model predictions can be made zero and (5) does not have a unique solution.

Let the entries to $A$ be sparse and measurement data rare (that is, $p \leq m$). Then the following program tries to recover this property of the matrix:

$$\begin{aligned} \min\ & \|\mathrm{vec}(A)\|_1 \\ \text{s. t.}\ & \hat{x}(t_{k+1}) = \hat{x}(t_k) + (t_{k+1} - t_k)Af(\hat{x}(t_k)), \\ & \forall k,\ k = 1, \ldots, p-1. \end{aligned} \tag{6}$$

Thus, if it is known that matrix $A$ is sparse then (6) could provide meaningful results with respect to the structure of $A$. Let us denote the solution of (6) by $A_{LP}$. The following remark shows how sparseness of $A$ might keep the error between $A$ and $A_{LP}$ small in the case when measurement data are rare.

*Remark 1:* Suppose that the initial $f(x_0)$ is in a "sufficiently random" configuration and that the interconnection topology has a constant number of nonzeros per reactant (lets say this constant is $s$). Then with high probability, there is a constant $C_1$ such that $C_1 s \log(n)$ experiments will suffice to determine the structure of $A$ assuming no noise and no error due to the Euler approximation. In the case that we have no noise, but there is additive error due to the Euler approximation

$$\sum_k \left( \frac{x_{t_{k+1}} - x_{t_k}}{t_{k+1} - t_k} - Af(x_{t_k}) \right)^2 < \gamma^2,$$

then solving the SOCP

$$\begin{aligned} \min\ & \|\mathrm{vec}(A)\|_1 \\ \text{s. t.}\ & \textstyle\sum_k \left( \frac{x_{t_{k+1}} - x_{t_k}}{t_{k+1} - t_k} - Af(x_{t_k}) \right)^2 < \gamma^2 \end{aligned}$$

with data from $C_1 s \log(n)$ experiments finds an $A_{SOCP}$ that satisfies

$$||A - A_{SOCP}||_2 \leq C_2 \gamma\,.$$

for a known constant $C_2$. This is a straightforward application of Theorem 1.1 in [20].

In the following, we provide an example to illustrate the results presented in this section.

### A. A linear dynamical system with a sparse but otherwise unknown interaction matrix

Consider the following linear dynamical system

$$\dot{x} = Ax,\ x \in \mathbb{R}^n,\ A \in \mathbb{R}^{n\times n}, \tag{7}$$

where matrix $A$ is sparse but otherwise unknown. We wish to identify the structure of $A$ from measurements as described above. Let the 'true' $A$ be given by

$$A_{\mathrm{true}} = \tag{8}$$

$$\begin{bmatrix} -.2 & 0 & 0 & 0 & -.08 & 0 & -.06 & .08 & 0 & -.07 \\ 0 & -.2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -.2 & .1 & 0 & 0 & 0 & 0 & 0 & .06 \\ 0 & 0 & .09 & -.2 & -.1 & 0 & 0 & 0 & 0 & 0 \\ .1 & 0 & 0 & 0 & -.18 & 0 & 0 & -.06 & .06 & 0 \\ 0 & 0 & 0 & 0 & 0 & -.2 & 0 & 0 & 0 & 0 \\ .02 & 0 & 0 & -.06 & .08 & 0 & -.23 & .05 & -.1 & 0 \\ 0 & 0 & 0 & .06 & 0 & 0 & 0 & -.2 & 0 & 0 \\ 0 & -.02 & .03 & -.07 & 0 & 0 & 0 & -.05 & -.2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & .09 & 0 & -.2 \end{bmatrix}.$$

The network in Figure 2 represents the interaction between variables given by the $A_{\text{true}}$. An arrow from node $i$ to $j$ indicates that $A_{ji}$ is nonzero. (Here, entries to the diagonal, which would result in self-loops, are ignored.) Solid arrows denote a positive entry and dash pointed arrows with a hammer head denote a negative entry.

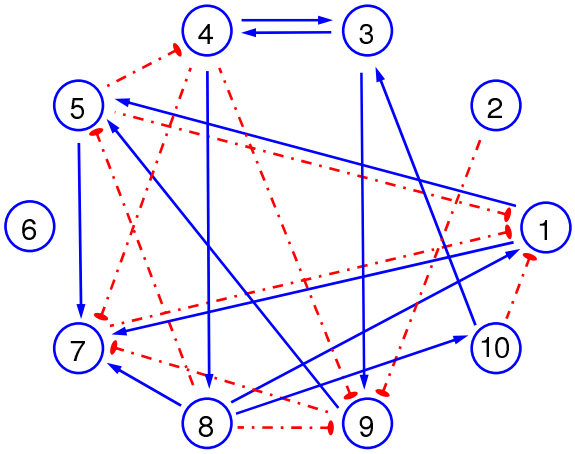


Fig. 2. Network of interactions represented by $A_{\text{true}}$. An arrow from node $i$ to $j$ indicates that $A_{ji}$ is nonzero. (Here, entries to the diagonal, which would result in self-loops, are ignored.) Solid arrows denote a positive entry and dash pointed arrows with a hammer head denote a negative entry.

Now, we produce a mock-up data set with 'measurements' taken every $\Delta t = 5$ between $t = 0$ and $t = 45$ (time is in arbitrary units). With this data, we wish to to solve the linear program given by (6) in order to estimate (8). To do so, we use the solver `SEDUMI` [21] and obtain the matrix $A_{\text{estimated}}$ (not shown). Figure 3 represents the interaction between variables given by matrix $A_{\text{estimated}}$. Most links that existed in the original matrix (8) were identified, only two are missing. Fifteen additional links were wrongly identified. It is important to note however that the all identified connections that overlap with connections given by $A_{\text{true}}$ have the right sign.

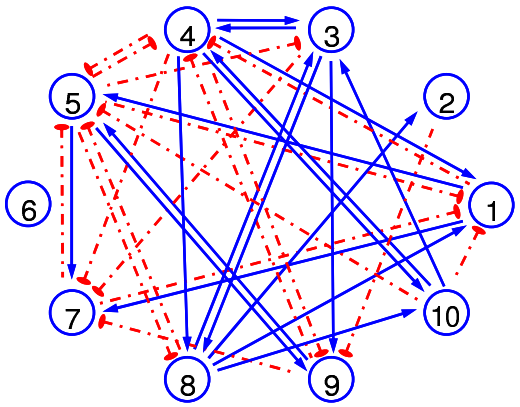


Fig. 3. Network of interactions represented by $A_{\text{estimated}}$: 2 links are missing, 15 additional links were wrongly identified, however, the connections that were identified and overlap with connections given by $A_{\text{true}}$ have the right sign.

Overall, this example shows that the linear program (6) is a powerful tool to identify interconnections between variables of a dynamical system from measurements, if the former are sparse, even when data are rare.

## III. Obtaining the structure of gene regulatory networks

Consider the model of a gene regulatory network as described in [22] and [23], where nodes represent genes. Knowledge about the three hierarchal levels of information mentioned previously are necessary to fully describe the network. The first level determines whether there is an interaction between proteins $X_1$ and $X_2$. An interaction of the form '$X_1 \rightarrow X_2$' means that protein $X_1$ activates the production of protein $X_2$ and '$X_1 \dashv X_2$' that $X_1$ inhibits it. This information belongs to the second level. The activation and repression Hill input functions are given mathematically by (see [22], p. 13):

$$\frac{kx_1^n}{1+kx_1^n}, \text{ and } \frac{1}{1+kx_1^n}, \tag{9}$$

respectively,[1] where $x_1$ is the concentrations of $X_1$. Knowledge about the magnitude of activation or repression coefficient $k$, $k > 0$, and exponent $n$, $n > 0$, is part of the third level of information.

If there exists more than one connection to a particular gene/node then all transcription factors associated with the connections could be necessary to fulfill a specific task (activation or inhibition) or it might be that any of them is sufficient to do the job; more complex combinations are also possible. In [22] (p. 255), a generalised input function of the following form is presented, which takes the possible interplay of different transcription factors into account:

$$f_i(x) = \frac{\sum_j b_{ij} x_j^{n_{ij}}}{1+\sum_j k_{ij} x_j^{m_{ij}}}. \tag{10}$$

Here, activation of protein $X_i$ by protein $X_j$ is represented by $n_{ij} = m_{ij} > 0$, and repression by $n_{ij} = 0$, $m_{ij} > 0$. The contribution of the different proteins is denoted by $b_{ij}$. The mathematical description of the dynamics of the concentrations of protein $X_i$ of an arbitrary large gene regulatory network is as follows:

$$\dot{x}_i = \gamma_i + f_i(x) - d_i x_i, \tag{11}$$

where $\gamma_i > 0$ is the basal production rate and $d_i > 0$ is the degradation rate.

In the following we extend the results of the previous section to a more complicated case, where the dynamical system is nonlinear in the unknowns. Let $\Delta t = t_{\ell+1} - t_\ell$. A discrete-time system that approximates (11) is:

$$x_i(t_{\ell+1}) = x_i(t_\ell) + \Delta t(\gamma_i + f_i(x_i(t_\ell)) - d_i x_i(t_\ell)). \tag{12}$$

Note that if $b_{ij}$, $k_{ij}$ and $m_{ij}$ are unknowns then (12) is not affine in the unknown parameters as is the case in (3). Now, we may rewrite (12) as follows:

[1] In [23], the notation $\frac{1}{K}$ is used instead of $k$. For clarity, we have adopted a 'simpler' notation.

$$(x_i(t_\ell)(1-\Delta t d_i) - x_i(t_{\ell+1}) + \Delta t \gamma_i) \circ (1 + \sum_j k_{ij} x_j^{m_{ij}}) + \Delta t \sum_j b_{ij} x_j^{\tilde{n}_{ij}} + \Delta t b_i = 0, \quad (13)$$

where, $\tilde{n}_{ij}$ corresponds to an exponent $n_{ij}$ such that $n_{ij} > 0$, $b_i = \sum_j b_{ij} x_j^{n_{ij}}$, for which $n_{ij} = 0$. For all $i, j$, let an entry to matrix $B$ be $b_{ij}$ for which $n_{ij} > 0$, and let an entry of matrix $K$ be $k_{ij}$. As before, given a set of measurements, which we denote by $\hat{x}$, this set can be used to approximate the structure of the gene regulatory network determined by $b_{ij}$, $b_i$ and $k_{ij}$ if the hill coefficients $m_{ij}$ (and thus, $n_{ij}$) are known and the basal production and degradation rates are known or considered an uncertainty. For instance, we can try to recover $B, K$ through a LP. The following LP puts emphasis on minimizing the 1-norm of vec$(B)$, $b$, and vec$(K)$, which are independent of each other, while we keep the Euler discretisation error, $\mu$, as small as possible.

$$\begin{array}{ll} \min & \|\text{vec}([B\ K\ b])\|_1 \\ \text{s. t.} & \mu > 0,\ (0 \le \epsilon_{1i} \le \gamma_i \le \epsilon_{2i}, 0 \le \varepsilon_{1i} \le d_i \le \varepsilon_{2i}, \forall i) \\ & -\mu < (\hat{x}_i(t_\ell)(1-\Delta t d_i) - \hat{x}_i(t_{\ell+1}) + \Delta t \gamma_i) \circ (1 \\ & + \sum_j k_{ij} \hat{x}_j^{n_{ij}}) + \Delta t \sum_j b_{ij} \hat{x}_j^{\tilde{n}_{ij}} + \Delta t b_i < \mu,\ \forall i, \ell, \\ & b_{ij} \ge 0,\ k_{ij} \ge 0,\ b_i \ge 0, \quad \forall i, j, \ell. \end{array} \quad (14)$$

(The requirements in brackets correspond to the case of uncertain production and degradation rates.) Now, note that per definition (10) is such that

$$k_{ij} = 0 \text{ if and only if } b_{ij} = 0 \text{ or } b_i = 0,\ \forall i, j. \quad (15)$$

The following remark deals with the case when the solution of (14) violates (15). The rationale behind the idea is that by following these rules we can determine unambiguously whether activation or repression takes place between two proteins.

*Remark 2:* Since requirement (15) cannot be implemented in a LP, we deduce the following from the solution of (14) about the connectivity of the network when (15) is violated:

- if $k_{ij} \neq 0$, $b_{ij} = 0$ and $b_i = 0$ then the production of $X_i$ is not affected by $X_j$; that is, it is the same case as when $k_{ij} = 0$,
- if $b_{ij} \neq 0$ and $k_{ij} = 0$ then $X_j$ enhances the production of $X_i$; i. e., it is the same case as when $k_{ij} \neq 0$,
- if $b_i \neq 0$ and $k_{ij} = 0$ for all $i$ then the production of $X_i$ is not affected by $X_j$; that is, it is the same case as when $b_i = 0$.

In the following, we provide examples applying (14) to mock-up data from simulation experiments.

### A. Sample gene regulatory network

Consider the artificial gene regulatory network given by

$$\begin{aligned} \dot{x}_1 &= \gamma_1 - d_1 x_1, \\ \dot{x}_2 &= \gamma_2 + \frac{b_{12} x_1}{1 + k_{12} x_1} - d_2 x_2, \\ \dot{x}_3 &= \gamma_3 + \frac{b_{43} x_4 + b_{13} x_1 + b_3}{1 + k_{43} x_4 + k_{13} x_1 + k_{53} x_5} - d_3 x_3, \\ \dot{x}_4 &= \gamma_4 + \frac{b_{54} x_5}{1 + k_{54} x_5} - d_4 x_4, \\ \dot{x}_5 &= \gamma_5 + \frac{b_{15} x_1 + b_5}{1 + k_{15} x_1 + k_{25} x_2} - d_5 x_5, \end{aligned} \quad (16)$$

where

$$B = \begin{bmatrix} 0 & 0.51 & 0.87 & 0 & 0.80 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.20 & 0 & 0 \\ 0 & 0 & 0 & 0.22 & 0 \end{bmatrix},$$

$$K = \begin{bmatrix} 0 & 0.31 & 0.87 & 0 & 0.15 \\ 0 & 0 & 0 & 0 & 0.77 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.97 & 0 & 0 \\ 0 & 0 & 0.79 & 0.44 & 0 \end{bmatrix},$$

$b_3 = 0.71$, $b_5 = 0.80$, $\gamma_i = 0.1$ and $d_i = 1$. The network is depicted in Figure 4, where solid lines with an arrow head denote activation and dash pointed lines with a hammer head denote inhibition.

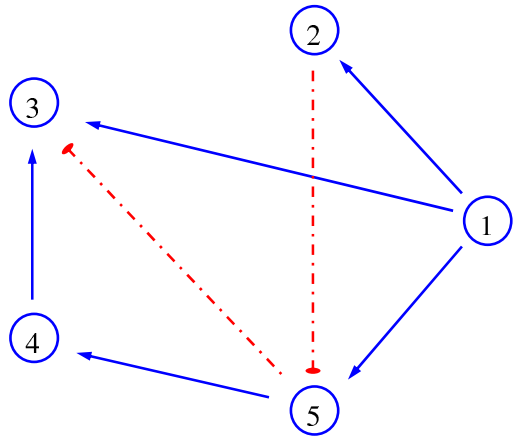


Fig. 4. Artificial gene regulatory network. Solid lines with an arrow head denote activation and dash pointed lines with a hammer head denote inhibition.

We assume that $d_i$ are known but, for all $i$, $\gamma_i = \gamma$ and $0.095 \le \gamma \le 0.105$. We take 'measurements' every $\Delta t = 0.05$ between $t = 0$ and $t = 5$ (time is in arbitrary units) from four different random initial conditions between 0 and 1 in order to obtain mock-up data. Solving (14) using the solver `SEDUMI` [21], we obtain the following results for matrices

$B$ and $K$:

$$B = \begin{bmatrix} 0 & 0.48 & 0.22 & 0 & 1.15 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.11 & 0 \end{bmatrix},$$

$$K = \begin{bmatrix} 0 & 0 & 0 & 0 & 0.61 \\ 0 & 0 & 0 & 0 & 0.75 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.32 & 0 & 0 \\ 0 & 0 & 0.35 & 0 & 0 \end{bmatrix};$$

and $b_3 = 0.64$, $b_5 = 0.80$ (all other $b_i = 0$). Following the rules given by Remark 2, we are able to reconstruct the network shown in Figure 4. As the example show, we were able to determine the interaction network given by (16) through the LP (14) even when degradation rates were considered uncertain.

## IV. CONCLUSIONS AND FUTURE RESEARCH

### A. Conclusions

In this paper, we first presented a methodology for robust determination of the interaction topology of dynamical systems, which are models for biological systems, and that are affine in the unknown parameters using time series data collected from experiments. We extended the results in [11] to large-scale and general networks; moreover, linear program (6) considered rareness of data in addition to sparseness of interconnections. We demonstrated the ability of our method to identify a network structure through examples. We extended our approach to the more complicated case of gene regulatory networks.

### B. Future Research

In Section III, we used a relatively simple mathematical model for a gene regulatory network. More realistic models would include additional complexities, first, by making the Hill coefficient in the activation and repression terms a free variable; and second, because when two transcription factors act on DNA either both are required (AND) or any of them is sufficient (OR) for action. Thus, a valuable research direction is to investigate this case and establish whether similar analysis techniques to the ones presented in this paper can be used. Moreover, a recent approach, the so called 'lasso' considers an objective function to minimize, which consists of the sum of the $L_1$-norm of the vector of unknowns and the least squares of the error (see for example reference [3], [24]). However, the effectiveness of this approach to determine sparse networks is still mainly heuristic and has to be investigated in more depth. Finally, an important future study will be to validate the approaches presented with experimental data.

## REFERENCES


[1] M. K. Stephen Yeung, J. Tegnér, and J. J. Collins. Reverse engineering gene networks using singular value decomposition and robust regression. *PNAS*, 99(9):6163–6168, 2002.

[2] J. Tegnér, M. K. Stephen Yeung, J. Hasty, and J. J. Collins. Reverse engineering gene networks: Integrating genetic perturbations with dynamical modeling. *PNAS*, 100(10):5944–5949, 2003.

[3] A. A. Julius, M. Zavlanos, S. Boyd, and G. J. Pappas. Genetic network identification using convex programming. *Technical Report MS-CIS-07-20*, 2007.

[4] E. J. Crampin, S. Schnell, and P. E. McSharry. Mathematical and computational techniques to deduce complex biochemical reaction mechanisms. *Progress in Biophysics and Molecular Biology*, 86:77–112, 2004.

[5] J. Ross, I. Schreiber, and M. O. Vlad. *Determination of Complex Reaction Mechanisms: Analysis of Chemical, Biological and Genetic Networks*. Oxford University Press, Oxford, UK, 2006.

[6] T. Chevalier, I. Schreiber, and J. Ross. Toward a systematic determination of complex reaction mechanisms. *J. Phys. Chem.*, 97(26):6776–6787, 1993.

[7] J. Gonçalves, R. Howes, and S. Warnick. Dynamical Structure Functions for the Reverse Engineering of LTI Networks. In *Proceedings of the 46th IEEE Conference on Decision and Control*, pages 1516–1522, New Orleans, LA, USA, 2007.

[8] B. N. Kholodenko, A. Kiyatkin, F. J. Bruggeman, E. D. Sontag, H. V. Westerhoff, and J. B. Hoek. Untangling the wires: A strategy to trace functional interactions in signalling and gene networks. *PNAS*, 99(20):12841–12846, 2002.

[9] B. N. Kholodenko and E. D. Sontag. Determination of functional network structure from local parameter dependence data. *oai:arXiv.org:physics/0205003*, 2002.

[10] W. Vance, A. Arkin, and J. Ross. Determination of causal connectivities of species in random networks. *PNAS*, 99(9):5816–5821, 2002.

[11] A. Papachristodoulou and B. Recht. Determining Interconnections in Chemical Reaction Networks. In *Proceedings of the 2007 American Control Conference*, pages 4872–4877, New York City, USA, 2007.

[12] A. Arkin, P. D. Shen, and J. Ross. Determination of causal connectivities of species in random networks. *Science*, 277:1275–1279, 1997.

[13] J. S. Almeida. Predictive non-linear modeling of complex data by artificial neural networks. *Current Option in Biotechnology*, 13:72–76, 2002.

[14] M. Wahde and J. Hertz. Coarse-grained reverse engineering of genetic regulatory networks. *Biosystems*, 55:129–136, 2000.

[15] A. Fujita, J. R. Sato, H. M. Garay-Malpartida, R. Yamaguchi, S. Miyano, M. C. Sogayar, and C. E. Ferreira. Modeling gene expression regulatory networks with the sparse vector autoregressive model. *BMC Systems Biology*, 1:39, 2007.

[16] M. Feinberg. Lectures on chemical reaction networks. Mathematics Research Centre, University of Wisconsin, 1979.

[17] M. Feinberg. Chemical reaction network structure and the stability of complex isothermal reactors–I. The deficiency zero and deficiency one theorems. *Chem. Eng. Sci.*, 42(10):2229–2268, 1987.

[18] V. Filkov, S. Skiena, and J. Zhi. Analysis Techniques for Microarray Time-Series Data. *Journal of Computational Biology*, 9(2):317–330, 2002.

[19] S. Hana, Y. Yoon, and K.-H. Choc. Inferring biomolecular interaction networks based on convex optimization. *Computational Biology and Chemistry*, 31:347–354, 2007.

[20] Emmanuel J. Candès, Justin Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications of Pure and Applied Mathematics*, 59:1207–1223, 2006.

[21] J. F. Sturm. Using `SeDuMi` 1.02, a `MATLAB` toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11-12:625–653, 1999. Available at `http://fewcal.kub.nl/sturm/software/sedumi.html`.

[22] U. Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman & Hall/CRC, Boca Raton, FL, 2006.

[23] S. Mangan and U. Alon. Structure and function of the feed-forward loop network motif. *PNAS*, 100(21):11980–11985, 2003.

[24] E. J. Candès and Y. Plan. Near-ideal model selection by $\ell_1$ minimization. *arXiv:0801.0345v2*, 2007.