Inferring biomolecular interaction networks based on convex optimization

Soohee Han\textsuperscript{a}, Yeoin Yoon\textsuperscript{b}, Kwang-Hyun Cho\textsuperscript{c,∗}

\textsuperscript{a} Bio-MAX Institute, Seoul National University, Seoul 151-818, Republic of Korea
\textsuperscript{b} Graduate Program in Immunology, College of Medicine, Seoul National University, Seoul 110-799, Republic of Korea
\textsuperscript{c} Department of Bio and Brain Engineering and KI for the BioCentury, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Republic of Korea

Received 4 May 2007; received in revised form 20 July 2007; accepted 10 August 2007

Abstract

We present an optimization-based inference scheme to unravel the functional interaction structure of biomolecular components within a cell. The regulatory network of a cell is inferred from the data obtained by perturbation of adjustable parameters or initial concentrations of specific components. It turns out that the identification procedure leads to a convex optimization problem with regularization as we have to achieve the sparsity of a network and also reflect any \emph{a priori} information on the network structure. Since the convex optimization has been well studied for a long time, a variety of efficient algorithms were developed and many numerical solvers are freely available. In order to estimate time derivatives from discrete-time samples, a cubic spline fitting is incorporated into the proposed optimization procedure. Throughout simulation studies on several examples, it is shown that the proposed convex optimization scheme can effectively uncover the functional interaction structure of a biomolecular regulatory network with reasonable accuracy.

© 2007 Elsevier Ltd. All rights reserved.

Keywords: Biomolecular regulatory network; Convex optimization; Inference; Estimation; Sparsity; Spline

1. Introduction

The high throughput measurement technologies in life science enable us to acquire a large amount of quantitative data on biomolecular substances in a living cell. Monitoring the quantitative variation of biomolecular components provides us with information on the intra-cellular stimulus-response processing steps. With the constant development of new technologies, systems theories based on mathematical approaches have been recently adopted to explore biological systems (Khammash and El-Samad, 2004; Sontag, 2004), which has formed a new area called systems biology (Wolkenhauer et al., 2003). One important issue in systems biology is to identify the functional interactions between biomolecular components such as genes and proteins (Wolkenhauer et al., 2004; Barabasi and Oltvai, 2004; Cho et al., 2005).

The identification of physical systems has been widely investigated for a long time and relatively well established (Ljung, 1987; Saligrama, 2005; Markovsky et al., 2005; Barker et al., 2004). When system parameters are not available and thereby it is difficult to apply any physical formula, the identification of a mathematical model from measured data is essential. A cellular dynamic system usually contains a lot of parameters and exhibits too complex dynamics, so it is in general difficult to derive a mathematical model from a physical formula. In this regard, the identification of a biological system is crucial in developing a mathematical model from measured experimental data (Ziv, 2004). In this paper, we present a systematic way of inferring a biological regulatory network which describes the functional interactions between biomolecular components, by using only a limited number of time-series data.

In order to probe intra-cellular interactions, an external perturbation of adjustable parameters or initial concentrations of specific components is often employed and the difference between a normal state and a perturbed state is analyzed. By quantifying \emph{a priori} knowledge on the regulatory relationships into probabilistic models, a substantial amount of work have been done to develop Bayesian approaches (Beal et al., 2005; Schafer and Strimmer, 2005; Werhli et al., 2006; Pournara and
Wernisch, 2004; Chen et al., 2006; Missal et al., 2006). On the other hand, for identification of a biomolecular regulatory network without such probabilistic models, the previous studies have mainly focused on least square criteria in a linearized model (Schmidt et al., 2005; Tegner et al., 2003; Bansal et al., 2006; Thomas et al., 2004; Li et al., 2006). Since the $2$-norm based cost function such as least square criteria puts a very small weight on small residuals, the corresponding optimal solution can have many nonzero elements. This implies that the resulting network can contain many false connections. We also note that biomolecular regulatory networks have a sparse structure. For instance, each node in a gene network interacts with only a small fraction of the other nodes in the network. An approach to obtain such a sparse solution has been developed in a heuristic manner (Yeung et al., 2002). In order to reduce the number of unknown variables to be estimated and thus enhance the accuracy of the identification result, responses to perturbations that directly influence only one component were considered in Sontag et al. (2004). However, in many practical cases, it is difficult to find and apply such a perturbation. In addition, we note that the aforementioned approaches were based on discretization of a continuous-time system and then recovering the continuous-time system via transformation. Since a discretization procedure might introduce additional numerical errors, it is desired to stay with continuous-time systems and estimate time derivatives directly from discrete-time data instead.

In this paper, an optimization-based inference scheme is proposed to unravel the functional interactions among biomolecular components within a cell. It turns out that the inference procedure leads to an efficient convex optimization problem with regularization. Since convex optimization has been well studied for a long time, a variety of efficient algorithms were developed and many numerical solvers are freely available (Grant et al., 2005; Sturm, 2004).

The sparsity of a solution is further considered in this paper by formulating a cost function with the sum-absolute-value norm such as least square criteria. If we have any prior information, we can impose it as a further constraint. Previously known interactions between components can be easily incorporated in this way. Moreover, this paper considers only a continuous-time system in order to minimize the numerical errors caused by discretization. In particular, we employ a cubic spline method to estimate time derivatives from measured discrete-time data.

In Section 2, a mathematical formulation of inferring a biomolecular regulatory network is described and the corresponding convex optimization problem is constructed. An inference algorithm based on convex optimization is then proposed. In Section 3, the identification results of the proposed inference scheme are illustrated by three examples. Finally, conclusions are made in Section 4.

2. Inference of Biomolecular Interaction Networks and Convex Optimization with Regularization

We consider a state vector $x(t) = [x_1(t) \ldots x_n(t)]^T$, the components of which represent concentrations, activities, or expressions of biomolecular components in a cellular network. The state $x(t)$ evolves along with time and constitutes the following nonlinear dynamic system:

$$\dot{x}(t) = f(x(t), p),$$

(1)

where $p$ is a vector of adjustable parameters such as kinetic rate constants, pH, and temperature. A system in the form of (1) can be considered as a network represented by a weighted directed graph. The nodes and edges of the network correspond to the biomolecular components and regulatory relationships between the components, respectively.

From (1), the state component $x_i(t)$ for each network node can be written as

$$\dot{x}_i(t) = f_i(x(t), p).$$

(2)

Note that the function $f_i(\cdot, \cdot)$ describes how the rate of change of $x_i$ depends on other components of the network. If all the interactions between biomolecular components within a cell are properly identified, we can reconstruct the function $f$ in terms of the so-called biomolecular kinetic equations.

If systems are assumed to be operating near a steady state, then the Jacobian matrix $A$ can be given by

$$A_{ij} = \frac{\partial f_i}{\partial x_j}.$$  

(3)

If $A_{ij}$ is zero, the component $x_j$ has no direct effect on the component $x_i$. In this case, there is no edge from the node $j$ to the node $i$ in the network. On the other hand, if $A_{ij} > 0$, the node $j$ activates the node $i$ by enhancing the net rate of $x_i$ production, and if $A_{ij} < 0$, the node $j$ inhibits the node $i$. The nonzero values of $A_{ij}$ specify the positive (activating) or negative (inhibiting) interaction strengths between network nodes. The higher the absolute value of $A_{ij}$ has, the stronger the effect of the node $j$ on the node $i$ is. In biomolecular networks, identifying the sign of nonzero elements in $A$ is even useful since only a very small number of sampled data are available from experiments while the underlying dynamics are highly nonlinear. Such qualitative information on the interactions (activation or inhibition) can be utilized in bio-medical applications by predicting an adverse effect of a new drug at a genomic level for instance.

The nonlinear dynamic system (1) can be approximated by a linearized system based on the Jacobian $A$ in (3) as follows:

$$\delta\dot{x}(t) = \frac{\partial f}{\partial x} \delta x(t) + \frac{\partial f}{\partial p} \delta p,$$

(4)

$$\delta\dot{x}(t) = A \delta x(t) + b \delta p,$$

(5)

where $\delta x(t)$ and $\delta p$ represent the differentials of a state and a parameter, respectively.

If the effect of perturbations on a network is partially known in advance, we can use this a priori knowledge in reducing the number of unknown variables to be estimated and thereby can enhance the accuracy of the identification. Suppose that a set of experimental perturbations that do not directly influence $x_i$ is selected. Each of these perturbations may directly affect one or more nodes other than $x_i$. For a formal description, for each $x_i$, we choose a set of parameters $p_j$ such that the function $f_i$ does not depend on $p_j$...
not depend on $p_j$:

$$\frac{\partial f_i(x, p)}{\partial p_j} = 0. \quad (6)$$

The parameters $p$ satisfying (6) can simplify the structure of $b$ in (5) and thus the number of variables to be estimated can be reduced. While the approach of Sontag et al. (2004) requires such parameters, these are not necessarily assumed in this paper.

In experiments, we assume that $L$ kinds of stimuli are applied and the concentrations of $N$ different biomolecular components $(x_1, x_2, \ldots, x_N)$ are measured at different $M$ time points for each stimulus. Let us denote $b$ obtained from $l$ th stimulus by $b_l$. For simplicity, we assume that each $b_l$ includes $\delta p$ in (5). We can arrange the measurements as follows:

$$X = [X_1 \ X_2 \ \cdots \ X_L],$$

$$X_l = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,M} \\
 x_{2,1} & x_{2,2} & \cdots & x_{2,M} \\
 \vdots & \vdots & \ddots & \vdots \\
 x_{N,1} & x_{N,2} & \cdots & x_{N,M} \end{bmatrix}, \quad (8)$$

where the subscripts $i, j$ denote the individual component and the experiment number, respectively. In terms of the augmented matrix (8), the system (1) can be approximated as

$$\dot{X} = AX + B,$$  

where $B = [B_1 \ \ldots \ B_L]$ with $B_l = [b_{1l} \ \ldots \ b_{Nl}]$. In this paper, $B$ is assumed unknown and should also be estimated from the measurements $X$. As mentioned before, the structure of $B$ can be further simplified by using, if available, the parameters satisfying (6).

The goal of reverse engineering biomolecular regulatory networks is to infer the matrices $A$ and $B$ from measurements $X$. Note however that all connectivity information can be found from $A$. We can have better estimates of $A$ if $A$ can be structurized in advance based on any available information. Previous studies suggest that most of the biomolecular regulatory networks have a sparse structure in which each network node interacts with only a small number of other nodes. Hence, we should look for a sparse matrix $A$ with as many zero components as possible while it fits the model (9). In many cases, we have a priori knowledge on the partial interaction structure and thus we can impose this as a constraint. In particular, if some positive or negative interactions among the components are already known, we can obtain a more accurate identification result by imposing these as constraints. Let $S_P$ and $S_N$ denote a set of a priori known index pairs $(i, j)$ representing that $x_j$ activates and represses $x_i$, respectively. It is noted that indices of the form $(i, i)$ belongs to $S_N$ since most of the biomolecular network nodes have their own self-regulatory term for homeostasis. In other words, $S_P$ and $S_N$ are constructed from previously known connections that might be uncovered from real experiments. Other connections not contained in $S_P$ or $S_N$ are to be mathematically estimated from measured data. As a priori knowledge, we may know the reliability of each experiment, leading to put more weight on some experiments and less weight on others. Such weighting can be effective in handling experimental noises.

Let us construct a convex optimization problem for the aforementioned identification issue as follows:

$$\min_{A,B} \| (\dot{X} - B - AX)W \|_F$$

subject to \( \text{card}(A) \leq k, A_{i,j} > 0, A_{r,s} < 0, \) \( \text{(11)} \)

for a given positive constant $k$, all $(i, j) \in S_P$ and all $(r, s) \in S_N$, where $W$ is a weighting matrix of appropriate dimension, $\text{card}(A)$ is the maximum number of nonzero elements in rows of $A$, and $\| \cdot \|_F$ is defined as

$$\|S\|_F = (\text{tr}(S^TS))^{1/2} = \left( \sum_{i=1}^{N} \sum_{j=1}^{M} S_{ij}^2 \right)^{1/2}. \quad \text{(12)}$$

If measurement noises are modeled with white Gaussian noise $V$ as $X = AX + B + V$, the optimal weighting matrix $W$ will be an inverse of $E[VV^T]$. Although it is difficult to know the exact information on $E[VV^T]$, we can make use of this weighting factor to reflect the confidence level of each experiment. The first inequality of (11) implies that the number of components influencing one node is limited by $k$. The problem of obtaining the derivative $\dot{X}$ in (10) from $X$ will be discussed later in this section. A solution of the optimization problem (10) should be chosen to fit the model (9) well enough while satisfying constraints (11). Since $M$ is usually much smaller than $N$ due to the high cost of perturbation experiments, the problem (10) becomes under-determined if the constraint (11) is not considered. In this case, the optimal solution is not unique. One of the solutions can be represented by

$$A = (\dot{X} - B)X^+,$$  

where $X^+$ is a pseudo-inverse of $X$ and $W$ is considered as a unit matrix.

One way of solving the optimization problem (10) is to check every possible sparsity pattern of $A$ with $k$ nonzero entries in each row. For a fixed sparsity pattern, we can find the optimal $\hat{A}$ and $\hat{B}$ by solving the problem, where $\hat{A}$ denotes the $k \times n$ matrix obtained by removing zero elements of $A$. This should be repeated for $(n!/(k!(n-k)!))^s$ sparsity patterns with $k$ nonzero elements in each row. In order to avoid such heavy combinatoric computations, we construct the following optimization problem with the sum-absolute-value regularization:

$$\min_{A,B} \| (\dot{X} - B - AX)W \|_F + \gamma \|A\|_{\text{sav}}, \quad \text{(14)}$$

where $\gamma$ is a positive constant and $\| \cdot \|_{\text{sav}}$ is defined as

$$\|M\|_{\text{sav}} = \sum_{i=1}^{N} \sum_{j=1}^{M} |M_{ij}|. \quad \text{(15)}$$

Note that if $A$ and $X$ in (14) are vectors, then the problem (14) reduces to the conventional $l_1$ norm regularization (Boyd and Vandenberghe, 2004). $\gamma \|A\|_{\text{sav}}$ in the optimization problem (14) plays the role of realizing a sparse solution. By varying the
parameter γ, we can sweep out the optimal trade-off curve between the model fitting and the sparsity of a solution. As γ increases, the sparsity of solution is given with a more weight.

Another good heuristic approach to the problem (10) is to solve the optimization problem (14) for different values of γ and then find the smallest value of γ. Let us discuss how to estimate the derivatives of X. Given a data set X, we apply the cubic spline interpolation to achieve a smooth interpolation function that is twice differentiable (Press et al., 2002). According to the cubic spline fitting, we have

$$x_{j,i} = \frac{x_{j,i+1} - x_{j,i-1}}{h} - \frac{h}{3} x_{j,i}.$$  

(16)

where $1 \leq j \leq N$, $1 \leq i \leq M - 1$, $1 \leq l \leq L$, $h$ is a sampling time, and $\bar{x}_j$ is computed from

$$\bar{x}_j = \frac{h}{6} x_{j,i+1} + \frac{2h}{3} x_{j,i} + \frac{h}{6} x_{j,i-1} = \frac{x_{j,i+1} - x_{j,i-1}}{h} - \frac{h}{3} x_{j,i}.$$  

(17)

Note that the constraint (17) makes the interpolation function twice differentiable and thereby achieves more smoothness. From (16), $\dot{X}$ can be computed by the following equation:

$$\dot{X} = \frac{1}{h} X P - \frac{h}{3} \ddot{X},$$  

(18)

where $P$ is given by diag($p$, $\ldots$, $p$) with $p$ defined as

$$p = \begin{bmatrix} -1 & 0 & \cdots & 0 & 0 \\ 1 & -1 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & -1 \\ 0 & 0 & \cdots & 1 & 1 \end{bmatrix}.$$  

(19)

and the notation ‘diag’ for diagonal augmentation of matrices. Note that the two last columns of $P$ in (19) are same since $x_{M+1}$ is not available. Augmenting the relation (17) from $i = 1$ to $i = N$ yields

$$\ddot{X} Q = \frac{6}{h^2} X R,$$  

(20)

where $Q$ and $R$ are given by $\text{diag}(q, \ldots, q)$ and $\text{diag}(r, \ldots, r)$ with $q \in \mathbb{R}^{M \times (M-2)}$ and $r \in \mathbb{R}^{M \times (M-2)}$ defined as

$$q = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 4 & 1 & \cdots & 0 \\ 1 & 4 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 4 \\ 0 & 0 & \cdots & 1 \end{bmatrix},$$

$$r = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ -2 & 1 & \cdots & 0 \\ 1 & -2 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -2 \\ 0 & 0 & \cdots & 1 \end{bmatrix},$$  

(21)

respectively. The relation (20) can be weakened by incorporating it into the cost function such that the optimisation problem is reformulated as

$$\min_{X,A,B} \left\| \frac{1}{h} X P - \frac{h}{3} \ddot{X} + AX \right\|_F + \left\| \dot{X} Q - \frac{6}{h^2} X R \right\|_F + \gamma \|A\|_{\text{sav}}.$$  

(22)

subject to $A_{ij} > 0$, $A_{i,s} < 0$, 

(23)

for all $(i, j) \in S_P$ and all $(r, s) \in S_N$. Note that the twice differentiability of an interpolation function is no longer guaranteed in the optimization problem (22), (23) although its smoothness is still achieved due to the term $\dot{X} Q - (6/h^2) X R$ in the cost function (22). The optimization problem (22) with the constraint (23) turns out as a tractable convex optimization problem, and thereby the globally optimal solution can be found with a high efficiency (Boyd and Vandenbergh, 2004).

Since the norm $\|A\|_{\text{sav}}$ considers all elements of the matrix $A$, the number of links for some nodes can be unnecessarily large. In order to address the sparsity of each component, an alternative way of formulation is introduced as follows:

$$\min_{X,A,B} \left\| \frac{1}{h} X P - \frac{h}{3} \ddot{X} + AX \right\|_F + \left\| \dot{X} Q - \frac{6}{h^2} X R \right\|_F + \gamma \|a_i^T\|_1,$$  

(24)

for $i = 1, \ldots, N$, where $\| \cdot \|_1$ is a 1-norm of a vector and $a_i^T$ is the $i$th row vector of $A$. The second term in (24) can be weighted against the model fitting represented by the first term, which enables us to alleviate the effect of noise. In this paper, however, we do not use such a weighting factor for simplicity.

After all, we can summarize the foregoing developments in the following algorithm:
Convex optimization based inference algorithm

Step 1. Determine $k$ in (11) and constraints on the structure $A$ and $B$ from any available a priori knowledge.

Step 2. Find out the sparsity pattern of the matrix $A$.

From $i = 1$ to $i = N$

By increasing $\gamma$, solve the optimization problem (24) and choose a proper number $\gamma^*$ so that $\text{card}(A) \leq k$ is satisfied.

Step 3. Solve the optimization problem (24) again with $\gamma = 0$ under the constraints that satisfy the sparsity pattern of the matrix $A$ obtained from Step 2.

In order to evaluate the inference result, we can consider the following cost function $C$:

$$
C = \frac{\|S\|_2}{\|N\|_2},
$$

where $S$ and $N$ denote the interaction strengths that are inferred as true positives and false positives, respectively. This cost function represents that we want to have an inference result with a high true positive ratio and a low false positive ratio. The same concept has been widely utilized in communication engineering as a signal to noise ratio (Haykin, 2000).

3. Examples

In this section, we consider three examples to illustrate the proposed inference scheme.

3.1. Artificial Gene Network with Simulated Data

We first consider an artificial gene network composed of five genes adopted from Schmidt et al. (2005) that was used as a test bed for inference of interaction matrices by estimating Jacobian. The activity level of ith gene is given by the concentration $x_i$ representing its mRNA abundance. The mathematical model and its parameters are given as follows:

$$
\begin{align*}
\dot{x}_1 &= V_{1d} \frac{1 + A_{1d}(x_4/K_{14a})^{n14}}{(1 + (x_4/K_{14a})^{n14})(1 + (x_2/K_{121})^{n12})} - V_{1d} \frac{x_1}{K_{1d} + x_1}, \\
\dot{x}_2 &= V_{2d} \frac{1 + A_{2d}(x_4/K_{24a})^{n24}}{(1 + (x_2/K_{241})^{n24})} - V_{2d} \frac{x_2}{K_{2d} + x_2}, \\
\dot{x}_3 &= V_{3d} \frac{1 + A_{3d}(x_2/K_{32a})^{n32}}{(1 + (x_2/K_{321})^{n32})(1 + (x_3/K_{351})^{n35})} - V_{3d} \frac{x_3}{K_{3d} + x_3}, \\
\dot{x}_4 &= V_{4d} \frac{1 + A_{4d}(x_3/K_{43a})^{n43}}{(1 + (x_3/K_{431})^{n43})} - V_{5d} \frac{x_5}{K_{5d} + x_5}.
\end{align*}
$$

Fig. 1. The interaction structure of the artificial gene network (Schmidt et al., 2005) where the arrows indicate activating (or positive) interactions and the blunted lines denote inhibiting (or negative) interactions. The corresponding mathematical model is described in (26). In this case, the maximum number of components influencing one node is 2. For an identification purpose, we assumed that this indegree number is smaller than or equal to 4. The exact Jacobian matrix computed from the mathematical model and the estimated Jacobian matrix obtained from the proposed method are shown in (27) and (30), respectively.

$$
\begin{align*}
-A_{1d} &\frac{x_4}{K_{1d} + x_4}, \dot{x}_5 = V_{5d} \frac{1 + A_{5d}(x_1/K_{51a})^{n51}}{1 + (x_1/K_{51a})^{n51}} - V_{5d} \frac{x_5}{K_{5d} + x_5}.
\end{align*}
$$

Table 1 shows how the inferred values change with $\gamma$. We found that the sparsity pattern gets clearer as $\gamma$ increases. In this

$$
A_{true} = \begin{bmatrix}
-6.46 & -2.94 & 0 & 2.43 & 0 \\
0 & -8.18 & 0 & 3.59 & 0 \\
0 & 1.68 & -14.68 & 0 & -0.85 \\
562.4 & -293.3 & 0 & 0 & -355.7
\end{bmatrix}.
$$

We assumed to take perturbations of 50% for $V_{1d}$, $V_{2d}$, $V_{3d}$, $V_{4d}$, and $V_{5d}$, as the inhibition or activation efficiency achievable in experiments is high enough nowadays. We performed five perturbation simulations and combined the obtained measurements. In each simulation, the maximal transcription rates of the five genes were perturbed by 50% and four samples were taken. The sampling time and $k$ are set to 0.01 h and 4, respectively. $\gamma^*$'s were chosen as 0.05, 0.01, 0.01, 10, and 0.4 for $i = 1$–5, respectively in order to achieve the sparsity pattern. White Gaussian noise with a variance of 0.001 was applied.
from true positive ones. Improved and therefore it becomes easier to distinguish false positive interactions with the weighting values. As a less weight is put on the data.

The relationship between a weighting matrix and noises

Table 2

The numbers in parentheses denote the corresponding true values. Note that the cost \( C \) increases with the value \( \gamma \). This implies that we can make a clearer decision on whether there exists an interaction or not.

In order to validate the effect of the weighting matrix \( W \), we applied noises with a large variance (0.1), to the second and third perturbation simulations. Table 2 shows how the cost \( C \) changes with the weighting values. As a less weight is put on the data corrupted with noises, the cost \( C \) becomes improved. This means that it is easier to distinguish false positive interactions from true positive ones.

The final identification result obtained by applying the proposed inference is as follows:

\[
A_{\text{estimated}} = \begin{bmatrix} -3.58 & -2.83 & 0 & 2.31 & 0 \\ 0 & -11.81 & 0 & 3.98 & 0 \\ 0 & 2.80 & -13.53 & 0 & -3.23 \\ 0 & 0 & 8.08 & -5.91 & 0 \\ 10.59 & -0.40 & 0 & 0.28 & -0.01 \end{bmatrix}
\]

(30)

It turns out that most signs of elements in \( A_{\text{estimated}} \) are identical with those in \( A_{\text{true}} \). The fifth row of \( A_{\text{estimated}} \) shows a rather big difference from that of \( A_{\text{true}} \). We note that the very large value of \( V_{5d} \) in (26) compared with other coefficients induces the fast dynamics of the corresponding differential Eq. (26) and this makes us difficult to obtain a good estimate with the given sampling time points. If we have more frequent sampling time points, we can have a better estimation result, but we should also expense as much cost and efforts to obtain the additional data.

3.2. Chemotactic Signaling Network of Dictyostelium discoideum with Simulated Data

Let us apply the proposed inference scheme to the chemotactic signaling network of Dictyostelium. The mathematical model of this system is given in Maeda et al. (2004) based on experimental measurements. The interaction structure is depicted in Fig. 2. The state variable and the positive/negative interactions of the network are given as follows:

\[
x = \begin{bmatrix} ACA \\ PKA \\ ERK2 \\ RegA \\ cAMPi \\ cAMPe \\ CAR1 \end{bmatrix},
A_{\text{true}} = \begin{bmatrix} - & - & 0 & 0 & 0 & 0 & + \\ 0 & - & 0 & 0 & + & 0 & 0 \\ 0 & - & - & 0 & 0 & 0 & + \\ 0 & 0 & - & - & 0 & 0 & 0 \\ + & 0 & 0 & - & - & 0 & 0 \\ + & 0 & 0 & 0 & 0 & - & 0 \\ 0 & 0 & 0 & 0 & + & - \end{bmatrix}
\]

(31)

We assumed to take perturbations of 1.0, 1.0, 0.2, 0.7, 0.2, 2.5, −2.0 from the original reaction coefficients 2.0, 1.5, 0.6, 1.3, 0.3, 4.9, and 4.5. We performed seven perturbation simulations. The sampling time and \( k \) were set to 0.2 h and 4, respectively, and eight samples are taken. \( \gamma^* \)’s were all chosen as 0.3. Noises were not applied in order to compare with other schemes. The goal is to estimate the Jacobian matrix such that we can unravel its sign pattern. The estimation result
Table 3
Comparison of the identification results obtained by the proposed schemes and previous methods

<table>
<thead>
<tr>
<th>Schemes</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>M-I (Bayesian network)</td>
<td>4/18 (22%)</td>
<td>18/24 (75%)</td>
</tr>
<tr>
<td>M-II (dynamic Bayesian network)</td>
<td>2/9 (22%)</td>
<td>25/33 (76%)</td>
</tr>
<tr>
<td>M-III (phase portrait scheme)</td>
<td>3/7 (43%)</td>
<td>28/35 (80%)</td>
</tr>
<tr>
<td>P-I (proposed scheme without threshold)</td>
<td>9/18 (50%)</td>
<td>23/24 (95%)</td>
</tr>
<tr>
<td>P-II (proposed scheme with threshold)</td>
<td>9/12 (75%)</td>
<td>29/30 (97%)</td>
</tr>
</tbody>
</table>

PPV and NPV stand for the positive predictive value and the negative predictive value, respectively. M-I, M-II, and M-III represent the schemes of Friedman et al. (2000), Ong et al. (2002), and Cho et al. (2006), respectively. The results obtained by the proposed schemes are denoted by P-I and P-II. In the case of P-II, any identified interaction with an estimated strength below 0.1 is considered to be zero (i.e., no interaction).

Table 3 shows the summary of identification results comparing the proposed scheme with previous ones where PPV and NPV stand for the positive predictive value and the negative predictive value, respectively. M-I, M-II, and M-III correspond to the schemes of Friedman et al. (2000), Ong et al. (2002), and Cho et al. (2006), respectively. The results obtained by the proposed schemes are shown under P-I and P-II. In the case of P-II, any interaction with a strength below 0.1 is considered to be zero. It turns out that the proposed scheme provides better identification results than any of the previous schemes.

3.3. Gene Regulatory Network of Yeast with Real Data

In Ronen and Botstein (2006), short-term perturbation experiments were undertaken for yeast in order to understand the dynamics of transcriptional responses to the variation of surrounding glucose concentrations. It was reported that more than 20 genes are involved in such responses. Among those, we choose four representative genes and illustrate how the proposed scheme can identify their regulatory network. The true gene regulatory network of the selected genes is shown in Fig. 3. The state variable and the signs of the corresponding Jacobian matrix are given by

\[
A_{\text{true}} = \begin{bmatrix}
- & - & 0 & 0 & + & 0 & + \\
0 & - & 0 & 0 & + & 0 & 0 \\
0 & - & - & +* & 0 & 0 & + \\
0 & +* & - & - & 0 & 0 & -* \\
+ & 0 & 0 & - & - & 0 & 0 \\
+ & -* & 0 & 0 & -* & - & 0 \\
+ & - & 0 & 0 & - & 0 & -
\end{bmatrix} \tag{33}
\]

In Ronen and Botstein (2006), two different sizes of pulses (0.2 and 2.0 g/l of glucose) were applied and the measurements were taken with uneven sampling times. Through interpolation with a cubic spline method and resampling, we obtained an evenly sampled data set. The sampling time and \( k \) were set to 0.033 h and 3, respectively, and all five samples were taken. \( \gamma^* \)'s were all chosen as 0.01.

By applying the proposed inference scheme, the signs of the Jacobian matrix are estimated as follows:

\[
A_{\text{estimated}} = \begin{bmatrix}
- & - & 0 & 0 & + & 0 & + \\
+ & - & + & 0 & 0 & 0 & 0 \\
+ & - & + & 0 & 0 & 0 & 0 \\
+ & - & 0 & 0 & - & - & 0 \\
0 & + & 0 & 0 & - & - & 0 \\
0 & 0 & + & 0 & - & - & 0 \\
0 & 0 & + & 0 & - & - & 0
\end{bmatrix} \tag{34}
\]

It can be seen that the PPV and the NPV are 3/8(37.5%) and 3/4(75%), respectively.

4. Conclusion

In this paper, we have proposed a convex optimization-based inference scheme to unravel the hidden functional interactions between biomolecular components within a cell. The intracellular regulatory networks are identified from time-series measurements obtained by perturbation of adjustable parameters or initial concentrations of specific components. The identification problem has led to a convex optimization problem with the
sum-absolute-norm regularization by which we can achieve a sparsity of the biomolecular network and can also reflect a priori information on the network structure. It was shown through three examples that the proposed scheme can successfully identify the biomolecular regulatory networks with reasonable accuracy.

Owing to the recent development of DNA microarray technologies, more refined and accurate time-series data are available and, in this respect, the proposed inference scheme can be useful for identifying the functional interaction structure of large-scale gene regulatory networks if the corresponding perturbation experiments are properly designed. Since the convex optimization adopted in this paper has a sound theoretic ground and a variety of solvers are freely available, the proposed inference scheme can be widely used for reverse engineering of various biomolecular interaction networks.

Acknowledgements

This work was supported from the Korea Ministry of Science and Technology through the Korean Systems Biology Research Grant (M10503010001-07N030100112), the Nuclear Research Grant (M20708000001-07B0800-00110), and the 21C Frontier Microbial Genomics and Application Center Program (Grant MG05-0204-3-0), in part from the Korea Ministry of Commerce, Industry & Energy through the Korea Bio-Hub Program (2005-B0000002), and from the NITR/KOREA FDA through the National Toxicology Program in Korea (KNTP).

References

Grant, M., Boyd, S., Ye, Y., 2005. CVX Software (http://www.stanford.edu/boyd/cvx/).