Available online at www.sciencedirect.com

ScienceDirect

Computational Biology and Chemistry

Computational Biology and Chemistry 31 (2007) 347–354

www.elsevier.com/locate/compbiolchem

Inferring biomolecular interaction networks based on convex optimization

Soohee Han[a], Yeoin Yoon[b], Kwang-Hyun Cho[c,*]

[a] *Bio-MAX Institute, Seoul National University, Seoul 151-818, Republic of Korea*
[b] *Graduate Program in Immunology, College of Medicine, Seoul National University, Seoul 110-799, Republic of Korea*
[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 *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.

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 *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

[*] Corresponding author at: Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology, 335 Gwahangno, Yuseong-gu, Daejeon 305-701, Republic of Korea. Tel.: +82 42 869 4325;
fax: +82 42 869 4310.
E-mail address: ckh@kaist.edu (K.-H. Cho).
URL: http://sbie.kaist.ac.kr (K.-H. Cho).

1476-9271/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2007.08.003

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 descritization 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 regularization. If we have any *a priori* 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 $x(t) = [x_1(t) \ldots x_n(t)]^{\mathrm{T}}$, 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:

$$\dot{x}(t) = f(x(t), p), \tag{1}$$

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

$$\dot{x}_i(t) = f_i(x(t), p). \tag{2}$$

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

$$A_{ij} = \frac{\partial f_i}{\partial x_j}. \tag{3}$$

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:

$$\delta\dot{x}(t) = \frac{\partial f}{\partial x}\delta x(t) + \frac{\partial f}{\partial p}\delta p, \tag{4}$$

$$\delta\dot{x}(t) = A\delta x(t) + b\delta p, \tag{5}$$

where $\delta x(t)$ and δp 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