Available online at www.sciencedirect.com
ScienceDirect
Computational Biology and Chemistry 31 (2007) 239–245
Computational Biology and Chemistry
www.elsevier.com/locate/compbiolchem

Inferring gene regulatory networks from temporal expression profiles under time-delay and noise

Shinuk Kim [a], Junil Kim [b], Kwang-Hyun Cho [a,c,*]

[a] *Bio-MAX Institute, Seoul National University, Gwanak-gu, Seoul 151-818, Republic of Korea*
[b] *Interdisciplinary Program in Bioinformatics, Seoul National University, Gwanak-gu, Seoul 151-747, Republic of Korea*
[c] *College of Medicine, Seoul National University, Jongno-gu, Seoul 110-779, Republic of Korea*

Received 15 January 2007; accepted 30 March 2007

Abstract

Ordinary differential equations (ODE) have been widely used for modeling and analysis of dynamic gene networks in systems biology. In this paper, we propose an optimization method that can infer a gene regulatory network from time-series gene expression data. Specifically, the following four cases are considered: (1) reconstruction of a gene network from synthetic gene expression data with noise, (2) reconstruction of a gene network from synthetic gene expression data with time-delay, (3) reconstruction of a gene network from synthetic gene expression data with noise and time-delay, and (4) reconstruction of a gene network from experimental time-series data in budding yeast cell cycle.

Keywords: Gene networks; Inference; Optimization; Ordinary differential equations; Noise; Time-delay

1. Introduction

While the recent developments of the high throughput measurement technologies such as DNA microarrays have made it possible to obtain a large volume of gene expression profiles, it is not answered yet how to assemble the data and construct a predictive model on the gene network. In this regard, inferring the interaction structure of a gene network is an imperative task in order to advance our understanding on the molecular mechanism of cellular functioning at a genome level.

Various approaches have been proposed to tackle such a gene network identification problem (Bansal et al., 2006; Brazhnik, 2005; Kholodenko et al., 2002; Kimura et al., 2005; Wahde and Hertz, 2000; Yeung et al., 2002). Although those approaches have been applied with some success, they are mostly demanding on the data along with a certain degree of *a priori* information. For instance, a popular approach using steady state gene expression data has been proved to be highly effective in simplifying the mathematical formulations to infer small microbial gene networks and thereby minimizing unknown parameters (Andrec et al., 2005; Gardner et al., 2003; Kholodenko et al., 2002; Sontag et al., 2004). However, the approach requires *a priori* knowledge on the genes involved in the network as well as the measurement of gene expressions at a steady state after transcriptional perturbations. On the other hand, an approach using time-series gene expression data can provide a clearer picture on the functional interaction between genes and the corresponding network. However, this approach also requires a perturbation of each gene of interest and subsequent measurements of the gene expression profiles at multiple time points.

Even though the measurement technologies have been revolutionized, noise and time-delay still have to be considered to infer gene networks under various experimental situations. However, such noise and time-delay have not been properly considered in previous studies (Dasika et al., 2004; Liu and Cao, 2006; Smolen et al., 2000; Xu et al., 2005). Hence, we propose in this paper a numerical scheme by which we can infer a gene network in consideration of such time-delay and noise.

To develop a mathematical formulation on inference of gene networks, we employ ordinary differential equations (ODEs) (Gardner et al., 2003; de Jong, 2002; Tegner et al., 2003). Based on this framework, we formulate the inference problem as a mathematical optimization problem and apply the

* Corresponding author. Present address: Bio-MAX Institute, Seoul National University, 3rd Floor, IVI, San 4-8, Bongcheon 7-dong, Gwanak-gu, Seoul 151-818, Republic of Korea. Tel.: +82 2 887 2650; fax: +82 2 887 2692.
E-mail address: ckh-sb@snu.ac.kr (K.-H. Cho).
URL: http://systemsbiology.snu.ac.kr (K.-H. Cho).

1476-9271/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2007.03.013

Broyden–Fletcher–Goldfarb–Shanno (BFGS) method (Press et al., 1992).

In particular, we consider the following two datasets: synthetic data produced from an artificial example (Sontag et al., 2004) and cell cycle experimental data (Bahler, 2005; Spell man et al., 1998). Especially, the proposed numerical scheme was applied to example networks of four nodes from G2/M and M/G1 transition phases and eight nodes representing the clusters of a budding yeast cell cycle. The remainder of this paper is organized as follows. Section 2 presents a mathematical model and proposes a numerical analysis scheme. Section 3 shows two examples illustrating the proposed scheme. Section 4 provides a summary of the method and discussions.

2. Model and method

Mathematical formulation and a schematic outline of the proposed method for inferring a dynamic gene network are described in this section.

2.1. Mathematical model

To analyze a gene regulatory system, the following nonlinear dynamic equation is employed:

$$\frac{dx}{dt} = f(x) \tag{1}$$

where $x = [x_1, x_2, \ldots, x_n]$ represents a set of gene expression levels.

If we take the first-order Taylor approximation, then

$$f(x) = f(x_0) + \frac{df(x)}{dx}(x - x_0) \tag{2}$$

where x_0 denotes the initial value of x. Based on this, we can have the following linear dynamic system (Chen et al., 1999):

$$f(x) = Ax \tag{3}$$

where A is the Jacobian matrix describing $df(x)/dx$. The component of A (i.e., $a_{ij} = df_i(x)/dx_j$) represents the gene interaction between x_i and x_j that is to be inferred. By substituting (3) into (1), we have

$$\frac{dx_i}{dt} = \sum_{j=1}^{n} a_{ij}x_j \quad i = 1, \ldots, n \tag{4}$$

where a_{ij} represents the influence of gene j on gene i. If we further consider the noise and time-delay that are usually involved in the interactions, we can rewrite this as follows:

$$\frac{dx_i(t)}{dt} = \sum_{j=1}^{n} a_{ij}(x_j(t - \tau_j) + n_{ij}) \tag{5}$$

where τ_j indicates the time-delay accompanied with the influence of x_j and n_{ij} denotes the intrinsic noise of the interaction between x_j and x_i.

2.2. Numerical scheme

In this section, we present a numerical scheme to infer gene regulatory networks. The proposed scheme consists of three main components: the direct solver, the optimization routine, and the objective function. Each component is described below in detail and the flow of overall algorithm is illustrated in Fig. 1.

2.2.1. Direct solver

We call given gene expression data 'reference data' and the data produced by solving the ODE system 'numerical data'. To solve (4), finite difference approximations are given as follows:

$$x_i(t_{k+1}) = x_i(t_{k-1}) + dt(a_{i1}x_1)(t_{k-1}) + a_{i2}x_2(t_{k-1}) + \ldots + a_{in}x_n(t_{k-1})) \tag{6}$$

where $i = 1, \ldots, n$ and $k = 1, \ldots, k_{max}$.

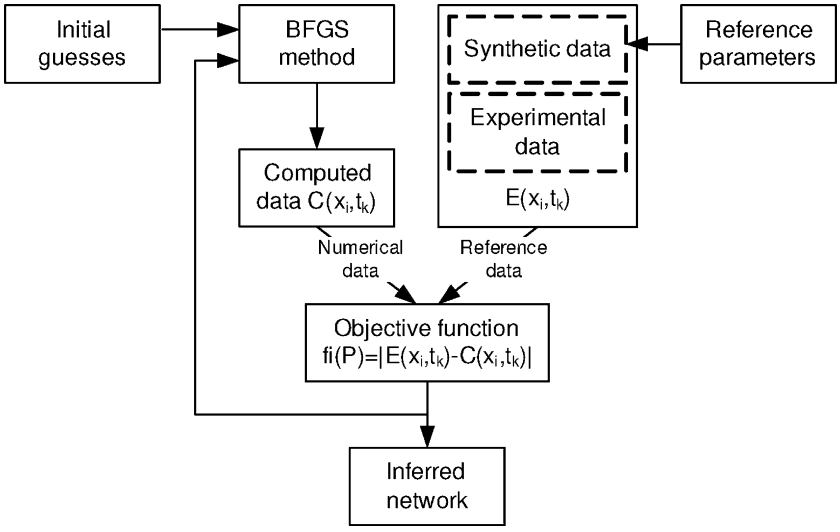

Fig. 1. A schematic diagram of the proposed numerical scheme.

2.2.2. Optimization routine

The data generated from the solver are used to optimize the unknown parameters. Among the various optimization algorithms, we employ the BFGS method as it is computationally most efficient. The basic idea is to construct an estimation of the inverse Hessian matrix to determine the next step iteration point (Press et al., 1992). Although estimating the Hessian matrix is complicated, the computational advantage outperforms any other available numerical optimization methods.

2.2.3. Objective function

The objective function is used to find the global minimum by integrating error norms. The objective function is constructed so that the norm of the difference between the reference data and the numerical data is equal to zero when parameters of (5) are accurately estimated. Hence, the objective function provides a stop condition for the iteration. The following three norms are considered for comparison:

$$L_1(A) = \sum_{i=1}^{n}\sum_{k=1}^{k_{\max}} |E_{i,k} - C_{i,k}|, \tag{7}$$

$$L_2(A) = \sum_{i=1}^{n}\sum_{k=1}^{k_{\max}} (E_{i,k} - C_{i,k})^2, \tag{8}$$

$$L_\infty(A) = \max_{i,k} |E_{i,k} - C_{i,k}| \tag{9}$$

where $A = [a_{11}, a_{12}, \ldots, a_{nn}]$, $E_{i,k}$ represents the reference data and $C_{i,k}$ is the numerical data of gene i at time k.

3. Results

To illustrate the proposed numerical scheme, we consider several examples—three examples based on synthetic data and two examples with real data sets. For the synthetic data example, we revisit the benchmark test example of four-node gene network (Sontag et al., 2004) where the gene interactions are summarized as follows:

$$A = \begin{bmatrix} -1.3 & -1.0 & 0.0 & 0.5 \\ 0.0 & -1.7 & 0.0 & 0.5 \\ -0.5 & 0.4 & -2.9 & 0.0 \\ 0.0 & 0.0 & 3.0 & -2.0 \end{bmatrix}$$

In order to compare the different inference results depending on norms employed, we considered L_1, L_2, and L_∞ norms in the objective function. It turns out that L_1 norm outperforms the others in the case of four-node networks, but L_2 norm brings out more reliable results for an eight-node network.

3.1. Construction of a gene network with noisy data

In order to examine the effect of noise in the reference data, we have implemented the BFGS method by adding 0, 1, 2, 5 and 10% noise to the reference data, respectively, and using L_1 norm for the objective function with zero initial conditions. The results are presented in Table 1. Without noise, the proposed numerical scheme can estimate the exact parameter values and thereby can infer a gene network with all the correct regulatory relations and interaction strengths. For a noise level below 2%, we can uncover 11 out of 16 regulatory relations with less than 10% estimation errors for interaction strengths. For 5 or 10% noise levels in the reference data, we can uncover 68.75% regulatory relations. These inference results can be, however, further improved if more data samples are used.

Table 1
Inference results for various noise levels in the reference data where 'a', 'b', 'c', 'd', and 'e' represent 0, 1, 2, 5, and 10% noise levels, respectively, and 's' represents the inference result of the previous method proposed by Sontag et al. (2004)

-1.3	-1.1	-1.0	-1.0	0.0	0.0	0.5	0.3
-1.0	-0.4	-1.1	-1.3	-0.1	-0.3	0.2	-0.4
-0.1	-2.0	-1.3	-1.1	-0.3	0.0	-0.9	0.5
0.0	0.0	-1.7	-1.7	0.0	0.0	0.5	0.6
0.0	0.0	-1.7	-1.8	0.0	-0.1	0.6	0.7
-0.1	0.0	-2.0	-0.3	-0.4	0.0	1.0	0.4
-0.5	-0.5	0.4	0.4	-2.9	-2.9	0.0	0.0
-0.5	-1.1	0.4	0.7	-2.9	-2.5	0.0	0.5
-0.7	-0.5	0.5	0.4	-2.7	-3.7	0.2	0.0
0.0	0.3	0.0	0.0	3.0	3.1	-2.0	-1.7
-0.6	-1.0	0.2	0.0	3.2	2.8	-1.4	-1.7
-0.1	0.0	0.0	0.0	2.6	2.6	-1.7	-1.5

a	b
c	d
e	s

3.2. Construction of a gene network with time-delayed data

Time-delay is one of the critical factors that should be considered in reconstruction of gene regulatory networks. We address this issue by introducing a time-delay parameter to the interac-

Table 2
Inference results of the cases with three different time-delays in the reference data where 'a' denotes the case without time-delay, 'b' indicates the case with randomly estimated time-delay, and 'c' represents the case with known exact time-delay

-1.3	-1.3	-1.1	-1.0	0.0	0.0	0.5	0.5
-1.3		-1.0		0.0		0.5	
0.0	0.0	-1.8	-1.7	0.0	0.0	0.5	0.5
0.0		-1.7		0.0		0.5	
-0.5	-0.5	0.4	0.4	-3.1	-3.0	-0.1	0.0
-0.5		0.4		-2.9		0.0	
0.0	0.0	0.0	0.0	3.3	3.1	-2.0	-2.0
0.0		0.0		3.0		-2.0	

a	b
c	

Table 3
Inference results for various noise levels with known exact time-delay where 'a', 'b', 'c', and 'd' represent 1, 2, 5 and 10% noise levels in the reference data, respectively

-1.1	-0.9	-1.1	-1.1	-0.1	-0.1	0.2	0.1
-0.2	-0.1	-1.3	-1.1	-0.3	0.0	-0.7	-1.1
0.0	-0.2	-1.7	-1.6	0.0	0.2	0.5	0.6
-0.5	0.2	-1.3	-1.7	0.4	0.0	0.7	0.1
-0.5	-0.5	0.4	0.4	-2.9	-2.9	0.0	0.0
-0.6	-0.6	0.5	0.5	-2.8	-2.8	0.0	0.0
-0.1	-0.2	0.1	0.1	3.1	3.2	-1.9	-1.9
0.0	-0.4	0.1	0.3	3.3	3.8	-2.0	-1.8

a	b
c	d

tion between x_j and x_i as follows:

$$\frac{dx_i}{dt} = \sum_{k=1}^{k_{max}} a_{ij} x_j(t_k - \tau_j), \quad i, j = 1, \ldots, 4 \tag{10}$$

To examine the effect of time-delay on the inference result, we considered three cases: the case without time-delay, the case with randomly estimated time-delay, and the case with known exact time-delay. Table 2 shows the inference results. We found that the proposed numerical scheme successfully uncovered gene interactions in case with known exact time-delay. Moreover, we also have reasonably good inference results under without time-delay or randomly estimated time-delay. This implies that the proposed numerical scheme can provide us with a reliable inference result with respect to unknown time-delay in cases with a small-scale network and bounded short time-delay.

3.3. Construction of a gene network with noise and time-delayed data

To address both time-delay and noise, we considered noisy synthetic data with time-delay and applied the proposed numerical scheme by using L_1 norm. The time-delay is assumed known and the inference results are shown in Table 3. For the noise level below 2%, 9 out of 16 regulatory relations were unconvered with less than 10% estimation errors for the interaction strengths. For the noise level above 5%, we could uncover 62.5% regulatory relations. These results are only slightly worse than the case with known exact time-delay. In the case of the 2% noise level without time-delay, the interactions were inferred as: (−0.6, −1.3, −0.3, −0.1, −0.5, −1.5, 0.3, 0.9, −0.5, 0.5, −3.0, −0.1, 0.0, 0.0, 3.4, −2.0). On the other hand, the interaction strengths were inferred as follows for the 2% noise level with randomly given time-delay: (−0.9, −1.1, −0.1, 0.0, −1.1, −1.6, 0.1, 0.5, −0.5, 0.4, −3.0, 0.0, −0.1, 0.0, 3.3, −1.9). In both cases, the proposed numerical scheme uncovered 11 and 10 out of 16 regulatory relations, respectively. However, the inferred interaction strengths in the two cases were much weaker than those with the known exact time-delay.

3.4. Construction of a gene network with real experimental data

We applied the proposed numerical scheme to infer a gene regulatory network of budding yeast cell cycle from microarray experimental data. In the budding yeast cell cycle, 17 regulatory genes were discovered (Bahler, 2005) and these genes were considered in our study. In particular, two network models were considered: four-node network composed of Clb1, Fkh2, Sic1, and Swi5, and eight-node network composed of Cln3, Whi5, Swi4, Cln1, Swi6, Clb1, Mcm1, and Mbp1. Among the three major phases of cell cycle transition, four nodes are selected from G2/M transition (entry into mitosis) and M/G1 transition (exit from mitosis), respectively. As the transition of these phases are not yet fully understood, the present study might provide a

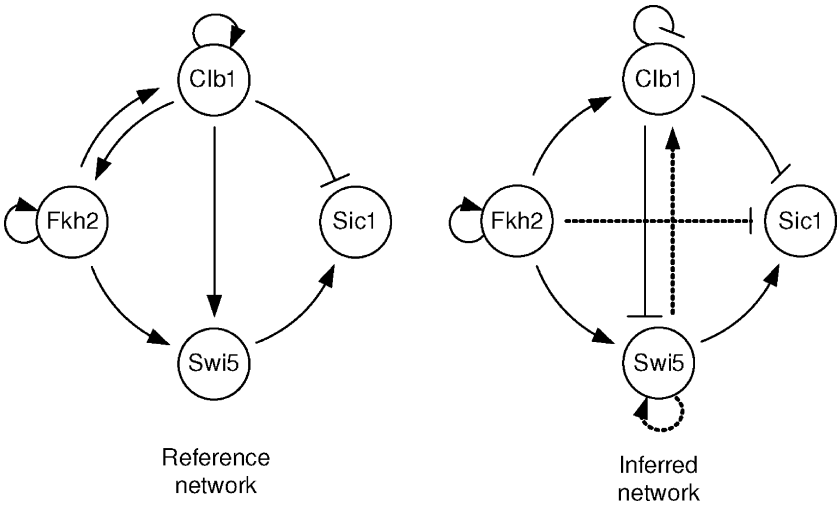

Fig. 2. The reference (left panel) and the inferred (right panel) gene regulatory networks of four nodes where arrows indicate activation and blunt lines denote inhibition. In the inferred network, solid lines indicate true positives while dotted lines indicate false positives.

useful insight into the hidden gene regulatory network of the transition mechanism. In the case of eight-node network, the 17 genes (scattered in all three phases) were classified into 8 clusters using k-means clustering algorithm (Draghici, 2003). One representative gene was then selected from each cluster for further consideration of a network of clusters.

The reference network and the inferred network are shown in Fig. 2, where arrows indicate activation and blunts denote inhibition. The real data sets and reproduced numerical data sets are compared in Fig. 3. The proposed numerical scheme has successfully reconstructed five out of eight probable connections with an identification accuracy of 62.5%. We found that 2 activating interactions were identified as inhibiting interactions. Compared with the reference network, the inferred network turns out to have three more new interactions which need further experimental verification. The reconstructed interaction network is as follows:

$$\begin{array}{c|cccc} & Clb1 & Fkh2 & Sic1 & Swi5 \\ \hline Clb1 & -47.13 & 44.97 & -5.13 & 40.52 \\ Fkh2 & 2.23 & 12.95 & -1.13 & -9.83 \\ Sic1 & -24.57 & -35.64 & 3.20 & 52.91 \\ Swi5 & -34.96 & 42.76 & -3.85 & 25.97 \end{array}$$

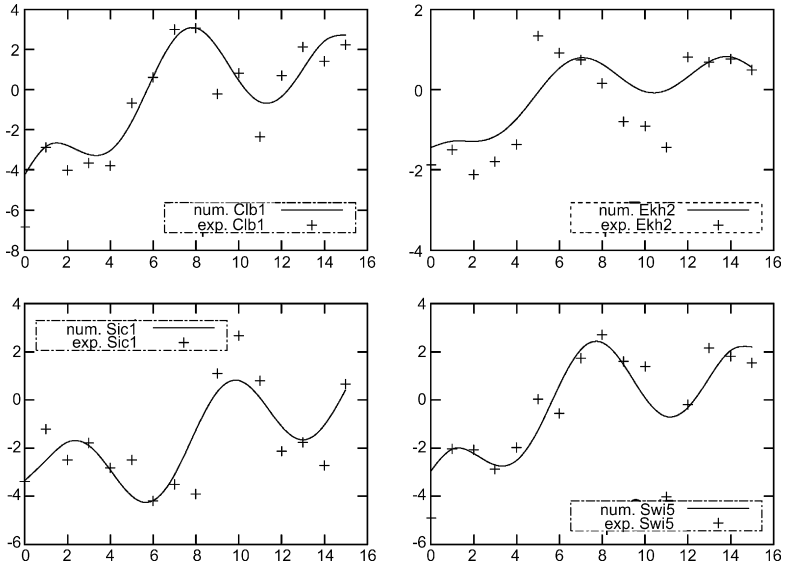

Fig. 3. Comparison of given real data sets (exp.) and reproduced numerical data sets (num.) of four-node model where the numerical data are produced from the inferred network model.

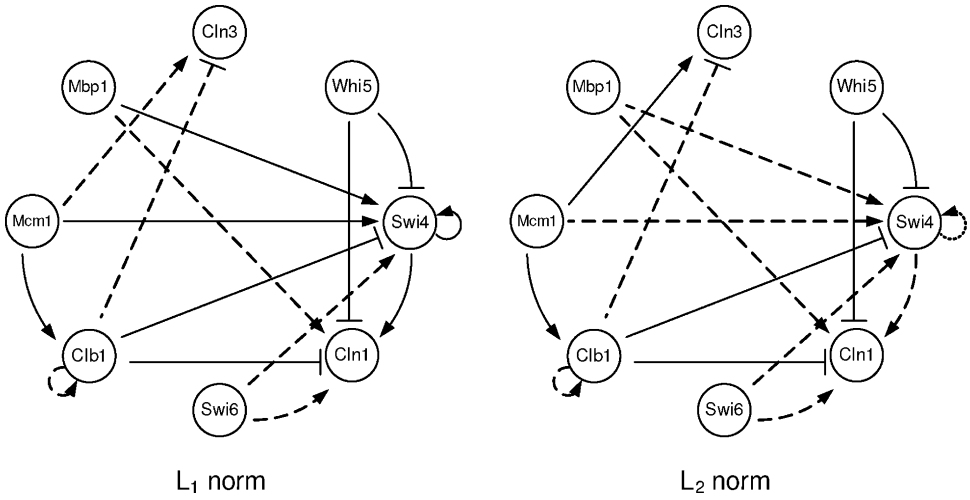

Fig. 4. The inferred network of eight nodes by using L_1 norm (left panel) and L_2 norm (right panel) where dotted lines denote the true interactions and solid lines indicate the correctly identified interactions.

Next, we have applied the proposed numerical scheme to the eight-node network using L_1 and L_2 norms. The results are shown in Fig. 4. If we consider only the interacting relationships, using L_1 norm was better in recovering the gene network than using L_2 norm. Fig. 5 shows the receiver operating characteristic (ROC) (Schroeder et al., 2006) curves explaining the reliability of the methods with respect to different thresholds used. Considering the true positive and false positive rates, we found that the proposed numerical scheme has successfully recovered the network in all the cases. However, if we consider the interaction relationships and strengths, we find that the proposed numerical scheme works better for a smaller network.

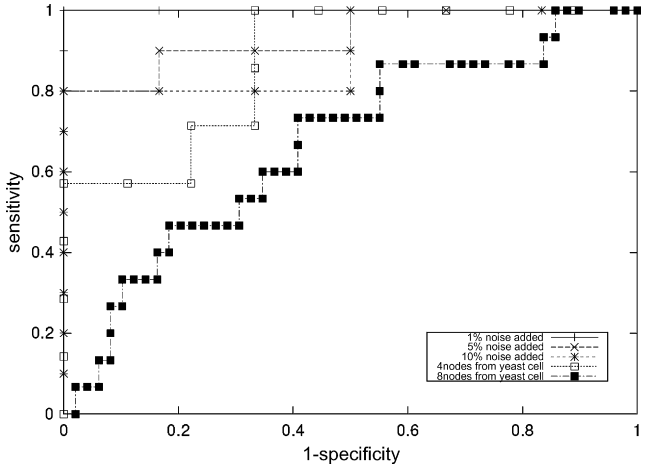

Fig. 5. Receiver operating characteristic (ROC) curves for the inference results on the synthetic data with the noise levels of 1, 5, and 10%. The ROC curves for the inference results on experimental data of four- and eight-node models are shown as well. The ROC curves illustrate whether the proposed numerical scheme is robust and reliable with respect to the threshold used. We found that the inference results on the synthetic data are more reliable for a smaller noise level while the inference results on the experimental data are less robust as the network size gets larger. All the inference results show at least consistency with respect to the variation of thresholds.

4. Discussion

Following the development of new measurement technologies and computational tools, there is a pressing need to develop a method for inferring gene regulatory networks from measured expression profiles. The present study has proposed an optimization based numerical analysis scheme that can identify gene interactions from time-series expression profiles without *a priori* knowledge on the network structure. In particular, we have taken the time-delay and noise into account such that the proposed numerical scheme can be applied to real data sets. The proposed approach has advantages of adopting a limited number of experimental data points and inferring a dynamic gene network model from the raw data. By applying the proposed numerical scheme, we could successfully uncover a small network of four nodes in the case of synthetic data. The gene regulatory network of budding yeast cell cycle needs to be further verified by another experiments. In the case of the larger network with eight nodes, the inference result was not so good as in the case of the smaller network, but we confirmed that the true positive and true negative interactions were successfully uncovered. Further studies are required to improve the proposed numerical scheme for application to large-scale networks.

Acknowledgments

This work was supported by a grant from the Korea Ministry of Science and Technology (Korea Systems Biology Research Grant, M10503010001-05N030100111), by the 21C Frontier Microbial Genomics and Application Center Program, Ministry of Science & Technology (Grant MG05-0204-3-0), Republic of Korea, and in part by 2005-B0000002 from the Korea Bio-Hub Program of Korea Ministry of Commerce, Industry & Energy. J. Kim and K.-H. Cho were supported by the second stage Brain Korea 21 Project in 2006.

References

Andrec, M., Kholodenko, B., Levy, R., Sontag, E., 2005. Inference of signaling and gene regulatory networks by steady-state perturbation experiments: structure and accuracy. J. Theor. Biol. 232, 427–441.

Bahler, J., 2005. Cell-cycle control of gene expression in budding and fission yeast. Annu. Rev. Genet. 39, 69–94.

Bansal, M., Gatta, G., Bernardo, D., 2006. Inference of gene regulatory networks and compound mode of action from time course gene expression profile. Syst. Biol. 22, 815–822.

Brazhnik, P., 2005. Inferring gene networks from steady-state response to single-gene perturbations. J. Theor. Biol. 237, 427–440.

Chen, T., He, H.L., Church, G.M., 1999. Modeling gene expression with differential equations. Pac. Symp. Biocomput. 4, 29–40.

Dasika, M.S., Gupta, A., Maranas, C.D., 2004. A mixed integer linear programming (MILP) framework for inferring time delay in gene regulatory networks. Pac. Symp. Biocomput., 474–486.

de Jong, H., 2002. Modeling and simulation of genetic regulatory systems: a literature review. J. Comput. Biol. 9 (1), 67–103.

Draghici, S., 2003. Data Analysis Tools for DNA Microarrays. Chapman & Hall/CRC.

Gardner, T., Bernardo, D., Lorenz, D., Collins, J., 2003. Inferring genetic networks and identifying compound mode of action via expression profiling. Science 301, 102–105.

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

Kimura, S., Ide, K., Kashihara, A., Kano, M., Hatakeyama, M., Masui, R., Nakagawa, N., Yokoyama, S., Kuramitsu, S., Konagaya, A., 2005. Inference of S-System models of genetic networks using a cooperative coevolutionary algorithm. Bioinformatics 21, 1157–1163.

Liu, Q., Cao, J., 2006. Improved global exponential stability criteria of cellular neural networks with time-varying delays. Math Comp. Model. 43, 423–432.

Press, W.H., Flannery, B.P., Teukolsky, S.A., Vetterling, W.T., 1992. Numerical Recipes in Fortran, second ed. Cambridge University Press.

Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., Lightfoot, S., Menzel, W., Granzow, M., Ragg, T., 2006. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol. Biol. 7 (3).

Smolen, P., Baxtes, D.A., Byrne, J.H., 2000. Modeling transcriptional control in gene networks-methods, recent results and future directions. Bull. Math Biol. 62 (2), 247–292.

Sontag, E., Kiyatkin, A., Kholodenko, B., 2004. Inferring dynamic architecture of cellular network using time series of gene expression, protein and metabolite data. Informatics 20, 1877–1886.

Spell man, P.T., Sherlock, G., Hang, M., Ayer, V.R., Anders, K., Essen, M.B., Brown, P.O., Bot stein, D., Fetcher, B., 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharides cerevisiae* by microarray hybridization 9, 3273–3297.

Tegner, J., Stephen Yeung, M.K., Collins, J., 2003. Reverse engineering gene networks: integrating genetic perturbations with dynamical modelling. PNAS 100, 5944–5949.

Wahde, M., Hertz, J., 2000. Coarse-grained reverse engineering of genetic regulatory networks. Biosystems 55, 129–136.

Yeung, M.K., Tegner, J., Collins, J.J., 2002. Reverse engineering gene networks using singular value decomposition and robust regression. PNAS 99, 6163–6168.

Xu, S., Lam, J., Ho, D., Zou, Y., 2005. Delay-dependent exponential stability for a class of neural networks with timedelays. J. Comput. Appl. Math. 183, 16–28.