Unraveling the functional interaction structure of a biomolecular network through alternate perturbation of initial conditions

Kwang-Hyun Cho a,b,⁎, Hyung-Seok Choi c, Sang-Mok Choo d,*

a College of Medicine, Seoul National University, Jongno-gu, Seoul, 110-799, Republic of Korea
b Bio-MAX Institute, Seoul National University, Gwanak-gu, Seoul, 151-818, Republic of Korea
c Interdisciplinary Program in Bioinformatics, Seoul National University, Gwanak-gu, Seoul, 151-747, Republic of Korea
d School of Electrical Engineering, University of Ulsan, Ulsan, 680-749, Republic of Korea

Received 7 November 2006; received in revised form 20 December 2006; accepted 20 January 2007

Abstract

Various approaches attempting to infer the functional interaction structure of a hidden biomolecular network from experimental time-series measurements have been developed; however, due to both experimental limitations and methodological complexities, a large majority of these approaches have been unsuccessful. In particular, with respect to the elucidation of such networks, there are (i) a dimensionality problem: too many network nodes with too few available sampling points, (ii) a computational complexity problem: exponential complexity if a priori information is unavailable for regulatory nodes, and (iii) an experimental measurement problem: no guidelines for an appropriate experimental design for distinguishing direct and indirect influences among network nodes. Here, we sought to develop a new methodology capable of identifying the correct functional interaction structure with only a few sampling points through relatively simple computations. We also attempted to provide guidelines for an experimental design capable of supporting this methodology by taking proper measurements of the direct influences among the network nodes.

In the present study, we considered an experiment where measurements were taken at two sampling time points with alternate perturbation (up-regulation or down-regulation) of initial conditions while keeping the same initial conditions for unperturbed network nodes, and propose a new method of identifying the functional interaction structure from such measurements. The proposed method is able to avoid the dimensionality problem caused by the practically limited number of sampling time points, and does not suffer from the computational complexity problem, as it only uses a simple algebra based on the Mean Value Theorem (see Supplementary mathematical descriptions) without any other complicated computation. In addition, we provide a detailed guideline for an experimental design that can take proper measurements of the direct influences among the network nodes through perturbation of initial conditions. The proposed method is particularly useful for cases investigating the local interaction structure around a specific network node of interest. An example, based on simulated data, is provided to illustrate the proposed method.

© 2007 Elsevier B.V. All rights reserved.

Keywords: Biomolecular networks; Functional interactions; Initial condition perturbations; Mean Value Theorem; Time-series measurements

1. Introduction

Unraveling the functional interaction structure of a biomolecular network is of crucial importance for investigating the biological function of a living system. As high throughput technology (e.g., DNA microarray) becomes available, identification of functional interaction structures of underlying biomolecular networks (e.g., gene regulatory networks) from time-series measurements is gaining attention. Several approaches have been developed for this purpose, including Boolean networks [17,19], Bayesian networks [3,10], dynamic

Abbreviations: TP, true positive; TN, true negative; FP, false positive; FN, false negative.
⁎ Corresponding authors. K.-H. Cho is to be contacted at the 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. S.-M. Choo, is to be contacted at the School of Electrical Engineering, University of Ulsan, Ulsan, 680-749, Republic of Korea. Tel.: +82 52 259 2312; fax: +82 52 259 1692.
E-mail addresses: ckh-sb@snu.ac.kr (K.-H. Cho), smchoo@ulsan.ac.kr (S.-M. Choo).
0165-022X/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jbbm.2007.01.008
Bayesian networks [15,21,31], differential equation models [13,28], and stochastic models [18]. Successful application of these methods depends heavily on the availability of experimental time-series data that accurately reflect the interaction properties of a network system, and in this regard, how to choose an appropriate number of sampling time points and the sampling rate has also been investigated [2]. Despite the aforementioned studies, a detailed set of guidelines for an appropriate experimental design that can provide suitable time-series data for such an identification purpose has not yet been developed. Consequently, various probabilistic or statistical approaches are reported more frequently in cases of insufficient data sets. In such cases, however, there remain a dimensionality problem (too many network nodes and only a very limited number of available sampling points), a computational complexity problem (inability to avoid exponential complexity when no extra information on regulatory network nodes is available), and an experimental measurement problem (no guideline for distinguishing direct and indirect influences among network nodes). Therefore, development of a new identification method, as well as a corresponding guideline for an experimental design, is necessary in order to overcome these problems.

With respect to the challenges presented above, an interesting area of recent progress was the use of a priori biological information to sort out potential regulatory network nodes in advance in order to overcome the computational complexity problem [31]. In most cases, however, such information is not always available, and thus, gene perturbation experiments have been employed to identify a set of genes affected by a specific gene perturbation. Unfortunately, conventional gene perturbation experiments are unable to provide data sets by which one can distinguish direct and indirect interactions, since both of these types of interactions are usually mixed together in the data set. To overcome this obstacle, various approaches have been proposed to discriminate those interactions from perturbation data [6,23,26,27,29] although in many cases identification of a gene regulatory network based on classical gene perturbation experiments has been unsuccessful. And additional approaches based on qualitative information of time-series data [5] or parameter perturbations [4,14,24] have been recently reported. In each of these approaches, however, there is a fundamental limitation, as they all attempt to identify only the direct interactions between network nodes from given time-series data in which both the direct and indirect interactions among the network nodes are already reflected.

Overcoming the limitation of identification should therefore become possible if experiments are designed such that the direct interactions among nodes are mainly reflected in the measured data, and a method of identifying the functional interaction structure based on the data obtained from those experiments is developed. To this end, we considered the following experiment. First, the initial values of network nodes (e.g., mRNA expression levels for a gene regulatory network) are measured, and sample values are taken. This process is repeated after perturbing the initial condition of one network node (e.g., up-regulating or down-regulating a particular gene of interest) while maintaining the same initial conditions for the other nodes. It should be noted that this is an important step that distinguishes the proposed perturbation experiment from conventional experiments [9,12,22,25,29]. This procedure is then repeated for all the other nodes whose functional interactions are to be investigated. The perturbation of initial conditions can be achieved through various experimental techniques. For example, considering a gene regulatory network, such perturbation can be done by regulating gene expression levels through gene-inducible systems (e.g., tetracycline on-off systems [11] or intracellular siRNA injection systems [7]). If a mathematical model describing this experimental procedure is formulated, we can then develop a theoretical framework of identifying the functional interaction structure based on the perturbation data.

Motivated from these ideas, we present a new method identifying the functional interaction structure of a biomolecular network and provide guidelines for experimental design. Although the proposed method uses only simple algebra based on the Mean Value Theorem (thereby allowing the method to avoid the computational complexity problem) the identification result is much better than previous approaches. Furthermore, the procedures are verifiable through rigorous mathematical definitions and proofs. The proposed method is particularly useful for cases where it is desirable to know whether a specific network node directly activates (induces) or inhibits (represses) the other network nodes. Hence, the method also allows for the efficient identification of the local sub-network structure of interest while simultaneously requiring much less experimental cost and computational time than previous holistic approaches. We note that, considering the number of perturbation experiments required, the proposed method can only be used to reconstruct relatively small networks, or small parts of larger networks. The proposed method is illustrated by an example of the Dictyostelium chemotactic signaling network based on the data generated from a mathematical model representing activation and inhibition kinetics of the network components [20].

2. Materials and methods

2.1. Inferring the functional interaction between network nodes from the data of incipient time steps

We consider the following dynamic equations for a biomolecular network:

\[
\frac{dx_i(t)}{dt} = f_i(x_1, x_2, \ldots, x_n), \ t \in [0, T], \ i = 1, 2, \ldots, n
\]  

(1)

where the variable \(x_i\) is the \(i\)-th network node denoting the biochemical quantity of an element (e.g., gene, protein, metabolite or a functional module), and the corresponding function \(f_i\) describes how the rate of change of \(x_i\) with respect to time depends on all of the variables in the network. From Eq. (1), we can identify the interaction structure of a network if we determine the sign of \(f_{ij}(x(t)) = \frac{\partial f_i(x(t))}{\partial x_j}\) (1≤i,j≤n) at some time \(t\), under the assumptions that there exist all first-order
partial derivatives of $f$ and the sign of $f_{ij}(\alpha(t))$ is time-invariant (i.e., the interaction structure is time-invariant) although $f_{ij}(\alpha(t))$ can be time-varying. These assumptions have been implicitly made in most of the related studies (e.g., the interaction structures of signaling networks such as MAPK cascades [14] and gene regulatory networks [24] were considered time-invariant in most literature). Specifically, we define that a given node $x_i$ directly affects a node $x_j$ if and only if $f_{ij} \neq 0$. In particular, if $f_{ij} > 0$, then this implies that the node $x_j$ directly activates (or induces) the node $x_i$ by increasing the net rate of $x_i$ and if $f_{ij} < 0$ the node $x_j$ directly inhibits (or represses) the node $x_i$. In addition, we define that a node $x_i$ does not affect a node $x_j$ if and only if $f_{ij} = 0$, which is also taken to mean that $x_j$ has no direct influence on $x_i$. Hence, we need only to develop an identification method and corresponding experimental design capable of inferring the sign of $f_{ij}$, as this determines the functional interaction structure of a given biomolecular network being considered.

To illustrate the main idea of the proposed method, let us consider a very simple example of a two node network ($n = 2$). In this case, we can represent the experimental situation in which the sign of $f_{1i}(i=1,2)$ can be determined as follows:

$$x_i^1(t) = f_i(x_i^0, x_j^0)(i = 1, 2) \text{ with initial values } x_i^0(0), x_j^0(0),$$

$$x_j^1(t) = f_j(x_i^0, x_j^0)(i = 1, 2) \text{ with initial values } x_i^1(0), x_j^1(0)$$

where $x_i^0(0) \neq x_j^0(0)$ and $x_j^0(0) = x_j^1(0)$. Note that the two differential Eqs. (2) and (3) represent the same network structure, as Eq. (3) was developed from Eq. (2) by replacing the initial condition of the first node (i.e., $x_i^0(0)$ into $x_i^1(0)$). The subscript $i$ of $x_i^0$, $x_i^1$ ($i = 1, 2$) denotes the $i$-th node and the superscript 1 of $x_i$ indicates that the initial condition of the first node was changed (perturbed) from the initial condition of $x_i^0$.

Thus, the superscript 0 of $x_i$ represents the basis of perturbation. From the Mean Value Theorem (see Supplementary mathematical descriptions for its details), we know that there exists a constant $\zeta$ such that

$$x_i^0(0) - x_i^1(0) = f_i(x_i^0(0), x_j^0(0)) - f_i(x_i^1(0), x_j^1(0))$$

$$= f_i(x_i^0(0) + \zeta x_i^1(0), x_j^0(0)) - f_i(x_i^0(0) - x_i^1(0)).$$

Hence, $f_i(x_i^0(0) + \zeta x_i^1(0), x_j^0(0)) = \{x_i^0(0) - x_i^1(0)\}/\{x_i^0(0) - x_i^1(0)\}$. If the interaction structure is indeed time-invariant, then the sign of $f_{1i}$ (a, b) is fixed at all nonnegative values of $a$ and $b$, and therefore, the sign of $f_{1i}$ is identical with that of $\{x_i^0(0) - x_i^1(0)\}/\{x_i^0(0) - x_i^1(0)\}$. Thus, if we can obtain experimental data by which the sign of $\{x_i^0(0) - x_i^1(0)\}/\{x_i^0(0) - x_i^1(0)\}$ can be deduced, then we can determine the sign of $f_{1i}$ and unravel the overall interaction structure by integrating these results. Here, the point is that in order to infer the sign of $f_{1i}$, we only need experimental measurements to determine the sign of $\{x_i^0(0) - x_i^1(0)\}/\{x_i^0(0) - x_i^1(0)\}$ and thus do not need any other information about the network nodes. As the proposed method is based on this principle, we can determine the sign of $f_{ij}$ (denoted by $\text{sign}(f_{ij})$) from the four measurements $x_i^1(0), x_j^1(0), x_i^0(t_1), x_j^0(t_1)$ using a simple algebraic equation for a sampling time point $t_1$ as follows (this procedure is summarized in the Supplementary experimental design sketch):

$$\text{sign}(f_{ij}) = \text{sign}\left[\{x_i^0(0) - x_i^1(0)\}/\{x_j^0(0) - x_j^1(0)\}\right]$$

$$= \text{sign}\left(\frac{x_i^0(t_1) - x_i^0(0)}{x_j^0(0) - x_j^1(0)}\right).$$

The proposed method can therefore overcome most of the difficulties commonly encountered by previous approaches, such as dimensionality problem, the computational complexity problem, and the experimental measurement problem. Moreover, the proposed method can determine the sign of $f_{1i}$ through a simple computation regardless of the number of network nodes. For the remainder of this paper, we denote the solution of $x_i^1 = f_i(x_i^1, \cdots, x_n^1)$ where the subscript $i$ indicates the $i$-th node and the superscript $k$ means that the initial condition of the $k$-th node has been changed (perturbed) from that of $x_i^0$.

### 2.2. Experimental design

The proposed experimental design procedure is as follows:

**Step 1** Obtaining the initial experimental measurements: for the system of $x_i^0 = f_i(x_i^1, \cdots, x_n^1)(1 \leq i \leq n)$, measurements are taken of the initial condition $x_i^0(0)$ at $t = 0$ and $x_i^0(t_1)$ at another sampling time point $t = t_1$. Note that $t_1$ should be small enough such that $x_i^0(0)$ approximates $x_i^0(t_1)$ (a relatively larger value of $t_1$ is allowable if the interaction process is slow). The measured value of $x_i^0(0)$ ($1 \leq i \leq n$) is to be used as a basis for the second measurements in Step 3. Note that in general, one can choose a sampling time point $t = t_k$ instead of $t = t_1$ (for further details see Supplementary generalization: inferring the sign of $f_{ij}$).

**Step 2** Choosing a node $x_i$: if we want to unravel the functional interaction structure of a whole network with $n$ nodes, then any node can be chosen as $x_i$. If it is desirable to investigate the functional interaction structure of a local sub-network around a particular node of interest, then this node should be designated as $x_i$.

**Step 3** Taking the second experimental measurements for the node $x_j$: measurements of $x_j^0(0)$ and $x_j^0(t_1)$ are taken under the same initial conditions as in Step 1 only after changing (perturbing) $x_j^0(0)$ into $x_j^1(0)$ through up-regulation or down-regulation of the initial condition of the $j$-th node.

**Step 4** Identifying the true interaction sign of $f_{ij}$: from $x_i^0(0), x_i^0(t_1), x_j^0(0), x_j^1(0)$ obtained in Step 1 and Step 3 respectively, $\text{sign}(f_{ij})$ ($1 \leq i \leq n$) can be inferred from $\text{sign}(f_{ij}) = \text{sign}\left[\{x_i^0(t_1) - x_i^0(0)\}/\{x_j^0(t_1) - x_j^0(0)\}\right]$. To account for the effect of noise in real data, we can introduce thresholding of the data prior to application of this formula. Let us consider a threshold level $\alpha$ reflecting the (generally unknown) noise effect of real data; we suggest
Table 1
Eight sets of simulated data: \( A^0 (1 \leq j \leq 7) \) were obtained by changing \( x_j^0 (0) \) into \( x_j^0 (0) + 20 \) while preserving the same initial conditions with those of \( A^0 \).

<table>
<thead>
<tr>
<th>( A^i )</th>
<th>( A^j )</th>
<th>( A^k )</th>
<th>( A^l )</th>
<th>( A^m )</th>
<th>( A^n )</th>
<th>( A^o )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1, 0.264</td>
<td>2.1, 1.7802</td>
<td>2.1, 1.022</td>
<td>2.1, 1.030</td>
<td>2.1, 1.468</td>
<td>2.1, 1.5408</td>
<td>2.1, 1.4582</td>
</tr>
<tr>
<td>2.2, 2.166</td>
<td>2.2, 2.240</td>
<td>2.2, 2.177</td>
<td>2.2, 2.016</td>
<td>2.2, 7.285</td>
<td>2.2, 2.170</td>
<td>2.2, 2.170</td>
</tr>
<tr>
<td>1.6, 1.442</td>
<td>1.6, 1.451</td>
<td>21.6, 17.563</td>
<td>1.6, 1.409</td>
<td>1.6, 1.200</td>
<td>1.6, 2.493</td>
<td>1.6, 2.369</td>
</tr>
<tr>
<td>0.8, 0.746</td>
<td>0.8, 0.749</td>
<td>0.8, 0.198</td>
<td>0.8, 0.750</td>
<td>0.8, 0.715</td>
<td>0.8, 0.695</td>
<td>0.8, 0.695</td>
</tr>
<tr>
<td>1.2, 1.189</td>
<td>1.2, 1.776</td>
<td>1.2, 1.228</td>
<td>20.8, 17.527</td>
<td>1.2, 1.224</td>
<td>1.2, 1.224</td>
<td>1.2, 1.237</td>
</tr>
<tr>
<td>0.3, 0.297</td>
<td>0.3, 1.461</td>
<td>0.3, 0.296</td>
<td>0.2, 0.277</td>
<td>0.3, 0.290</td>
<td>0.3, 0.380</td>
<td>0.3, 0.380</td>
</tr>
<tr>
<td>1.7, 1.636</td>
<td>1.7, 2.841</td>
<td>1.7, 1.640</td>
<td>1.7, 1.637</td>
<td>1.7, 1.626</td>
<td>1.7, 28.185</td>
<td>21.7, 15.035</td>
</tr>
</tbody>
</table>

The two numbers in each element of \( A^i (0 \leq j \leq 7) \) represent the sample values at the initial sampling time point (\( t_0 = 0 \)) and at the second sampling time point (\( t_1 = 0.1 \)), respectively.

specifying this \( \alpha \) value together with the inference result.

Also, we can repeatedly apply the proposed method for different levels of \( \alpha \) if we do not have any a priori information on the noise effect.

**Step 5** Repeating the above procedure from Step 3 to Step 4: if we only want to unravel the direct interaction of the \( j \)-th node on the \( i \)-th node in particular, then we stop here. Otherwise, in cases where we want to further investigate the other functional interaction such as the direct interaction of the \( j \)-th node on the \( k \)-th node (\( k \neq i \)) then the above procedure from Step 3 to Step 4 is repeated for \( x_j^k \).

**Step 6** Repeating the above procedure from Step 3 to Step 5: in order to further investigate other functional interaction such as the direct interaction of \( x_m \) (\( m \neq j \)) on the \( i \)-th node, then the above procedure from Step 3 to Step 5 is repeated by replacing \( j \) with \( m \).

As one can see from the above procedure, it is necessary to perform the calculations only once if we want to unravel the direct interaction of the \( i \)-th node on the \( j \)-th node. That is, we only need to take four measurements of \( x_j^0 (0), x_j^1 (0), x_j^1 (t_1), x_j^1 (t_1) \) in order to probe whether or not the \( j \)-th node is a regulator of the \( i \)-th node. Hence, tailor-made experiment can be designed depending on the interaction substructure one wishes to unravel.

It should be noted that we can devise a more generalized experimental design by relaxing the sampling time point \( t_1 \) and the solution \( x_j^0 \), although these were fixed and used as a basis in the above procedure (for further details see Supplementary generalization: generalized experimental design).

**3. Results**

**3.1. Illustrative example: the chemotactic signaling network of Dictyostelium**

In order to illustrate the proposed method and to verify its usefulness and accuracy for a real biomolecular network, we considered the chemotactic signaling network of Dictyostelium [20] as a simple but real example (for an additional example, see Supplementary additional example which illustrates the proposed method through the synthetic oscillatory network adopted and modified from [8]). The network, which includes ERK2, ecAMP (external cAMP), icAMP (internal cAMP), CAR1 (cAMP receptor), ACA (adenylyl cyclase), PKA (cAMP-dependent protein kinase A), and RegA, was proposed on the basis of biochemical and genetic evidence [16,20]. A mathematical model representing activation and inhibition kinetics for each of the components was employed (see Supplementary mathematical model of Dictyostelium network), and we made use of this model to generate the required data set in Table 1, which we then used to identify the interaction structure of the Dictyostelium chemotactic signaling network (see Fig. 1). \( A^0 \) in Table 1 represents the initial condition plus one more sampled data at \( t_1 = 0.1 \), while \( A^i (1 \leq j \leq 7) \) refers to the data obtained by up-regulating the initial condition of the \( j \)-th node by +20 while keeping the same initial conditions for all the other nodes.

We assumed that any noise introduced would be proportional to the difference \( x(t_1) - x(0) \), which represents the relative increase with respect to a basal level \( x(0) \). Thus, in Table 1, we added random noise of a uniform distribution to the artificial data \( x(t_1) \) obtained from \( x_i = f_j(x) \) up to 20% of \( x(t_1) - x(0) \). Next, we removed data whose variation was less than twice (i.e., we set a threshold level of \( \alpha = 2 \), since a two-fold change is widely accepted as a basis for comparison) of the control (\( A^0 \) in Table 1) at \( t = t_1 \) and replaced the filtered values with those values of
corresponding nodes in $A^0$. We inferred the interaction structure by applying the sign($f_{ij}$) = \[ \text{sign} \left( \frac{x_i(t_1) - x_i(t_0)}{x_j(t_0) - x_j(t_0)} \right) \] (see Materials and methods) to the filtered data.

It should be noted that the $i$-th node $x_i^j$ denotes either $ACA_i^j$, $PKA_i^j$, $ERK2_i^j$, $REGA_i^j$, $icAMP_i^j$, $ecAMP_i^j$, or $CAR1^j$ presented in Table 1. The sign of the $(i, j)$ element in Fig. 2 should therefore be identical to the sign of $f_{ij}$. For example, sign($f_{61}$) = $\oplus$ in Table 2, as the $(6,1)$ element is +1.164 in Fig. 2. This implies that the first ($j = 1$) column node $ACA$ directly up-regulated the sixth ($i = 6$) row node ecAMP. Note that Fig. 2 was obtained by applying the above sign formula to the filtered data of Table 1 with a threshold level $\alpha = 2$ with respect to $A^0$. All the sign($f_{ij}$)'s in Table 2 were obtained in a similar way and the true signs were also denoted in the parentheses. As summarized in Table 2, 46 of the 49 total interactions were correctly identified, and the identified interaction structure is illustrated in Fig. 3. The three incorrect identifications, marked by grey boxes in Table 2, were caused by either the mixing of direct and indirect influences in the assumed data set ($(1,6)$ element) or relatively low expression levels compared to the given threshold ($(5,1)$ and $(3,7)$ elements).

### 3.2. Comparison of the proposed method and previous methodologies

To compare the proposed method with previous methodologies, we applied the Boolean network, the Bayesian network, and the dynamic Bayesian network approaches to the aforementioned example of the Dictyostelium chemotactic signaling network. The identification results are summarized in Table 3 (see Supplementary identification results for further details). We confirmed that the proposed method provided

![Fig. 3. Illustration of the identified interaction structure: solid lines denote the correct identifications and dotted lines indicate incorrect identifications. Note that the 'incorrect' identifications are not actually wrong, as incorrect interaction in this case means that there was no such 'direct' interaction. For instance, considering the incorrect interaction between ecAMP → ACA, there is indeed an indirect interaction from ecAMP to ACA via ecAMP → CAR1 → ACA. The reason for this incorrect assignment stems from nature of the (assumed) data measurements in which the direct influences, as well as some indirect influences, were mixed together according to the variation of initial conditions.](image)

<table>
<thead>
<tr>
<th>Boolean network</th>
<th>Bayesian network</th>
<th>Dynamic Bayesian network</th>
<th>Proposed method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identification rate</td>
<td>28/49 (57%)</td>
<td>14/49 (29%)</td>
<td>26/49 (53%)</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>3/17 (18%)</td>
<td>9/17 (53%)</td>
<td>6/17 (35%)</td>
</tr>
<tr>
<td>Specificity</td>
<td>25/32 (78%)</td>
<td>5/32 (16%)</td>
<td>20/32 (63%)</td>
</tr>
</tbody>
</table>

A Boolean network was implemented using the REVEAL algorithm with the maximal 6 input nodes (101 sampling time points were used for each node with the simulated data following the proposed perturbation experiment design. The average of the data was used for the threshold of discretization). The Bayesian network was implemented using the deal package (R package) and employing the greedy search algorithm based on 1000 parameter perturbations with a degree (i.e., the number of insertion/deletion/turning) of 100. The dynamic Bayesian network was implemented using the Mocapy Toolkit for parameter learning, the Gibbs sampling for the MCMC method, and the EM algorithm for optimization of parameters. In this case, we also used the deal package (R package) for structure learning and employed the greedy search algorithm based on 100 parameter perturbations with a degree of 7. All computations were performed using an HP xw8200 Workstation with two intel Xeon 3.6 GHz processors and 8 GB memory.
better identification results than any of the other previous methods with respect to the absolute measure of correct identifications, sensitivity, and specificity.

3.3. Remarks

Another aspect of the proposed method is that it can be used in combination with previous methodologies (e.g., Boolean network, probabilistic Boolean network, Bayesian network, and dynamic Bayesian network) to determine a set of possible regulators for each network node prior their application, thereby increasing their efficiency. As most of the previous methodologies are forced to assume all of the network nodes as potential regulators if a priori knowledge is unavailable, an unmanageable computational complexity problem is often encountered during attempts to apply such methods to real, large-scale data or even small networks consisting of several tens of nodes. This problem can be resolved by employing the proposed method to pre-filter the set of possible regulators for each node. For instance, if any of the previous methods are applied to the example in this study, then a total of $2^2 - 1 = 127$ states of possible regulators for a given target node must be considered. However, if the proposed method for pre-filtering is applied prior to the previous methods, the number of possible regulators for each network node is reduced to between $2^3 - 1 = 3$ and $2^4 - 1 = 15$ (see Table 2).

Currently, a real data set that has been obtained by the proposed perturbation experiment is unavailable. To illustrate the limitation of conventional perturbation experiments and the way of choosing $t_i$, we considered a real data set of *E. coli* obtained through over-expression of the $\sigma$ factor rpoH [22]. The data consist of gene expression profiles of 4290 annotated open reading frames (ORFs) out of the complete genome sequence of *E. coli* K-12 after exposure to a heat shock. The heat shock response of *E. coli* is controlled at a transcriptional level by the alternative $\sigma$ factors rpoH ($\sigma$32) and rpoE ($\sigma$E); the direct target genes of rpoH are known [9,12,29]. The identification result is summarized in Supplementary Table 8, but we find that the result is rather inferior, as the data were not obtained from the proposed perturbation experiment. Note that with the proposed method, we can choose $t_i$ as less than 10 min, as the indirect interactions are reflected in the data after 10 min (Supplementary Table 7).

4. Discussion

As more and more diverse experimental techniques that take measurements of interactions between intracellular molecules such as genes, proteins, metabolites, or functional modules become available, various approaches have been developed to infer the underlying functional interaction structure of bimolecular networks from measured data sets. Furthermore, there are a number of good reasons for giving increased attention to the identification of such interaction networks, as we can elucidate the nature of more complex biological functions if all of the information about the regulatory factors from an identified interaction structure is available. Such results could be used for identifying more efficacious drug targets by predicting the inhibitory effect of the potential target molecule, as well as the unexpected adverse effects based on the interaction structure. In practice, however, inferring the functional interaction structure of a complex large-scale biomolecular network is hindered by dimensionality and computational complexity problems. In addition, there has continually been a more fundamental problem to such attempts at inferring interaction structures, as it is necessary to identify only direct interactions between network nodes from given experimental data in which both the direct and the indirect influences among the network nodes are mixed.

To resolve the aforementioned problems in identification of interaction structures, we needed to develop not only a new identification methodology, but also a corresponding experimental design. To overcome the dimensionality problem and the computational complexity problem, we proposed a simple algebraic equation capable of identifying the direct influence of the $j$-th node on the $i$-th node (i.e., $\text{sign}(f_{ij})$) based only upon four measurements. This proposed formula does not require any complicated computations such as solving a large number of equations [14,24] or exhaustive computations of probabilities, instead requiring only simple calculus and thereby avoiding the computational complexity problem. Moreover, the proposed method does not suffer from the dimensionality problem, as it requires only two sampling measurements (i.e., the initial measurement at $t=0$ and another measurement at $t=t_i$) and does not rely on time-series measurements with many sampling time points. By this experimental design, the proposed method is able to rely only on the data mainly reflecting the direct interactions between network nodes. In addition, we can also design tailor-made experiments with minimal cost and time to unravel a specific region of a larger network. On the other hand, we acknowledge that the number of perturbation experiments required for this method can be very large in cases where one wishes to reconstruct an entire interaction network, as it is necessary to perturb each network node alternatively. In such cases, the number of perturbation experiments can be reduced by using a priori biological information, such as sequence information or binding motifs of transcription factors in gene regulatory networks [30]. One could also employ the perturbation strategy proposed by Tegnér et al. [26] in which it was proposed to consecutively perturb only some of the nodes that are most insensitively affected at each perturbation.

As we confirmed with the example in this paper (and another example in Supplementary information), the proposed method is capable of successfully inferring the functional interaction structure of a bimolecular network in a very efficient manner. In addition, the proposed method can also be used to pre-filter possible regulators through combination with previous identification methodologies in order to obtain better identification results or in applications of a large scale network system with many nodes. In particular, if we use the proposed method to filter out possible regulators for each network node and apply any of the previous identification methods, we can resolve most of the computational complexity problem and obtain much more accurate identification results for large-scale data sets.
5. Simplified description of the method and its (future) applications

We have proposed choosing an optimal sampling time point $t_1$, such that the majority of direct interactions among network nodes are reflected in the measured data. However, as the optimal value of $t_1$ depends on the hidden dynamics of interaction processes, choosing $t_1$ becomes difficult in cases where there is no available a priori information about the dynamics. In such cases, it may be necessary to infer an optimal $t_1$ (see Supplementary real data set of E. coli). Moreover, consideration of time-delay effects accompanied with the interaction mechanisms and a more thorough statistical extension of the proposed method to take account of measurement noises remain as further studies. Combining the proposed method with other bioinformatic approaches such as motif search may also add value to its applications and should be further investigated. For example, it is possible to further refine the inference result of a gene regulatory network by removing the false direct targets, such as those whose core motifs differ from the perturbed transcription factor, from the possible direct regulation targets. Perturbation dynamics is another issue that should be studied further. For example, perturbations made with siRNA can have their own dynamics caused by transfection. As such perturbation dynamics are not quantitatively known in many cases, this area should be addressed in the future.

Acknowledgments

This work was supported by the 2006 Research Fund of the University of Ulsan.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jbmb.2007.01.008.

References