Structural Equation Modeling for Systems Biology

Sachiyo Aburatani
Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), Japan

Hiroyuki Toh
Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), Japan

INTRODUCTION

Inference of gene regulatory network is a useful approach to understand the control systems in living cells. To obtain the better insights into the gene regulation, we developed statistical approach, based on Structural Equation Modeling (SEM) in combination with factor analysis and improved four-step procedure. This approach allowed us to reconstruct a model of transcriptional regulation that involves protein-DNA interactions from only the gene expression data, in the absence of protein information. The significant features of SEM are the inclusion of latent variables within the constructed model and the ability to infer the network, including its cyclic structure. Furthermore, the SEM approach allows us to strictly evaluate the inferred model by using fitting scores. In this article, we'll show the details of SEM approach for detecting the causality between genes and other cellular components.

BACKGROUND

One of the essential systems in a living cell is gene expression which is controlled for surviving and adapting to various conditions by translating proteins when needed. Actually, many cellular events, such as cell division, cell differentiation and specific reactions to several conditions, are accomplished by the cooperative expression of related genes. The system of gene expression is usually started by transcription, which is the process of creating complementary RNA copies of a protein-coding gene. Thus, clarifying the transcriptional control system is an important task for uncovering the cellular mechanism. The problem for clarifying the transcriptional regulation is that transcription levels of protein-coding genes are adjusted by various intercellular components, including DNA, RNA and proteins (Brazhnik et al., 2002). Transcriptional regulation in eukaryotes is more complicated by combinatorial interactions between several transcription factors, intracellular signals and extracellular conditions. Furthermore, the gene expressions during developmental stage, such as embryogenesis in a fertilized egg, are known to be controlled spatially and temporally by the regulated translation of stored maternal mRNAs and the accommodation of protein activity (Oh et al., 2000; Ogura et al., 2003). Since the transcription levels of genes are controlled in a complicated manner, revealing the transcriptional control mechanisms among the genes is one of the central themes in systems biology.

One of the most common mechanisms of gene expression control employs protein-DNA interactions. The proteins that bind to a DNA sequence to control transcription from DNA to mRNA are known as transcription factors (Latchman, 1997; Karin, 1990). They are crucially involved in regulating their target genes, by binding to a specific adjacent DNA sequence (Mitchell & Tjian, 1989). To gain a better understanding of the transcriptional regulation mechanisms among genes, a gene regulatory network is useful, and detailed knowledge about how these transcription factors conduct this regulation is essential when inferring a gene regulatory network.

DOI: 10.4018/978-1-4666-5888-2.ch044

MAIN FOCUS OF THE ARTICLE

Issues, Controversies, Problems

Investigations of gene regulatory systems or complex functional networks among DNA, RNA, proteins and other cellular components in a living cell conventionally follow a standard protocol. After a DNA sequence is completed, the mRNA level is measured by a cDNA microarray, to reveal the gene expression profiles under various conditions. From this information, various algorithms, including Boolean and Bayesian networks, have been developed to infer complex gene networks (Akutsu et al., 2000; Friedman, et al., 2000). In our previous investigation, we developed an approach based on graphical Gaussian modeling (GGM) combined with hierarchical clustering (Aburatani et al., 2003). We can infer the huge network among all of the genes by the GGM approach, since this approach is suitable for massive amounts of gene expression data (Aburatani & Horimoto, 2005). However, GGM infers only the undirected graph, whereas the Boolean and Bayesian models infer the directed graph, which shows causality. Although all of these approaches are feasible for establishing relationships among genes, it is difficult to reveal the critical interactions between genes and the other cellular components, owing to the insufficient information about the other cellular components in the gene expression profiles. Since estimation of regulatory network between genes and the other cellular components is absolutely essential to uncover the mechanism of gene expression control, an alternative approach is needed.

Solutions and Recommendations

To infer the relationships between genes and the other cellular components, structural equation modeling (SEM) is one possible technique (Bollen, 1989), which has been successfully used to elucidate causal relationships in disparate fields. In general, SEM approach has been utilized in econometrics, sociology and psychology (Haavelmo, 1943; Duncan, 1975; Pearl, 2001) fields for analyzing questionnaire data and measured numerical data. In biological field, SEM has been applied to quantify trait loci (QTLs) for association and linkage mapping (Liu, et al., 2008; Aten et al., 2008), as well as to identify genetic networks from microarray or SNP data (Shieh et al., 2008; Xiong et al., 2004). One of the significant features of SEM is the inclusion of latent variables into the constructed model, which allows the inferred model to include cellular components as latent variables, and genes as observed variables. Additionally, this method is able to infer the complicated network, including cyclic structure. In the constructed model, linear relationships among variables are assumed to minimize the differences between the fitted covariance matrix and the calculated sample covariance matrix. Some fitting indices are defined for evaluating the model adaptability, and thus the most suitable model can be selected by SEM.

Recently, we developed a new statistical approach, based on SEM in combination with factor analysis and altered four-step procedure, to infer hierarchical transcriptional regulation system mediated by transcription factor proteins from only the gene expression profiles, in the absence of protein information (Aburatani, 2011). Our SEM approach is explained below.

Factor Analysis

To assume the model structure, we selected the optimal number of factors for inclusion in the network model as latent variables, by performing a factor analysis. Factor analysis is a statistical method for describing the variability among observed variables in terms of a potentially lower number of latent variables (Spirtes et al., 2001). The initial assumption is that any observed variables may be related to any latent variables. Let us assume that there are p latent variables and q observed variables $x_1, x_2, \dots x_q$, with means $u_1, u_2, \dots u_q$. Note that the number p of latent variables is always smaller than the number q of observed variables. Each observed variable is expressed as linear combinations of p latent variables, as follows:

$$x_i\text{-}u_i=a_{i1}f_1+a_{i2}f_2+\dots+a_{ip}f_p+e_i \qquad (1)$$

where x_i is the vector of the expression levels of the gene i, a_{ip} is the partial regression weight of the latent variable f_j, and e_i is an independently distributed error term with zero mean and finite variance. In matrix form, Equation (1) is expressed as