Available online at www.sciencedirect.com
SCIENCE DIRECT®
BioSystems 83 (2006) 125–135
BioSystems
www.elsevier.com/locate/biosystems

Prediction of regulatory pathways using mRNA expression and protein interaction data: Application to identification of galactose regulatory pathway

A. Darvish, K. Najarian *

College of Information Technology, University of North Carolina at Charlotte, Charlotte, NC 28223, USA

Received 31 December 2004; received in revised form 11 May 2005; accepted 8 June 2005

Abstract

We propose a novel technique that constructs gene regulatory networks from DNA microarray data and gene–protein databases and then applies Mason rule to systematically search for the most dominant regulators of the network. The algorithm then recommends the identified dominant regulator genes as the best candidates for future knock-out experiments. Actively choosing the genes for knock-out experiments allows optimal perturbation of the pathway and therefore produces the most informative DNA microarray data for pathway identification purposes. This approach is more practically advantageous in analysis of large pathways where the time and cost of DNA microarray data experiments can be reduced using the proposed optimal experiment design. The proposed method was successfully tested on the galactose regulatory network.

Keywords: DNA microarray; Knock-out experiments; Gene-pathway

1. Introduction

One of the most challenging tasks in bioengineering is the estimation of regulatory interactions among the genes involved in a particular pathway using the data provided by high-throughout assays. This problem is often too complicated to have a definite solution. In the recent years, several methods have been introduced to address this problem. These methods include Boolean networks (Akutsu et al., 1999; Liang et al., 1998), Bayesian networks (Friedman et al., 2000; Setter et al., 2003), dynamic Bayesian networks (Thieffry and Thomas, 1998), linear models (Dhaeseleer et al., 1999), compartment modeling using differential equations (Chen et al., 1999), techniques based on control theory (Kholodenko et al., 2002), full biochemical interaction models (Arkin et al., 1998) and methods using metabolic regulation concepts (Klipp et al., 2002; Heinrich and Schuster, 1996; de la Fuente et al., 2002; Vo et al., 2004). The above-mentioned methods have span wide range of computational complexity. While some of these techniques such as Boolean networks are too abstract and simplified, others are computationally complex. The complex models include techniques modeling all biochemical interactions among genes based on a large number of differential equations.

The main advantage of the Boolean models is that they can handle a large number of genes and incorporate their binary interactions in the resulting network. However, the Boolean models express the activation of each gene as a binary value. This oversimplification results

* Corresponding author.
E-mail addresses: adarvish@uncc.edu (A. Darvish), knajaria@uncc.edu (K. Najarian).

0303-2647/$ – see front matter © 2005 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2005.06.013

to the loss of some important details about the pathway under study. In other words, considering each gene as an active or inactive unit does not allow the discovery of many aspects of the gene regulatory networks that can only be studied when the gene expressions are described as continuous values (Savageau, 1998). Moreover, the Boolean models heavily rely on suitable choices of threshold values to convert real-valued microarray data to binary values (i.e. active or inactive). This thresholding step poses another restriction on the use of the Boolean models as it is often difficult to determine an optimal threshold value to identify whether a particular gene is active or not. Since different genes can be active with different expression levels, the optimal choices of these threshold values can differ dramatically from one gene to another.

On the other side of the spectrum, the models attempting to incorporate detailed biochemical interactions among the genes in a pathway are often limited to discovery of gene networks with a small number of genes. This is due to the fact that these models require the estimation of a large number of parameters from a small set of microarray data (Arkin et al., 1998). Specifically, considering the fact that often very few replicates of gene expression data are available, the statistical signal processing theory (Kay, 1993) states that a model with many estimated parameters trained with these few data points may not be reliable. In other words, based on the fundamental principles of estimation theory, since the parameters estimated using only a few replicate points may not be reliable, a model trained by only a few data point can simply overfit the training data (Cadzow, 1994). Some more detailed models based on Bayesian networks theory are also known to suffer from the complications involved in estimating a large number of parameters. In other words, although Bayesian models are much less complex than differential equations based compartmental models, they too require the estimation of a large number of parameters often from a small number of data points.

The majority of gene regulatory models developed in the recent years can be considered as specialized versions of the reverse engineering approach. These methods attempt to avoid over-simplification of the problem while obtaining more accurate and realistic models. A group of such methods have been specialized to process DNA microarray data containing mRNA of the genes before and after the perturbation of the pathway. In the recent works conducted by Kholodenko et al. (2002) and de la Fuente et al. (2002), some concepts in control theory and metabolic control are utilized to quantitatively model the effects of the changes in the expression value of one gene on the expression level of other genes. These methods utilize some fundamental concepts of metabolic control analysis (MCA) (Heinrich and Rapoport, 1974; Kacser and Burns, 1973; Fell, 1996) to process the variations in expression value of the regulatee genes before and after pathway perturbations (often knock-out of regulator genes) to estimate the regulatory network. MCA simply uses control coefficients to determine the relationship among parameters describing biochemical reactions (e.g. rate constants) and the state variables of the system (e.g. fluxes and metabolite concentrations).

The network created by almost all methods described above will include some direct links between each pair of genes but fail to discover indirect interactions between the two genes. To see this disadvantage more clearly, consider the simple network shown in Fig. 1.

As can be seen in Fig. 1, the network explicitly describes the direct effect of gene C on gene B with the direct links between these two genes. However, the network of Fig. 1 also indicates that gene C has some indirect effects on gene B through genes D and A. Quantitatively describing these indirect effects constitutes one of the challenges in analysis of gene networks that will be addressed in this paper. The present paper introduces a systematic method to estimate and model the overall effects of gene C on gene B. Estimating the overall interactions among a pair of genes (both direct and indirect) plays an important role in many practical biology studies. For instance, one of the most practically important problems in designing gene knock-out experiments is to establish an optimal procedure to identify the most informative gene(s) to be deleted in the next step. Pe'er et al. (2002) addressed the gene-knock out procedure from a different point of view, i.e. they followed the principles of local modeling of regulatory interactions developed by Friedman et al. (2000) to select a set of active regulators from a pool of candidates as the genes to be knocked out in the next experiment. This method utilizes a local scoring function such as mutual information to evaluate the models of regulators. Pournara and

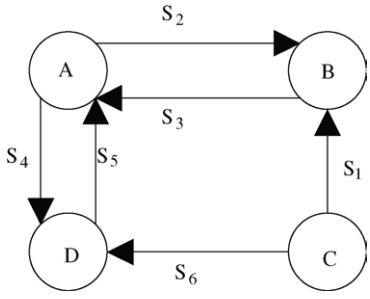

Fig. 1. A typical gene regulatory network.