Available online at www.sciencedirect.com
ScienceDirect
Computers and Chemical Engineering 31 (2007) 950–961
Computers & Chemical Engineering
www.elsevier.com/locate/compchemeng

A network model for gene regulation

Rishi R. Gupta *, Luke E.K. Achenie

Department of Chemical Engineering, University of Connecticut, Storrs, CT 06269, USA

Received 2 December 2005; received in revised form 8 April 2006; accepted 3 August 2006
Available online 26 September 2006

Abstract

Advances in microarray technology have resulted in an exponential rise in gene expression data. Partially as a result of this, full genome sequences have been reported for many organisms. In addition several methods have been developed (a) for assigning functionality to previously unknown genes and (b) for measuring the output (i.e. gene expression) of the gene regulatory network. The knowledge of the gene regulatory network further gives insights about gene pathways. This information leads to many potential applications in medicine and molecular biology, examples of which are identification of metabolic pathways, complex genetic diseases, drug discovery and toxicology analysis. Also, gene regulatory networks allow comparison of expression patterns of many uncharacterized genes; this comparison provides clues to gene function. A variety of models (such as neural networks, Boolean networks, and Bayesian) have been proposed in recent times. Although each of these models have individual strengths, none of them addresses important issues such as time delay, or make use of available biological information. In the work presented here we demonstrate that networks can efficiently model natural biological processes, specifically gene regulatory systems. Through the modeling approach, we have inferred gene regulatory networks using a time course data set for (a) lambda bacteriophage infection, (b) osteoblast study, and (c) rat central nervous system (CNS) development. The results compare favorably with experimental results from the literature.

Keywords: Microarray; Boolean algebra; Network; Time series; Gene regulation

1. Introduction

According to the central dogma of molecular biology, genes (composed of deoxyribo-nucleic acid, DNA) are the basic units of heredity (Lander, 1996). The genetic code (represented as a sequence of chemical monomers) is decoded in each living cell into the functional molecular network of proteins and RNA molecules (RNA, ribonucleic acid). The proteins and RNA are termed as the functional molecules of life. These are responsible for the physical, chemical and biological properties of the cell, which in turn is manifested globally in the behavioral nature of a living organism like humans.

1.1. Gene expression and regulation

The biochemical process by which genes are first transcribed into RNA molecules (transcription) and then converted to the protein molecules (translation) is referred to as "gene expression". This process of gene to protein information translation is not a static phenomenon. It varies dynamically with time depending on several factors such as the stage of development of the cell (or organism), environmental conditions, etc. For example the heat shock genes are over expressed (that is more proteins are synthesized) when their parent cell is suddenly exposed to a high temperature environment (Yura & Nakahigashi, 1999; Yura, Nagai, & Mori, 1993). In other words, the expression level of heat shock genes remains at a basal rate under normal or standard conditions; as soon as these conditions change the expression level increases and finally returns to the normal level of expression as the cell recovers from the shock.

When the expression level of a gene increases (such as in the heat shock example), it is said to be switched "ON" during that time interval. Similarly a gene is switched "OFF" when its expression level decreases from the normal rate (DeRisi, Iyer, & Brown, 1997; McAdams & Arkin, 1999; McAdams & Shapiro, 1995).

For a known bio-molecular system it is important to investigate the effects of perturbing various components (which define cellular mechanisms) within the system. Also, it is important to

* Corresponding author. Tel.: +1 860 486 4600/2756; fax: +1 860 486 2959.
E-mail addresses: rishi@engr.uconn.edu (R.R. Gupta), achenie@engr.uconn.edu (L.E.K. Achenie).

0098-1354/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compchemeng.2006.08.008

use experimental tools, which can measure these effects simultaneously and efficiently for as many genes as possible. This would help to measure the gene regulation for all the genes at the same condition. Microarray technology (a recent advancement in biotechnology) helps address these issues. It provides biologists with the ability to measure the expression levels of thousands of genes in a single experiment; therefore it has led to a paradigm shift in the way biologists approach their experimental model. For example the technique makes it possible to examine how the cell or tissue uses a repertoire of genes to control a biological process (Eisen, Spellman, Brown, & Botstein, 1998). It can be used to perform assays for measuring the transcriptional changes in gene functions (Chu et al., 1998). These assays can also be used to reveal genes that characterize tissue types (Bittner et al., 2000), developmental stage (Spellman et al., 1998), or response to environmental conditions (Iyer et al., 1999).

Unfortunately the success of microarrays has exposed a key economically driven weakness: for a time-course experiment, the number of time points (at which experiments are carried out) is much smaller than the amount of gene expression data at a given time point. This is due to the fact that there is a high-cost associated with a microarray measurement at a given time point. Besides cost, experimental errors, duplication and time plays a critical role in every microarray measurement. As a result computational models have been developed to partially offset this limitation; these models mimic biological knowledge (D'haeseleer, PhD. dissertation, 2000; D'haeseleer, Wen, Fuhrman, & Somogyi, 1999; D'haeseleer, Liang, & Somogyi, 2000; Marnellos & Mjolsness, 1998; Marnellos, Deblandre, Mjolsness, & Kintner, 2000; Vohradsky, 2001). Thus there is a definite need to address the dimensionality issue while developing a gene regulatory model.

A gene regulatory model (or the genetic pathway) can be modeled as a network of elements or nodes, which interact over a period of time to regulate the expression of one another (Brownstein, Trent, & Boguski, 1998; Fields, Kohara, & Lockhart, 1999; Kanehisa & Goto, 2000; Loomis & Sternberg, 1995; Nowak, 1995; Palsson, 1997; Strohman, 1997; Thieffry, 1999). In this work, we employ enhanced Boolean algebra to model regulatory effects on the expression of a gene. A Boolean algebra in the strict sense means either 0 or 1. We introduced "−1" to address the inhibition or negative interaction among various elements in the network. We refer to the resulting network as the enhanced Boolean network. The paper is organized as follows. In Sections 1.2–1.4 we provide a brief description of Pearson's Correlation and enhanced Boolean networks. Sections 2 and 3, describes the model formulation, case studies results and discussions, respectively. Finally we provide conclusions in Section 4.

1.2. Pearson's correlation

The correlation between two variables reflects the degree to which the variables are related. The most common measure of correlation is the Pearson correlation. Pearson's correlation reflects the degree of linear relationship between two variables:

$$r = \frac{\sum XY - \sum X \sum Y/N}{\sqrt{\left(\sum X^2 - \left(\sum X\right)^2/N\right)}\sqrt{\left(\sum Y^2 - \left(\sum Y\right)^2/N\right)}}$$

where, r represents the Pearson's correlation between two random variables "X" and "Y".

In the equation described above, the numerator represents the covariance of X and Y, whereas, the denominator is the standard deviations for X and Y. Thus, it is necessary, for the correlation to be defined, to have both the standard deviations finite and nonzero. Also, it is assumed that both "X" and "Y" are approximately normally distributed, and their joint distribution is bivariate normal. The correlation is +1 in the case of an increasing linear relationship, −1 in the case of a decreasing linear relationship, and some value in between in all other cases, indicating the degree of linear dependence between the variables. The closer the coefficient is to either −1 or 1, the stronger the correlation between the variables.

In this paper, we have employed Pearson's correlation to determine the linear relationship among various super genes (or clusters). This aids in establishing the network among the set of super genes and proves to be a useful statistical tool.

1.3. Boolean networks

Boolean networks allow large regulatory networks to be analyzed efficiently, by making strong simplifying assumptions on the structure and dynamics of a genetic regulatory system (Kaufman, 1969b). In the traditional Boolean network formalism (Fig. 1), a gene has 2 expression states, namely (a) on and (b) off. Also, transitions between the activation states of the genes are assumed to occur synchronously. However, it is usually the case that transitions are asynchronous. In addition, the Boolean model does not account for the fact that the extent of a gene's regulatory effect (on the transcription of other genes) may be a function of the gene expression level. The model proposed in this paper attempts to address these issues.

1.4. Network model

A network model is a parallel computational model comprised of a large number of adaptive processing units (such as neurons in a neural network model). The neurons communicate through a large set of interconnections with variable strengths (weights) in which the learned information is stored. A very important feature of these networks is their adaptive nature, in the sense that learning by example replaces conventional programming. This feature makes such computational models very appealing in application domains where one has little or incomplete understanding of the problem to be solved, but where training data are readily available. In our model we employ an enhanced Boolean network, which reduces the need (as shown by the model proposed in this work) for a large amount of data for network training as explained in the next section.

One important application of the network model is for gene identification. The latter is tackled by two complimentary