Gene Regulatory Network modelling: a state-space approach

Fang-Xiang Wu

Department of Mechanical Engineering,
Division of Biomedical Engineering,
University of Saskatchewan,
57 Campus Dr. Saskatoon, SK, S7N 5A9, Canada
E-mail: faw341@mail.usask.ca

Abstract: This study proposes a state-space model with control portion for inferring Gene Regulatory Networks (GRNs). The proposed model views genes as the observation variables, whose expression values depend on the current internal state variables and control variables, and views the means of clusters of gene expression as the control variables of the internal state equation. Bayesian Information Criterion (BIC) and Probabilistic Principal Component Analysis (PPCA) are used to estimate the internal states from observation data. The proposed approach is applied to two gene expression datasets. Computational results show that inferred GRNs possesses the characteristics of the real-life GRNs.

Keywords: Gene Regulatory Network; GRN; time-course gene expression data; state-space approach; Bayesian Information Criterion; BIC; Probabilistic Principal Component Analysis; PPCA; stability; robustness; periodicity; observability; controllability; data mining; bioinformatics.

Reference to this paper should be made as follows: Wu, F-X. (2008) 'Gene Regulatory Network modelling: a state-space approach', *Int. J. Data Mining and Bioinformatics*, Vol. 2, No. 1, pp.1–14.

Biographical notes: F-X. Wu received his BSc and MSc Degrees in Applied Mathematics from the DaLian University of Technology China; in 1990 and 1993, respectively, his first PhD in Control Theory and its Applications from Northwestern Polytechnical University in 1998, China; and his second PhD in Biomedical Engineering from the University of Saskatchewan, Canada, in 2004. He is currently an Assistant Professor at the Department of Mechanical Engineering, and an associate faculty member of Division of Biomedical Engineering, University of Saskatchewan, Canada. His current research interests include computational bioengineering, systems biology, gene regulatory networks, applications of control theory to biological systems and mass spectrometers.

1 Introduction

In a biological developmental process, a large number of genes and proteins in cells either directly or indirectly interact with one another. Such interactions make up a dynamic GRN. A GRN acts as a complex dynamic system for controlling cellular functions. For a normal cell life cycle to take place, a cell needs to have a correctly

working GRN for control in place. Many of the known regulatory factors that control mRNA levels work at the level of transcription (other control mechanisms are not considered here). Most of these regulatory factors are components of protein complexes that regulate the transcription of other genes. Insights into the nature and function of various networks are of interest to many researchers, as it has proved that many diseases (such as cancer and AIDS) stem from the malfunction of GRNs of the corresponding cell lines.

A GRN could have many components (genes and proteins), and the mechanism of gene regulation is unclear. Therefore, in order to study a GRN, it is necessary to have the observation data of gene expression and protein expression from the corresponding cell. With advances in the measurement technology for gene expression and in genome sequencing, it has become possible to measure the expression level of thousands of genes simultaneously in a cell at a series of time points over a specific biological process. Such time-course gene expression data provides insights into dynamic GRNs although there is no counterpart technology to measure time-course protein expression level. Actually, with time-course gene expression data, several modelling methods have been proposed for inferring GRN such as Boolean network model (Akutsu et al., 1999; Liang et al., 1998; Somogyi and Sniegoski, 1996), differential/difference equation models (Chen et al., 1999; D'haeseleer et al., 1999).

Boolean network model (Somogyi and Sniegoski, 1996) simply views a gene's expression state to be either completely 'on' or 'off'. These states are often represented by the binary values 1 and 0, respectively, and the state of a gene is determined by a Boolean function of the states of all possible genes in the network. The functions can be represented in tables, or as rules. For example, if gene A is 'on' AND either gene B OR C is 'off' at time t, then gene D is 'on' at time $t + \Delta t$. As the system evolves from one state (or time point) to the next, the states of all genes in the network are used as input to rules which specify which genes will be 'on' at the next state or time point. Somogyi and Sniegoski (1996) have showed that such Boolean networks have features similar to those in biological systems, such as global complex behaviour, self-organisation, stability, redundancy, and periodicity. Liang et al. (1998) have described an algorithm for inferring genetic network architectures from the rule table of a Boolean network model. Their computational experiments have showed that a small number of state transition pairs are sufficient to infer the original observations. Recently, Akutsu et al. (1999) have devised a much simpler algorithm for the same problem and proved that if the in-degree of each node (i.e., the number of input nodes to each node) is bounded by a constant h, only $O(\log n)$ state transition pairs (from possible 2^n pairs) are necessary and sufficient to identify the original Boolean network of n nodes (genes) correctly with high probability. However, the Boolean network models depend on simplifying assumptions about biology systems. For example, by treating gene expression as either completely 'on' or 'off', these models ignore those genes that have a range of expression levels and can have regulatory effects at intermediate expression levels. Therefore they ignore those regulatory genes that influence the transcription of other genes to variable degrees.

In addition to Boolean networks models, differential/difference equation models have also been applied to inferring gene expression. Chen et al. (1999) have proposed a differential equation model of gene expression. Due to the lack of gene expression data, the model is usually underdetermined. Using the additional requirements that the GRN should be sparse, they have showed that the model can be constructed in $O(n^{h+1})$ time, where n is the number of genes and/or proteins in the model and h is the number of

maximum nonzero coefficients (connectivity degree of genes in a regulatory network) allowed for each differential equation in the model. In order that the parameters of the models are identifiable, both Chen et al. (1999) and Akutsu et al. (1999) assume that all genes have a fixed maximum connectivity degree h(often small). These assumptions are debatable. In biological reality, some genes are known to have many regulatory inputs, while others are not known to have more than a few. D'haeseleer et al. (1999) have proposed a linear difference model for mRNA expression levels during Central Nervous System (CNS) development and injury. To deal with the lack of gene expression data, a non-linear interpolation scheme is adopted to guess the shapes of gene expression profiles between the measured time points. Such an interpolation scheme is ad hoc. Therefore, the reasonableness of the model built from such interpolated data is suspicious. In addition, there exists a problem of dimensional disaster when the number of genes in a model is large.

This paper describes a state-space approach to modelling GRNs. The state-space approach is one of the most powerful methods to modelling a dynamic system and has been widely employed for engineering control systems (Chen, 1999). A state-space model consists of internal variables, external input (control) variables, and output (observation) variables. In a state-space model, the observation variables typically depend on the internal variables, while the change in the internal variables is completely determined by the current internal variables plus any external inputs. In the existing models such as Boolean network, differential and difference models, genes are viewed as the internal state variables as well as observation variables of a DGRN, and their expression levels are the values of both the internal state variables and the observation variables. This viewpoint has suffered from the underestimation of the model parameters as pointed out previously. Actually, not all genes (their products, proteins) directly regulate gene expressions in a network since only a part of genes are translated into regulatory factors (proteins) which regulate gene expression while others are translated into structure proteins which do not participate gene regulation (Alberts et al., 1998; Baldi and Hatfield, 2002; Liebler, 2002). Recently we have propose a state-space model for GRNs (Wu et al., 2004a), in which genes are viewed as the observation variables and gene expression dynamics is governed by a group of the internal variables. Further we have extended this model to the one with time-delayed regulatory relationships (Wu et al., 2004b, 2005). However, the previous model (Wu et al., 2004a) does not take the control portion of the system into consideration, and instead just describes the influence of internal states on gene expression level and assumes that the internal states evolve autonomously. Actually, expression levels of all genes affect the internal states in turn.

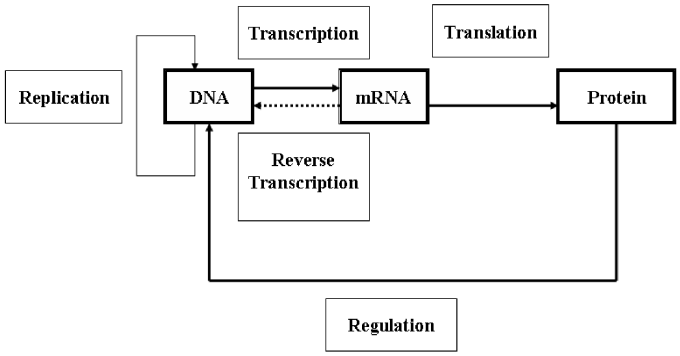
This study extends our earlier work (Wu et al., 2004a) by proposing a state-space model with the control portion for inferring GRNs. In the proposed model, genes are viewed as the observation variables, whose expression values depend on the current internal state variables and control variables. BIC and PPCA are used to estimate the internal state variables from its dynamic observation data – time-course gene expression data. To reflect the influence of gene expression on the internal variables, the means of clusters of gene expression are viewed as control input variables, which are obtained by the k-means cluster analysis method. The proposed approach is applied to two time-course gene expression datasets. The computational results show that two inferred GRNs possess the characteristics of the real-life GRNs, such as stability, robustness, and periodicity, observability and controllability.

2 Models and estimation

2.1 Biological model

In living systems, the double stranded macromolecules deoxyribonucleic acids (DNAs) encode the genetic information. The genetic information of an organism is distributed onto both strands of a DNA molecule. The term *gene* refers to those segments of one strand of DNA on which genetic information can be copied onto another class of macromolecules messenger ribonucleic acids (mRNAs) in a process called *transcription*. The mRNA is then used as a template to build proteins from 20 different amino acids in a process called *translation*. In turn, some of proteins regulate the transcription process. This whole scenario shown in Figure 1 is called the central dogma in molecular biology, and is the biological model of GRN in this study. A gene is said to be expressed if its genetic information is copied onto an mRNA in the transcriptional process.

Figure 1 The dogma of molecular biology



In any biological process, not all genes in a cell are expressed simultaneously, and instead only those genes pertinent to the biological process are expressed. Depending on the type of cells and the stage of cellular developmental process, the expression rates and abundances of genes vary widely. It has long been recognised that mRNA plays a pivotal role in determining the type and quantity of proteins produced by cells (Baldi and Hatfield, 2002). Indeed, the differences in protein content of different cells are the reflection of differences in the mRNA species expressed and of their levels of expression (abundance) during cellular development and maintenance. Once such differences in mRNA populations among types of tissues/cells are appreciated, it becomes important to quantify these differences and use them to understand GRNs.

2.2 Mathematic model

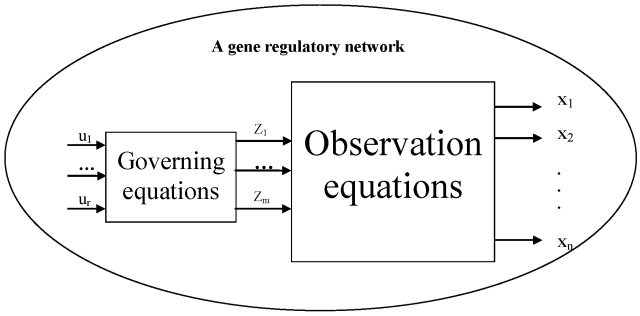
The following state-space model is proposed to describe a GRN

$$\begin{cases} \mathbf{z}(t+1) = \mathbf{A} \cdot \mathbf{z}(t) + \mathbf{B}\mathbf{u}(t) + \mathbf{n}_1(t) \\ \mathbf{x}(t) = \mathbf{C} \cdot \mathbf{z}(t) + \mathbf{n}_2(t). \end{cases} \tag{1}$$

The meaning of the variables follows as: in terms of linear system theory (Chen, 1999), equation (1) is called the state-space model of a dynamic system which can be

diagrammatically expressed by Figure 2. The vector $\mathbf{x}(t) = [x_1(t) \ldots x_n(t)]^T$ consists of the observation variables of the system and $x_i(t)$ $(i = 1, \ldots, n)$ represents the expression level of gene i at time point t, where n is the number of genes in the network. The vector $\mathbf{z}(t) = [z_1(t) \ldots z_p(t)]^T$ consists of the internal state variables of the system and $z_i(t)$ $(i = 1, \ldots, p)$ represents the expression value of internal element (variable) i at time point t which directly regulates gene expression, where p is the number of the internal state variables. The vector $\mathbf{u}(t) = [u_1(t) \ldots u_r(t)]^T$ represents the external input (control variable) of the internal state governing equation. The matrix $\mathbf{A} = [a_{ij}]_{p \times p}$ is the time translation matrix of the internal state variables or the state transition matrix. It provides key information on the influences of the internal variables on each other. The matrix $\mathbf{B} = [b_{ik}]_{p \times r}$ is the control matrix. The entries of the matrix reflect the strength of a control variable to an internal variable. The matrix $\mathbf{C} = [c_{ik}]_{n \times p}$ is the observation matrix which transfers the information from the internal state variables to the observation variables. The entries of the matrix encode information on the influences of the internal regulatory elements on the genes. Finally, the vectors $\mathbf{n}_1(t)$ and $\mathbf{n}_2(t)$ stand for system noise and observation noise. In model (equation (1)) the upper equation is called the internal state governing equation while the lower one is called the observation equation.

Figure 2 A state-space model for Gene Regulatory Networks, where $x_i(i = 1, \ldots, n)$ are the observation variables, $z_i(i = 1, \ldots, p)$ are the state variables, and $u_i(i = 1, \ldots, r)$ are the control variables



2.3 Internal variables and estimation of observation matrix

Let **X** be the gene expression data matrix with n rows and m columns, where n and m are the numbers of the genes and the measuring time points, respectively. The constructing of model (equation (1)) using microarray gene expression data **X** may be divided into three phases. Phase one identifies the internal state variables and their expression matrix, and estimates the elements of observation matrix **C**; Phase two defines the control internal variables; and Phase three estimates the elements of matrices **A** and **B**.

The internal states are latent variables in GRNs. They could be any unobserved molecules in cell which participate the process of gene regulation. Form the biological model in Figure 1, it is reasonable to assume that the latent variables are some regulatory factors (protein). Many statistical methods (Everitt and Dunn, 1992) have been developed to find the expression of latent variables from the observation data. In this study, the maximum-likelihood algorithm for PPCA (Tipping and Bishop, 1999) is employed to extract the internal variables from the observation data (time-course gene expression data). The PPCA model can be expressed by

$$\mathbf{X} = \mathbf{C} \cdot \mathbf{Z} + N \tag{2}$$

where $\mathbf{X}$ is the $n \times m$ observation data matrix, each row of which is an observation sample; $\mathbf{C}$ is the $n \times p$ transformation matrix, each row of which is a realisation of latent variables; and $\mathbf{Z}$ is the $p \times m$ loaded matrix, each row of which represents the expression profile of an internal state, and N is the $n \times m$ noise matrix consisting by n observation noise vector $\mathbf{n}_2(t)$ in model (equation (1)) at a series of m observation time points. Assume that the sample mean is shifted to zero. The log-likelihood of PPCA model (Tipping and Bishop, 1999) is expressed by

$$L = -\frac{n}{2}\left\{m(\ln 2\pi) + \log\left|\mathbf{D}\right| + tr(\mathbf{D}^{-1}\mathbf{S})\right\} \tag{3}$$

where $\mathbf{D} = \mathbf{Z}^T\mathbf{Z} + \sigma^2\mathbf{I}$ and $\mathbf{S} = \mathbf{X}' \times \mathbf{X}/n$. For the given number of internal variables, p, the global maximum log-likelihood of the PPCA model is calculated by

$$L_p = -\frac{n}{2}\left\{\sum_{j=1}^{p}\log(\lambda_j) + (m-p) \times \log\left(\sum_{j=p+1}^{m}\lambda_j/(m-p)\right) + m(\log(2\pi)+1)\right\} \tag{4}$$

when

$$\mathbf{Z}_p = \mathbf{R}(\mathbf{Q}_p - \sigma^2\mathbf{I}_p)^{1/2}\mathbf{U}_p^T \tag{5}$$

where λ_j ($j = 1, \ldots, p$) are the first p largest eigenvalues of the sample variance matrix $\mathbf{S}$, the matrix $\mathbf{Q}_p$ is a $p \times p$ diagonal matrix, whose diagonal elements are these λ_j ($j = 1, \ldots, p$), $\mathbf{U}_p$ is a $m \times p$ matrix, each column of which is a corresponding eigenvector of $\mathbf{S}$, $\mathbf{I}_k$ is a $p \times p$ identity matrix, $\mathbf{R}$ is an arbitrary $p \times p$ orthogonal matrix, and

$$\sigma^2 = \sum_{j=k+1}^{m}\lambda_j/(m-p).$$

From equation (4), the values of the maximum log-likelihood for the PPCA model increase with the increased numbers of internal state variables, p. The redundant internal state variables may result in a complicated model. Since the PPCA has a solid probabilistic foundation, BIC is employed to determine the number of internal state variables (Burnham and Anderson, 1998; Schwarz, 1978). For each model, the BIC is defined as:

$$\mathrm{BIC}(p) = 2 \cdot L_p - \log(n) \cdot v_p \tag{6}$$

where n is the sample size (the number of genes), and $v_p(=mp + 1)$ is the number of parameters in the PPCA model. Since the terms $nm(\log(2\pi) + 1)/2$ and $\log(n)$ in equation (7) is a constant for a given dataset, the calculation of BIC can be simplified as

$$\mathrm{BIC}(p) = -n\left\{\sum_{j=1}^{p}\log(\lambda_j) + (m-p) \times \log\left(\sum_{j=k+1}^{m}\lambda_j/(m-p)\right)\right\} - \log(n) \cdot mp. \tag{7}$$

By this definition, the model with the largest BIC is chosen.

Note that if $\{\mathbf{C}, \mathbf{Z}\}$ is an optimum solution of equation (2), $\{\mathbf{C}\mathbf{T}^{-1}, \mathbf{T}\mathbf{Z}\}$ is also its optimum solution, where $\mathbf{T}$ is any $p \times p$ non-singular matrix. However, it can be proved that the state-space models from $\{\mathbf{C}, \mathbf{Z}\}$ and $\{\mathbf{C}\mathbf{T}^{-1}, \mathbf{T}\mathbf{Z}\}$are algebraically equivalent (Chen, 1999). Therefore, one can always normalise the expression profiles of the internal state variables. For the optimal number of internal state variables, p, since $\mathbf{R}(\mathbf{Q}_p - \sigma^2\mathbf{I}p)^{1/2}$ is a $p \times p$ non-singular matrix, we take

$$\mathbf{Z} = \mathbf{U}_p^{\mathrm{T}} \tag{8}$$

as the expression profiles of the internal state variables. Further, the corresponding transformation matrix $\mathbf{C}$ can be calculated by formulae $\mathbf{C} = \mathbf{X} \cdot \mathbf{Z}^{+}$.

2.4 Control variables

In state-space model (equation (1)), the control variables together with current internal states determine the next internal states. From the viewpoint of biology, the overall expression level of all genes in a cell determines the overall protein expression level (Alberts et al., 1998; Liebler, 2002; Tozeren and Byers, 2004). By cluster analysis, gene expression patterns can be found from expression profiles of all genes involved in a process (Eisen et al., 1998). Genes that have the same expression patterns have the same or similar function. Similarly, often proteins work in a team (not independently) for a specific function as regulating expression of a certain class of genes (Alberts et al., 1998; Liebler, 2002). Therefore, it is reasonable to assume that the control variables are the means of gene clusters which reflect the overall gene expression and that the number of control variables is the same as that of the internal variables.

There are many cluster methods which can be used here to find the means of gene clusters (Duda et al., 2001; Eisen et al., 1998; Tavazoie et al., 1999; Yeung et al., 2001), but none of these is perfect. For simplicity, the *k*-means clustering method is employed to find the control variable expression (the means of gene clusters) in this study. To avoid falling the local optimal clustering, the *k*-means cluster algorithms will be run a number of times with different initial partitions. The resultant means of the best one among all runs are adopted to be the expression of control variables.

2.5 Estimation of state transition matrix and control matrix

Using the expression data of the internal variables and the control variables, one can estimate the parameters of the state transition matrix and control matrix in the internal state governing equation:

$$\mathbf{z}(t+1) = \mathbf{A} \cdot \mathbf{z}(t) + \mathbf{B}\mathbf{u}(t) + \mathbf{n}_1(t) \tag{9}$$

by minimising the system noise $\mathbf{n}_1(t)$. This is equivalent to minimise the cost function

$$CF = \sum_{j=1}^{m} \|\mathbf{z}(t_j) - \mathbf{v}(t_j)\|^2 \tag{10}$$

where the time-variant vector $\mathbf{v}(t)$ has the same dimensions as the internal state vector $\mathbf{z}(t + 1)$ and is calculated by the following difference equation

$$\mathbf{v}(t+1) = \mathbf{A} \cdot \mathbf{v}(t) + \mathbf{B}\mathbf{u}(t) \tag{11}$$

with the initial state value $\mathbf{v}(0) = \mathbf{z}(t_0)$, and control values $u(0)$, …, $u(t)$.

For equally spaced measurements, the minimisation of the cost function (equation (10)) can be solved by the least square method for the linear regression problem (Harvey, 1993). For unequally spaced measurements, the problem becomes non-linear, and it is necessary to determine matrices **A** and **B** by using an optimisation technique such as those in Chapter 10 of Press's text (Press et al., 1992). In this study assume that each row of matrix B has at most one non-zero element. The matrix **A** contains p^2 unknown elements and the matrix **B** contains p unknowns while the matrix **Z** contains $m \cdot p$ known expression data points. If $p > m$, equation (9) will be underdetermined. Fortunately, using BIC the number of chosen internal variables p generally is less than the number of time points m. Therefore matrices **A** and **B** in model (equation (1)) could unambiguously be estimated from time-course gene expression data.

3 Model evaluation

In this study, the inferred GRNs will be evaluated in the following aspects: the prediction power, stability, robustness, periodicity, controllability, and observability.

- *The prediction power (error)*. Let $\hat{\mathbf{X}}$ be a data matrix with the same size as the original data matrix **X**, which is computed from the model inferred from the data matrix **X**. The prediction error reflects how well $\hat{\mathbf{X}}$ approximates **X**. The prediction error (P_E) is defined as:

$$P_E = \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{X}(i,:) - \hat{\mathbf{X}}(i,:)\|^2 / \|\mathbf{X}(i,:)\|^2 \tag{12}$$

  where $\mathbf{X}(i, :)$ is the expression profile of gene i in the data matrix **X**. $\|\mathbf{X}(i, :)\|$ is the Euclidean norm of the vector $\mathbf{X}(i, :)$. Intuitively, the smaller the prediction error, the stronger the prediction power is. The prediction error P_E defined in equation (12) is invariant with respect to the scale of **X**. Therefore, it is more reasonable to evaluate the models using formulae (12) than the one defined by Wessels et al. (2001). We will use the prediction error in equation (11) to evaluate the models.

- *Stability*. Due to the limited energy and storage within a cell, concentrations of gene expression products such as mRNA should remain bounded. All real-life gene networks are therefore stable. Consequently, the inferred gene network models should also be stable in order to be realistic. For our model, this is equivalent to the governing equation (9) being stable. It has been proven (Chen, 1999) that the equation (9) is stable if and only if all eigenvalues of the state transition matrix **A** lie inside the unit circle in the complex plane.

- *Periodicity*. Certain biological processes are periodic. The cell-cycle and circadian clock, for example, repeat at well-defined and reliable time intervals. Studies have shown that GRNs associated with these periodic biological processes are themselves rhythmic (Langmead et al., 2002; Kauffman, 1993; Wichert et al., 2004). Therefore, the inferred GRNs associated with these periodic biological processes should be periodic at its stable states. Accordingly, the periodicity of system (equation (1)) at its stable state is determined by its dominant eigenvalues of the state transition matrix **A** whose moduli are the largest.

- *Robustness.* The robustness of a GRN is understood as its insensitivity to noise or disturbance. It is obvious that a real-life GRN has robustness (Tavazoie et al., 1999; Wessels et al., 2001). Therefore, the inferred GRN should be robust. The stability of a linear system implies robustness to a certain degree (Chen, 1999). Note that the stability, robustness, and periodicity of system (equation (1)) are all related to the eigenvalues of the state transition matrix **A**.
- *Controllability.* A dynamic control system is said to be controllable if any state could be transferred to any preset state by appropriate control actions (Chen, 1999). For example, if a GRN is controllable, it can always be transferred to a normal state if the network malfunctions or deviates from the normal state. The inferred network should be controllable if a real-life GRN is. It has been proven that the linear system (equation (1)) is controllable (Chen, 1999) if and only if

  $$\text{rank}([B, AB, \ldots, A^p B]) \geq p. \tag{13}$$

  A control system is said to be directly controllable if $\text{rank}(B) \geq p$. A directly controllable system can more easily transfer one state to the other one than a controllable system can.
- *Observability.* A dynamic control system is said to be observable (Chen, 1999) if the internal state could be estimated by the observation data of the systems. For an observable GRN, one can always estimated its internal states from the observation data (i.e., time-course gene expression data) even though one can not directly 'see' the behaviour of the internal variables. Its inferred network should be observable if a real GRN is. Because of the proposed modelling method in this study, all inferred GRNs are observable.

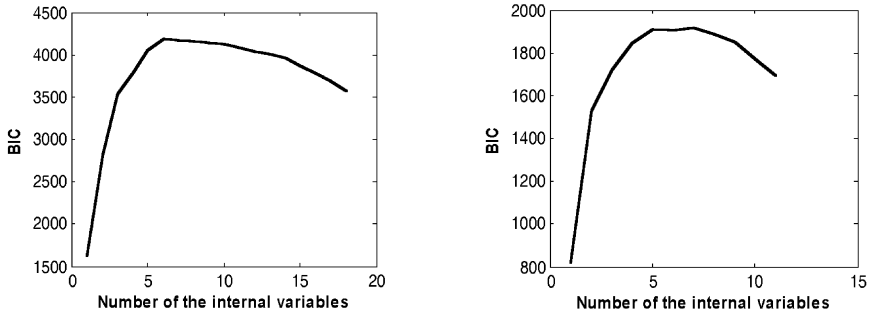
4 Computational experiments and results

In this section, the proposed methodology was applied to two publicly available microarray datasets. The first dataset (ALPHA) is from Spellman et al. (1998) and consists of the expression data of 701 cell-cycle related genes for the 18 equally spaced time points with no missing data. The dataset is available at http://cellcycle-www.stanford.edu. The second dataset (BAC) is from Laub et al. (2000) and consists of the expression data of 1501 genes for 11 equally spaced time points. This dataset created from the original dataset available at http://caulobacter.stanford.edu /CellCycle, by excluding genes whose profiles have missing data and are of little variation. As the mean values and magnitudes for genes and microarrays mainly reflect the experimental procedure (Eisen et al., 1998) the expression profile of each gene is normalised to have the median of zero and the length of one, and then for the expression values on each microarray as so to have the mean of zero and the length of one. Such normalisations also make the PPCA simple (Tipping and Bishop, 1999).

The maximum likelihood algorithm for PPCA (Tipping and Bishop, 1999) is employed to analyse the two datasets. For a variety of number k of internal state variables, BIC(k) is calculated by equation (6). Figures 3(a) and (b) depict the profiles of BIC with respect to the number of internal variables for Dataset ALPHA and Dataset BAC, respectively. Obviously, the number of internal variables is six for ALPHA from

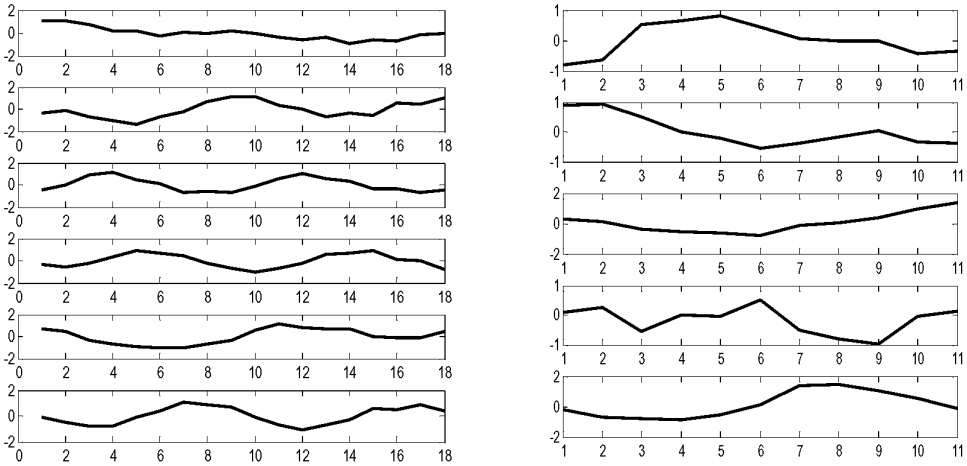
Figure 3(a) while it is five for BAC from Figure 3(b). The expression profiles of the internal state variables, **Z**'s, are calculated according to equation (8), and further so are the corresponding transformation matrices, **C**'s. The *k*-means cluster algorithm in MatLab is run 1000 times for each dataset. The resultant means of the best one over 1000 runs are considered as the profiles of control variables. Figure 4(a) and (b) plot the profiles of control variables for Datasets ALPHA and BAC, respectively. Figure 4 shows that the profiles of all control variables except for the first one are strong periodic for Dataset ALPHA while the profiles of all control variables are strong periodic for Dataset BAC. This result is not surprising because both datasets are collected from the cell-cycle division process which is periodic.

Figure 3 Plots of BIC with respect to the number of internal variables for datasets: (a) ALPHA and (b) BAC



(a) (b)

Figure 4 Profiles of control internal variables for datasets: (a) ALPHA and (b) BAC



(a) (b)

In order to estimate parameters in the state transition matrix **A** and control matrix **B** in model (equation (1)), the linear regression method is employed as gene expression profiles in each of both datasets are equally spaced. The estimates of matrices **A** and **B** follow as: for dataset ALPHA

$$\mathbf{A} = \begin{bmatrix} 0.7506 & 0.5163 & 0.0837 & 0.1047 & 0.0809 & 0.0910 \\ -0.7591 & 0.4087 & -0.2794 & 0.0588 & 0.1763 & 0.1705 \\ 0.8161 & 0.4035 & 0.7096 & 0.4321 & 0.2824 & 0.1147 \\ 1.5038 & -0.8739 & 0.1333 & -0.1210 & -0.2877 & -0.3445 \\ -2.4316 & -0.1057 & -0.3163 & -0.1836 & -0.0597 & 0.0110 \\ 1.4272 & -4.5559 & 0.9813 & 0.3010 & -0.1987 & -0.5031 \end{bmatrix} \tag{14}$$

and

$$\mathbf{B} = \begin{bmatrix} 0.1371 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.1472 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.3169 \\ 0 & 0.6408 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.9452 & 0 & 0 & 0 \\ 0 & 0 & 0 & -2.0348 & 0 & 0 \end{bmatrix} \tag{15}$$

and for Dataset BAC:

$$\mathbf{A} = \begin{bmatrix} 0.5676 & 0.4080 & -0.2257 & -0.0929 & -0.0589 \\ -0.2210 & 1.0866 & -0.2986 & 0.0071 & -0.0217 \\ -0.2352 & -0.1825 & 0.1203 & -0.9955 & 0.5100 \\ -1.5691 & 1.7586 & 0.7706 & -0.7890 & 0.4104 \\ 0.3304 & -0.6586 & 0.2063 & -0.0913 & 0.1812 \end{bmatrix} \tag{16}$$

and

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 0 & -0.0945 \\ 0 & -0.3388 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.3386 & 0 \\ 0 & 0 & -1.6543 & 0 & 0 \\ -0.4799 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{17}$$

The prediction power for the inferred GRNs is checked by using formulae (12). The values of the prediction error (P_E) are 5.0161×10^{-4} and 3.1550×10^{-4} for the inferred networks from ALPHA and BAC, respectively. This indicates that the proposed method performs very well for inferring GRNs. If the rank of control matrix is greater than or equal to the number of internal states, the system (equation (1)) is directly controllable, and otherwise one must use equation (13) to check the controllability of the system (equation (1)). From the expressions (15) and (17), the inferred GRNs from both datasets are directly controllable. This means that the states of GRNs can be modulated to any expected states by using appropriate input control strategy. Especially, by adding more mRNA transcripts of some genes the states of the system can be modulated because by this way the expression values of related genes can be changed (Gardner et al., 2003; Spieth et al., 2004) and thus the control variables (i.e., the cluster means of gene expression) can be changed.

To inspect stability, robustness, and periodicity of these two inferred GRNs for genes in these two datasets, the eigenvalues of the state transition matrices **A** in equations (14) and (16) are calculated, respectively. For the inferred network from ALPHA, matrix **A** in equation (14) has six eigenvalues: $-0.1111 \pm 0.9684i$, 0.9193, 0.3853, 0.1693 and –0.0666. For the inferred network from BAC, matrix **A** in equation (16) has five eigenvalues: $0.9621 \pm 0.1894i$, $-0.4606 \pm 0.7469i$ and –0.1636. For each inferred network, all of the eigenvalues of its state transition matrix lie inside the unit circle in the complex plane. This means that the inferred regulatory networks from both datasets are stable and robust. Furthermore, the dominant eigenvalues of the inferred networks from ALPHA and BAC are pairs of conjugate complex number: $-0.1111 \pm 0.9684i$ and $0.9621 \pm 0.1894i$, respectively. Accordingly, this implies that the network behaves periodically (Durbin and Koopman, 2001). This result is not surprising again because the networks are inferred from cell-cycle regulated gene expression data.

5 Conclusion and discussion

This paper has proposed a state-space approach to modelling GRNs. The model for inferring GRNs consists of two parts: internal states governing equation and observation equation. The novelty of this study is that the control variables are incorporated into the internal states governing equation which governs the dynamics of the internal variables. The proposed approach is applied to two previously published gene expression datasets to investigate the characteristics of the inferred GRNs. The results have showed that the inferred networks can grasp the characteristics of real-life GRNs and are consistent with biological knowledge. For example, genes are regulated by some regulatory internal variables (Baldi and Hatfield, 2002; Liebler, 2002; Tozeren and Byers, 2004), and the inferred GRNs are stable, robust, controllable, observable and of periodic behaviour (Wessels et al., 2001; Kauffman, 1993; Kitano, 2002).

Compared to previous models such as Boolean network model (Akutsu, 1999; Liang et al., 1998; Somogyi and Sniegoski, 1996), and difference/differential equation (Chen et al., 1999; D'haeseleer et al., 1999), the proposed model (equation (1)) has the following characteristics. Firstly, gene expression profiles are the observation variables rather than the internal state variables. Secondly, from a biological angle, the model (equation (1)) can capture the fact that genes may be regulated by internal regulatory factor (Baldi and Hatfield, 2002). Thirdly, the model (equation (1)) takes the control portion of state-space model into consideration. However, the proposed approach does have some shortcomings, for example, the inherent linearity which can only capture the primary linear components of a biological system which may be non-linear, and the ignorance to time delays in a biological system resulting, for example, from the time necessary for transcription, translation, and diffusion. These shortcomings will be address in the future work.

In addition, one important exercise of future work is to study the biological relevance of the internal variables. According to the principle of gene regulation process, the internal variables should reflect the information of the regulatory factors (proteins) (Alberts et al., 1998; Liebler, 2002; Tozeren and Byers, 2004) which involve the regulation process of genes in the network. Fortunately, with advances in proteomics (Liebler, 2002) it is possible to employ the expression data of proteins involving a gene regulatory process to investigate the biological meaning of the internal variables

in the model. This goal requires close collaboration with molecular biologists as we are under way.

Acknowledgements

This study is supported by Natural Sciences and Engineering Research Council of Canada (NSERC). I would like to thank Dr. W.J. Zhang, Dr. A.J. Kusalik and Dr. H. Wang for their helpful suggestions.

References

Akutsu, T., Miyano, S. and Kuhara, S. (1999) 'Identification of gene networks from a small number of gene expression patterns under the Boolean network model', *Pacific Symposium on Biocomputing*, Vol. 4, pp.17–28.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Bray, D., Hopkin, K., Roberts, K. and Walter, P. (1998) *Essential Cell Biology*, Garland Science, New York.

Baldi, P. and Hatfield, G.W. (2002) *DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling*, Cambridge University Press, New York.

Burnham, K.P. and Anderson, D.R. (1998) *Model Selection and Inference: A Practical Information-theoretic Approach*, Springer, New York.

Chen, C.T. (1999) *Linear System Theory and Design*, 3rd ed., Oxford University Press, New York.

Chen, T., He, H.L. and Church, G.M. (1999) 'Modeling gene expression with differential equations', *Pacific Symposium on Biocomputing*, Vol. 4, pp.29–40.

D'haeseleer, P., Wen, X., Fuhrman, S. and Somogyi, R. (1999) 'Linear modeling of mRNA expression levels during CNS development and injury', *Pacific Symposium on Biocomputing*, Vol. 4, pp.41–52.

Duda, R.O., Hart, P.E. and Stork, D.G. (2001) *Pattern Classification*, Wiley Press, New York.

Durbin, J. and Koopman, S.J. (2001) *Time-series Analysis by State Space Model*, Oxford University Press, New York.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) 'Cluster analysis and display of genome-wide expression patterns', *Proc. Natl. Acad. Sci. USA*, Vol. 95, pp.14863–14868.

Everitt, B.S. and Dunn, G. (1992) *Applied Multivariate Data Analysis*, Oxford University Press, New York.

Gardner T.S., di Bernardo, D., Lorenz, D. and Collins, J.J. (2003) 'Inferring genetic networks and identifying compound mode of action via expression profiling', *Science*, Vol. 301, pp.102–105.

Harvey, A.C. (1993) *Time Series Models*, 2nd ed., The MIT Press, Cambridge, MA.

Kauffman, S.A. (1993) *The Origins of Order: Self-Organization and Selection in Evolution*, Oxford University Press, Oxford.

Kitano, H. (2002) 'Computational systems biology', *Nature*, Vol. 420, pp.206–210.

Langmead, C.J., Yan, A.K., McClung, C.R. and Donald, B.R. (2002) 'Phase-independent rhythmic analysis of genome-wide expression patterns', *Proceedings of the Sixth Annual International Conference on Research in Computational Molecular Biology*, Washington DC, USA, Vol. 1, pp.205–215.

Laub, M.T., McAdams, H.H., Feldblyum, T., Fraser, C.M. and Shapiro, L. (2000) 'Global analysis of the genetic network controlling a bacterial cell cycle', *Science*, Vol. 290, pp.2144–2148.

Liang, S., Fuhrman, S. and Somogyi, R. (1998) 'REVEAL, a general reverse engineering algorithm for inference of genetic network architectures', *Pacific Symposium on Biocomputing*, Vol. 3, pp.18–29.

Liebler, D.C. (2002) *Introduction to Proteomics*, Humana Press, Totowa, NJ.

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1992) *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed., Cambridge University Press, Cambridge, UK.

Schwarz, G. (1978) 'Estimating the dimension of a model', *Annals of Statistics*, Vol. 6, pp.461–464.

Somogyi, R. and Sniegoski, C.A. (1996) 'Modeling the complexity of genetic networks: understanding multigenic and pleiotropic regulation', *Complexity*, Vol. 1, pp.45–63.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998) 'Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization', *Mol. Biol.*, Vol. 9, pp.3273–3297.

Spieth, C., Streichert, F., Speer, N. and Zell, A. (2004) 'Iteratively inferring gene regulatory networks with virtual knockout experiments', *Proc. 2nd European Workshop on Evolutionary Bioinformatics*, Coimbra, Portugal, April 5–7, pp.104–112.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. and Church, G.M. (1999) 'Systematic determination of genetic network architecture', *Nature Genetics*, Vol. 22, pp.281–285.

Tipping, M.E. and Bishop, C.M. (1999) 'Probabilistic principal component analysis', *Journal of the Royal Statistical Society, Series B*, Vol. 61, pp.611–622.

Tozeren, A. and Byers, S.W. (2004) *New Biology for Engineers and Computer Scientists*, Pearson Education Inc., New Jersey.

Wessels, L.F.A., van Someren, E.P. and Reinders, M.J.T. (2001) 'A comparison of genetic network models', *Pacific Symposium on Biocomputing*, Vol. 6, pp.508–519.

Wichert, S., Fokianos, K. and Strimmer, K. (2004) 'Identifying periodically expressed transcripts in microarray time series data', *Bioinformatics*, Vol. 20, pp.5–20.

Wu, F.X., Zhang, W.J. and Kusalik, A.J. (2004a) 'Modeling gene expression from microarray expression data with state-space equations', *Pacific Symposium on Biocomputing*, Vol. 9, pp.581–592.

Wu, F.X., Zhang, W.J. and Kusalik, A.J. (2004b) 'State-space model with time delays for gene regulatory networks', *Journal of Biological Systems*, Vol. 12, pp.483–499.

Wu, F.X., Poirier, G.G. and Zhang, W.J. (2005) 'Inferring gene regulatory networks with time delays using a genetic algorithm', *IEE Proc. Systems Biology*, Vol. 152, pp.67–74.

Yeung, K.Y., Fraley, C., Murua, A., Rafterym, A.E. and Ruzzo, W.L. (2001) 'Model-based clustering and data transformations for gene expression data', *Bioinformatics*, Vol. 17, pp.977–987.