Modeling Gene Expression from Microarray Expression Data with State-Space Equations

F.X. Wu, W.J. Zhang, and A.J. Kusalik

MODELING GENE EXPRESSION FROM MICROARRAY EXPRESSION DATA WITH STATE-SPACE EQUATIONS

F. X. WU[1], W. J. ZHANG[1], A. J. KUSALIK[1,2]

[1] Division of Biomedical Engineering, [2] Department of Computer Science,
University of Saskatchewan, 57 Campus Dr., Saskatoon, SK, S7N 5A9, CANADA
faw341@mail.usask.ca; *zhangc@engr.usask.ca*; *kusalik@cs.usask.ca*

We describe a new method to model gene expression from time-course gene expression data. The modelling is in terms of state-space descriptions of linear systems. A cell can be considered to be a system where the behaviours (responses) of the cell depend completely on the current internal state plus any external inputs. The gene expression levels in the cell provide information about the behaviours of the cell. In previously proposed methods, genes were viewed as internal state variables of a cellular system and their expression levels were the values of the internal state variables. This viewpoint has suffered from the underestimation of the model parameters. Instead, we view genes as the observation variables, whose expression values depend on the current internal state variables and any external input. Factor analysis is used to identify the internal state variables, and Bayesian Information Criterion (BIC) is used to determine the number of the internal state variables. By building dynamic equations of the internal state variables and the relationships between the internal state variables and the observation variables (gene expression profiles), we get state-space descriptions of gene expression model. In the present method, model parameters may be unambiguously identified from time-course gene expression data. We apply the method to two time-course gene expression datasets to illustrate it.

1. Introduction

With advances in DNA microarray technology[1,2] and genome sequencing, it has become possible to measure gene expression levels on a genomic scale[3]. Data thus collected promise to enhance fundamental understanding of life on the molecular level, from regulation of gene expression and gene function to cellular mechanisms, and may prove useful in medical diagnosis, treatment, and drug design. Analysis of these data requires mathematical tools that are adaptable to the large scale of the data, and capable of reducing the complexity of the data to make it comprehensible. Substantial effort is being made to build models to analyze it.

Non-hierarchical clustering techniques such as k-means clustering are a class of mixture model-based approaches[4]. They group genes with similar expression patterns and have already proven useful in identifying genes that contribute to common functions and are therefore likely to be coregulated[5,6,7,8]. However, as pointed out by Holter et al.[9], whether information about the underlying genetic architecture and regulatory interconnections can be derived from the analysis of gene expression patterns remains to be determined. It is also important to note that models based on clustering analysis are static and thus can not describe the dynamic evolution of gene expression.

Boolean network can be applied to gene expression, where a gene's expression (state) is simplified to being 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 other genes. The functions can be represented in tables, or as rules. And example of the latter is "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 proceeds from one state (or time point) to the next, the pattern of currently expressed/non-expressed genes is used as input to rules which specify which genes will be "on" at the next state or time point. Somogyi and Sniegoski[10] showed that such Boolean networks have features similar to those in biological systems, such as global complex behaviour, self-organization, stability, redundancy, and periodicity. Liang et al.[11] described an algorithm for inferring genetic network architectures from the rules table of a Boolean network model. Their computational experiments showed that a small number of state transition pairs are sufficient to infer the original observations.

Akutsu et al.[12] 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 (of discrete variables), dynamic models (of continuous variables) have also been applied to gene expression. Chen et al.[13] 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 gene regulatory network should be sparse, they 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[13] and Akutsu[12] assume that all genes have a fixed maximum connectivity degree h (often small). These assumptions obviously contradict biological reality. For instance, some genes are known to have many regulatory inputs, while others are not known to have more than a few.

Another shortcoming of the previous work is that the fixed maximum connectivity degree h of Chen et al.[13] is chosen in an *ad hoc* manner. De Hoon et al.[14] considered Chen's differential model and used Akaike's Information Criterion (AIC) to determine the connectivity degree h of each gene. In their method, not all

genes must have a fixed connectivity. However, they do not present an efficient algorithm to identify the parameters of their differential equation model; the brute-force algorithm used in the paper[14] has a computational complexity of $O(2^{n^2})$, where n is the number of genes in the model. The authors claim that their method can be applied to find a network among individual genes. However, for biologically realistic regularity networks, the computational complexity is prohibitive. For instance, De Hoon et al. do not build any gene expression models among individual genes and instead choose to group the genes into several clusters and only study the interrelationships between the clusters.

D'haeseleer et al.[15] proposed a linear model for mRNA expression levels during CNS (stands for Central Nervous System) development and injury. To deal with the lack of gene expression data, the authors used a nonlinear interpolation scheme 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, while authors built a linear model for 65 measured mRNA species, there exists a problem of dimensional disaster when the number of genes in a model is large, for example, about 6000 (the number of genes in yeast).

Recently we have investigated strategies[16] for identifying gene regulatory networks from gene expression data with a state-space description of the gene expression model. We have found that modeling gene expression is key to inferring the regulatory networks among individual genes. Therefore, in the paper we focus on modeling gene expression.

The contributions of this paper are as follows:

- A state-space description of a gene expression dynamic model is proposed, where gene expression levels are viewed as the observation variables of a cellular system, which in turn are linear combinations of the internal variables of the system.
- Factor analysis is used to separate the internal variables and calculate their expression values from the values of the observation variables (gene expression data), where Bayesian Information Criterion (BIC) is used to determine the number of the internal variables
- The method is applied to two time-course gene expression datasets. The results suggest that it is possible to determine unambiguously a gene expression dynamic model from limited of time-course gene expression data.

2. Methods

Chen et al.[13] theoretically model biological data with the following linear differential equations:

$$\frac{d}{dt}\mathbf{x}(t) = \mathbf{\Lambda} \cdot \mathbf{x}(t) \tag{1}$$

where the vector $\mathbf{x}(t) = [x_1(t) \quad \cdots \quad x_n(t)]^T$ contains the mRNA and/or protein concentrations as a function of time t, the matrix Λ is constant and represents the extent or degree of regulatory relationships among genes and/or proteins, and where n is the number of genes and/or proteins in the model. The superscript "T" in the formula indicates the transposition of a vector.

D'haeseleer et al.[15] proposed the following linear difference equations to model gene expression data:

$$\mathbf{x}(t + \Delta t) = \mathbf{W} \cdot \mathbf{x}(t) \tag{2}$$

where the vector $\mathbf{x}(t) = [x_1(t) \quad \cdots \quad x_n(t)]^T$ contains gene expression levels as a function of time t, the matrix $\mathbf{W} = [w_{ij}]_{n \times n}$ represents regulatory relationships and degrees among genes, and n is the number of genes in the model. In detail, $x_i(t + \Delta t)$ is the expression level of gene i at time $t + \Delta t$, and w_{ij} indicates how much the level of gene j influences gene i when time goes from t to $t + \Delta t$.

Models (1) and (2) are equivalent. When Δt tends to zero, model (2) may be transformed into model (1). On the other hand, to identify the parameters in model (1), one must descretize it into the formalism of model (2). Since gene expression data from DNA microarray can only be obtained at a series of discrete time points with the present experimental technologies, difference equations are employed to model gene expression data in this paper. In addition, in DNA microarray experiments usually only the gene expression levels are determined, while the concentrations of resulting proteins are unknown. Therefore this work only considers constructing a system describing a gene expression dynamic model.

In Boolean network model, model (1) or model (2) genes are viewed as state variables in a cellular system. This makes parameter identification of the models impossible without other additional assumptions when using microarray data. In addition, previous models assume that regulatory relationships among genes are direct; for example, gene j directly regulating gene i with the weight w_{ij} in model (2). In fact, genes may not be regulated in such a direct way in a cellular system and may be regulated by some internal regulatory elements[17].

The following state-space description of a gene expression model is proposed to model gene expression evolution

$$\begin{cases} \mathbf{z}(t + \Delta t) = \mathbf{A} \cdot \mathbf{z}(t) + \mathbf{n}_1(t) \\ \mathbf{x}(t) = \mathbf{C} \cdot \mathbf{z}(t) + \mathbf{n}_2(t) \end{cases} \tag{3}$$

where, in terms of linear system theory[18], equations (3) are called the state-space description of a system. The vector $\mathbf{x}(t) = [x_1(t) \quad \cdots \quad x_n(t)]^T$ consists of the

observation variables of the system and $x_i(t)$ $(i = 1, \cdots, n)$ represents the expression level of gene i at time t, where n is the number of genes in the model. The vector $\mathbf{z}(t) = [z_1(t) \quad \cdots \quad z_p(t)]^T$ consists of the internal state variables of the system and $z_i(t)$ $(i = 1, \cdots, p)$ represents the expression value of internal element i at time t which directly regulates gene expression, where p is the number of the internal state variables. The matrix $A = [a_{ij}]_{p \times p}$ 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 $C = [c_{ik}]_{n \times p}$ is the transformation matrix between the observation variables and the internal state variables. The entries of the matrix encode information on the influences of the internal regulatory elements on the genes. Finally, the vectors $\boldsymbol{n}_1(t)$ and $\boldsymbol{n}_2(t)$ stand for system noise and observation noise. For simplicity, noise is ignored in this development.

Let $\mathbf{X}(t)$ 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 building of model (3) from microarray gene expression data $X(t)$ may be divided into two phases. Phase one identifies the internal state variables and their expression matrix $\mathbf{Z}(t)$ with p rows and m columns from the data matrix $\mathbf{X}(t)$ and computes the transformation matrix C such that

$$\mathbf{X}(t) = \mathbf{C} \cdot \mathbf{Z}(t) \tag{4}$$

Phase two builds the difference equations of the internal states; i.e. determine the state transition matrix A from the expression matrix $\mathbf{Z}(t)$.

In the process of building model (3), phase one, i.e. to establishing equations (4), is key. There are many methods that may be used to get decomposed equations (4) describing the gene expression data. For example, one may employ cluster analysis[14,19], where the means of the clusters may be viewed as the internal variables. One may also employ singular value decomposition[9,20], where the characteristic modes or eigengenes may be viewed as the internal variables. However, in typical applications of cluster analysis and singular value decomposition, the number of such internal variables is chosen in *ad hoc* fashion, with the result that matrix C and the expression data matrix of the internal variables $\mathbf{Z}(t)$ are decided subjectively rather than from the data themselves. Note that the matrices C and $\mathbf{Z}(t)$ are dependent. After $\mathbf{Z}(t)$ is identified, $\mathbf{C}$ may be calculated by formula $\mathbf{C} = \mathbf{X}(t) \cdot \mathbf{Z}^+(t)$, where $\mathbf{Z}^+(t)$ is a unique Moore-Penrose generalized inverse of the matrix $\mathbf{Z}(t)$.

Next, maximum likelihood factor analysis[4,21,22] is used to identify the internal state variables, and BIC is used to determine the number of the internal state variables,

where $\mathbf{X}(t)$ is the $n \times m$ observed data matrix, C is the $n \times p$ unobserved factor-score matrix and $\mathbf{Z}(t)$ is the $p \times m$ loaded matrix. In fact, both the generalized likelihood ratio test (GLRT) and the Akaike's information criterion (AIC) method[23] also may be used to determine the number of the internal variables, but they have a similar drawback, as the sample size increases there is an increasing tendency to accept the more complex model[24]. The BIC takes sample size into account. Although the BIC method was developed from a Bayesian standpoint, the result is insensitive to the prior distribution for adequate sample size. Thus a prior distribution does not need to be specified[24,25], which simplifies the method. For each model, the BIC is calculated as

$$BIC = -2 \cdot \begin{bmatrix} \log-\text{likelihood of the} \\ \text{estimation model} \end{bmatrix} + \log(n) \cdot \begin{bmatrix} \text{number of the estimated} \\ \text{parameters in the model} \end{bmatrix} \tag{5}$$

where n is the sample size. As with AIC, the model with the smallest BIC is chosen. BIC avoids the overfitting of a model to data.

After obtaining the expression data matrix of the internal variables $\mathbf{Z}(t)$ and the transformation matrix C in phase one, we develop the difference equations in model (3)

$$\mathbf{z}(t + \Delta t) = \mathbf{A} \cdot \mathbf{z}(t) \tag{6}$$

from the data matrix $\mathbf{Z}(t)$ in phase two. The matrix A contains p^2 unknown elements while the matrix $\mathbf{Z}(t)$ contains $m \cdot p$ known expression data points. If $p > m$, equations (6) will be underdetermined. Fortunately, using BIC the number of chosen internal variables p generally is less than the number of time points m. Therefore matrix A is identifiable.

To determine matrix A, the time step Δt is chosen to be the highest common factor among all of the experimentally measured time intervals so that the time of the j th measurement is $t_j = n_j \cdot \Delta t$, where n_j is an integer. For equally spaced measurements, $n_j = j$. We define a time-variant vector $v(t)$ with the same dimensions as the internal state vector $\mathbf{z}(t)$ and with the initial value $v(t_0) = z(t_0)$. For all subsequent times, $v(t)$ is determined from $v(t + \Delta t) = A \cdot v(t)$. For any integer k, we have

$$v(t_0 + k \cdot \Delta t) = A^k \cdot v(t_0) \ . \tag{7}$$

The p^2 unknown elements of the matrix A are chosen to minimize the cost function (the sum of squared relative errors)

$$CF = \sum_{j=1}^{m} \left\| z(t_j) - v(t_j) \right\|^2 / \sum_{j=1}^{m} \left\| z(t_j) \right\|^2 \tag{8}$$

where $\left\| \bullet \right\|$ stands for the Euclidean norm of a vector. For equally spaced measurements, the problem is a linear regression one and the solution to minimizing the cost function (8) can be a least square one. For unequally spaced measurements, the problem becomes nonlinear, and it is necessary to determine matrix A by using an optimization technique such as those in chapter 10 of Press's text[26].

3. Applications
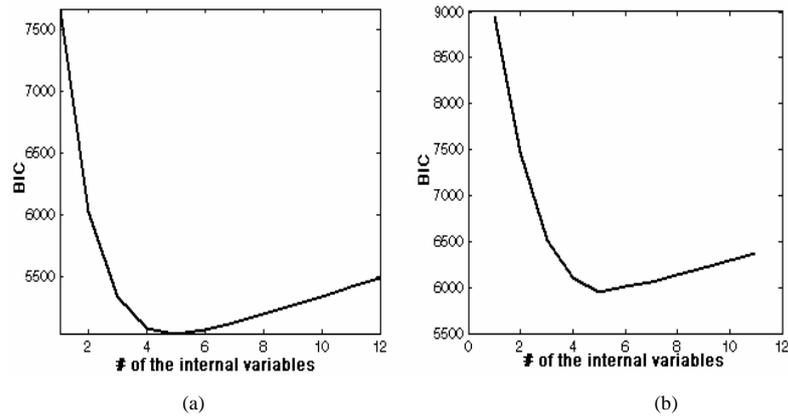
Figure 1. Profiles of BIC with respect to the number of the internal variables for (a) CDC15 data and (b) BAC data.

In this section, the proposed methodology was applied to two publicly available microarray datasets. The first dataset (CDC15) is from Spellman et al.[27] and consists of the expression data of 799 cell-cycle related genes for the first 12 equally spaced time points representing the first two cycles. The dataset is available at http://cellcycle-www.stanford.edu, and missing data were imputed by the mean values of the microarrays. The second dataset (BAC) is from Laub at al.[28] and consists of the expression data of 1590 genes for 11 equally spaced time points with no missing data. The dataset is available is at http://caulobacter.stanford.edu /CellCycle. As the mean values and magnitudes for genes and microarrays mainly reflect the experimental procedure, we normalize the expression profile of each gene to have length one and then for expression values on each microarray as so to have mean zero and length one. Such normalizations also make factor analysis simple[22].

Table 1. The internal variable expression matrices

CDC					BAC				
-0.2065	0.2914	-0.5766	0.2401	-0.0886	-0.4478	0.0733	-0.5429	0.0938	-0.1839
-0.7472	0.0812	-0.4848	0.1591	-0.0418	-0.6954	0.2965	-0.4481	0.0018	-0.2020
-0.5397	-0.6201	-0.2144	0.1406	-0.0389	-0.8355	0.4048	0.0408	-0.2612	0.0739
0.2695	-0.7875	-0.0898	0.0950	0.1159	-0.7904	0.2241	0.1674	0.0162	0.0252
0.7960	-0.3190	-0.2828	-0.0038	0.1283	-0.7850	0.2158	0.2685	0.0289	0.0021
0.6692	0.4116	-0.3365	-0.0460	0.1430	-0.8141	-0.0381	0.2671	0.2602	-0.1303
-0.4139	0.4091	-0.3770	-0.4557	-0.0130	-0.7410	-0.4120	0.1512	0.0618	-0.0864
-0.7042	-0.2534	-0.0028	-0.4060	0.0820	-0.6371	-0.5639	0.0442	-0.2583	-0.1583
-0.3371	-0.6247	0.0893	-0.1332	-0.0618	-0.5635	-0.4091	-0.1484	-0.2821	0.0947
0.5592	-0.4646	-0.1469	-0.0957	-0.3433	-0.7409	-0.2597	-0.2584	0.1761	0.3170
0.7490	0.0429	-0.1504	-0.1983	-0.2431	-0.7777	-0.0906	-0.1943	0.1666	0.1007
0.0216	0.5261	0.2677	0.2599	-0.1465					

The EM algorithm for maximum likelihood factor analysis[23] was employed for the two datasets. The gene expression profile for one gene is one sample observation and the identified parameters are the $p \cdot m$ elements of the matrix $\mathbf{Z}(t)$ and the variances of m residue errors[23]. Figure 1 depicts the profiles of BIC with respect to the number of internal variables. Clearly from Figure 1, 5 is the best choice as the number of internal variables for both datasets. The expression matrices for the internal varaibles are listed in Table 1, where each column describes one internal variable.

Table 2. The state transition matrix of the internal variables

CDC15					BAC				
$A = [$0.4378	-1.0077	0.5009	0.1851	-0.1189	$A = [$0.4378	-1.0077	0.5009	0.1851	-0.1189
0.6649	0.5244	0.2475	0.1511	-0.1356	0.6649	0.5244	0.2475	0.1511	-0.1356
-0.0702	0.1734	0.6794	-0.3092	-0.5279	-0.0702	0.1734	0.6794	-0.3092	-0.5279
-0.0699	-0.0103	0.1786	0.6163	-0.5190	-0.0699	-0.0103	0.1786	0.6163	-0.5190
0.0161	0.0316	-0.0700	0.1358	0.6662]	0.0161	0.0316	-0.0700	0.1358	0.6662]

In order to determine the state transition matrices in the models from the internal expression matrices, we solve two optimization problems (8), for the two datasets. As both datasets are equally spaced measurements, the least square method can be used to obtain the two state transition matrices A in the models shown in Table 2. Figure 2 gives a comparison of the internal variable expression profiles in Table 1 and their calculated profiles from the model (3) for (a) CDC15 and (b) BAC,

respectively. The values of the cost functions are 0.2321 and 0.0761 for the CDC15 dataset and the BAC dataset, respectively. That is, at each time point the average relative errors between the internal variable profiles in Table 1 and their calculated values by model (3) are 0.0622 and 0.0372 for the CDC15 dataset and the BAC dataset, respectively. Therefore, two state transition matrices in Table 2 are plausible.

(a)   (b)
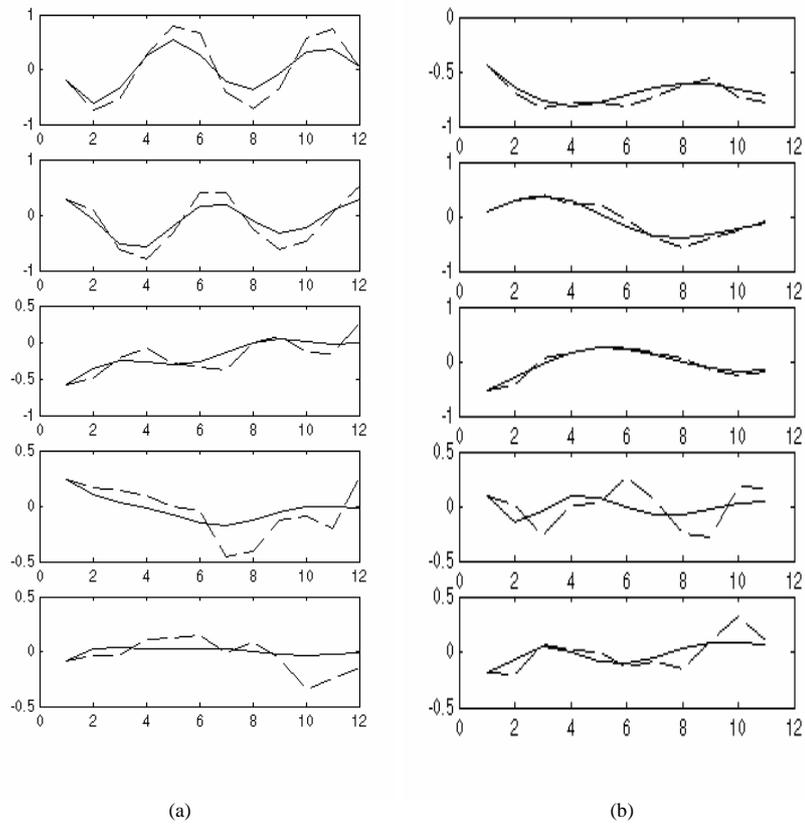
Figure 2. A comparison of the internal variable expression profiles in table 1 and their calculated profiles from the model (3) for (a) CDC15 and (b) BAC. The solid lines correspond to the profiles in table 1 and the dash lines to the calculated profiles from the model (3).

Since an exponential or a polynomial growth rate of a gene expression is unlikely to happen, the gene expression systems are assumed to be a stable system[13]. This means that all eigenvalues of the state transition matrix A in model (3) should lie

inside the unit circle if model (3) describes a gene expression dynamic system. Five eigenvalues of the state transition matrix A for the CDC15 dataset are $0.4262 - 0.8488i$, $0.4262 + 0.8488i$, 0.5509, $0.7605 - 0.2950i$, and $0.7605 + 0.2950i$, all of which lie inside the unit circle. Five eigenvalues of the state transition matrix A for BAC dataset are 1.0282, $0.6835 - 0.4997i$, $0.6835 + 0.4997i$, $0.3092 - 0.5769i$, and $0.3092 + 0.5769i$. All of these except for the first one lie inside the unit circle. However, the first eigenvalue is very close to 1. Since these two systems are (almost) stable, they are robust to system noise, for example, the squared summable noises. Therefore, these two models are sound to gene expression dynamic systems.

4. Discussion

This paper proposes a method to model gene expression dynamics from measured time-course gene expression data. The model is in the form of the state-space description of linear systems. Two gene expression models for two previously published gene expression datasets were constructed to show how the method works. The results demonstrate that some of features of the models are consistent with biological knowledge. For example, genes may be regulated by internal regulatory elements[17], and gene expression dynamic systems are stable and robust[29].

Compared to previous models, our model (3) has the following characteristics. First gene expression profiles are the observation variables rather than the internal state variables. Second, and from a biological angle, our model (3) can capture the fact that genes may be regulated by internal regulatory elements[17]. Finally, although it contains two groups of equations (one is a group of difference equations and the other, algebraic equations), the parameters in model (3) are identifiable from existing microarray gene expression data without any assumptions on the connectivity degrees of genes[11,12,13,14] and the computational complexity to identify them is simple.

The main shortcomings of this approach are: 1) the inherent linearity which can only capture the primary linear components of a biological system which may be nonlinear; 2) the ignorance to time delays in a biological system resulting, for example, from the time necessary for transcription, translation, and diffusion; 3) the failure to handle external inputs and noise. In the future work, we will address these shortcomings, especially the latter one. In addition, the present approach will be applied to more datasets and the biological relevance of the internal variables will be demonstrated. This last goal requires closer collaborations with biologists.

We can not expect to obtain perfect gene expression models which can completely explain organismal or suborganismal behaviours from existing gene expression data at this time. On the other hand, any subjective assumptions-enforced models may result in misinterpreting organismal or suborganismal behaviours. Using the present methodology one may sufficiently explore the data to

construct sound models, which is what data can tell us. We believe that our method, along with the results of the application to two datasets, advances gene expression modelling from time-course gene expression datasets.

Acknowledgements

Reference

1. Pease, A. C., *et al*. "Light-Generated Oligonucleotide Arrays for Rapid DNA Sequence Analysis" *Proc. Natl. Acad. Sci. USA* **91**: 5022-5026, (1994).
2. Schena, M., *et al*. "Quantitative monitoring of gene expression patterns with a complementary DNA microarray" *Science* **270**: 467-470, (1995).
3. Sherlock, G., *et al*. "The Stanford Microarray Database" *Nucleic Acids Research* **29**: 152-155, (2001).
4. Everitt, B. S. and Dunn, G. "Applied Multivariate Data Analysis" New York: *Oxford University Press*, (1992).
5. Tavazoie, S., *et al*. "Systematic determination of genetic network architecture*", Nature genetics* **22**: 281-285, (1999).
6. Yeung, K.Y, *et al*. "Model-based clustering and data transformations for gene expression data", *Bioinformatics* **17**: 977-987, (2001).
7. Ghosh, D. and Chinnaiyan, A. M. "Mixture modelling of gene expression data from microarray experiments" *Bioinformatics* **18**: 275-286, (2002).
8. McLachlan, G. J., Bean, R. W., and Peel, D. A. "Mixture model-based approach to the clustering of microarray expression data"**,** *Bioinformatics* **18**: 413-422, (2002).
9. Holter, N. S., *et al*. "Dynamic modeling of gene expression data" *Proc. Natl. Acad. Sci. USA* **98**: 1693-1698, (2001).
10. Somogyi, R. and Sniegoski, C. A. "Modeling the complexity of genetic networks: Understanding multigenic and pleiotropic regulation" *Complexity* **1**: 45-63, (1996).
11. Liang, S., *et al*. "REVEAL, A general reverse engineering algorithm for inference of genetic network architectures" *Pacific Symposium on Biocomputing* **3**: 18-29, (1998).
12. Akutsu, T., *et al*. "Identification of gene networks from a small number of gene expression patterns under the Boolean network model" *Pacific Symposium on Biocomputing* **4**: 17-28, (1999).

13. Chen, T., He, H. L., and Church, G. M. "Modeling Gene Expression with Differential Equations" *Pacific Symposium on Biocomputing* **4**: 29-40, (1999).

14. de Hoon, M. J. L., *et al*. "Inferring Gene Regulatory Networks from Time-Ordered Gene Expression Data of Bacillus Subtilis Using Differential Equations" *Pacific Symposium on Biocomputing* **8**: 17-28, (2003).

15. D'haeseleer, P., *et al*. "Linear Modeling of mRNA Expression Levels During CNS Development and Injury" *Pacific Symposium on Biocomputing* **4**: 41-52, (1999).

16. Wu, F. X., *et al*. "Reverse engineering gene regulatory networks using the state-space description of microarray gene expression data" in preparation.

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

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

19. van Someren, E. P., Wessels, L. F. A., and Reinders, M.J.T. "Linear modeling of genetic networks from experimental data" In *Proceedings of the Eight International Conference on Intelligent Systems for Molecular Biology* (ISMB 2000), La Jolla, California, USA, (2000).

20. Alter, O., Brown, P. O., and Botstein, D. "Singular value decomposition for genome-wide expression data processing and modeling" *Proc. Natl. Acad. Sci. USA* **97**: 10101-10106, (2000).

21. Lawley, D. N. and Maxwell, A. E. "Factor Analysis as a Statistical Method" 2ed, London: *Buuterorth*, (1971).

22. Bubin, D. B. and Thayer, D. T. "EM algorithms fro ML factor analysis" Psychometrika **47**: 69-76, (1982).

23. Burnham, K. P. and Anderson, D. R., "Model selection and inference: a practical information-theoretic approach" New York: *Springer*, (1998).

24. Raftery, A. E. "Choosing models for cross-classification" *American Sociological Review* **51**: 145-146, (1986).

25. Schwarz, G. "Estimating the dimension of a model" *Annals of Statistics* **6**: 461-464, (1978).

26. Press, W. H. *et al*. "Numerical Recipes in C: The Art of Scientific Computing" 2nd edition, Cambridge, UK: *Cambridge University Press*, (1992).

27. Spellman, P. T., *et al*. "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization" *Mol. Biol.* **9**: 3273-3297, (1998).

28. Laub, M. T. "Global analysis of the genetic network controlling a bacterial cell cycle" *Science* **290**: 2144-2148, (2000).

29. Hartwell, L. H., *et al*. "From molecular to modular cell biology" *Nature* **402**: C47 – 52, (1999).