BIOINFORMATICS ORIGINAL PAPER Vol. 24 no. 7 2008, pages 932–942
doi:10.1093/bioinformatics/btm639

Gene expression

Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models

Osamu Hirose[1,†], Ryo Yoshida[2,*,†], Seiya Imoto[1], Rui Yamaguchi[1], Tomoyuki Higuchi[2], D. Stephen Charnock-Jones[3], Cristin Print[4] and Satoru Miyano[1]

[1]Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan, [2]Institute of Statistical Mathematics, Research Organization of Information and Systems, 4-6-7 Minami-Azabu, Minato-ku, Tokyo, 106-8569, Japan, [3]Department of Pathology, Cambridge University, Tennis Court Road, Cambridge, CB2 1QP, UK and [4]Department of Molecular Medicine and Pathology, University of Auckland, Private Bag 92019, Auckland, New Zealand

Received on May 3, 2007; revised on December 10, 2007; accepted on December 28, 2007

Advance Access publication February 21, 2008

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: Statistical inference of gene networks by using time-course microarray gene expression profiles is an essential step towards understanding the temporal structure of gene regulatory mechanisms. Unfortunately, most of the current studies have been limited to analysing a small number of genes because the length of time-course gene expression profiles is fairly short. One promising approach to overcome such a limitation is to infer gene networks by exploring the potential transcriptional modules which are sets of genes sharing a common function or involved in the same pathway.

Results: In this article, we present a novel approach based on the state space model to identify the transcriptional modules and module-based gene networks simultaneously. The state space model has the potential to infer large-scale gene networks, e.g. of order 10^3, from time-course gene expression profiles. Particularly, we succeeded in the identification of a cell cycle system by using the gene expression profiles of *Saccharomyces cerevisiae* in which the length of the time-course and number of genes were 24 and 4382, respectively. However, when analysing shorter time-course data, e.g. of length 10 or less, the parameter estimations of the state space model often fail due to overfitting. To extend the applicability of the state space model, we provide an approach to use the technical replicates of gene expression profiles, which are often measured in duplicate or triplicate. The use of technical replicates is important for achieving highly-efficient inferences of gene networks with short time-course data. The potential of the proposed method has been demonstrated through the time-course analysis of the gene expression profiles of human umbilical vein endothelial cells (HUVECs) undergoing growth factor deprivation-induced apoptosis.

Availability: Supplementary Information and the software (TRANS-MNET) are available at http://daweb.ism.ac.jp/~yoshidar/software/ssm/

Contact: yoshidar@ism.ac.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Time-course experiments of gene expression facilitate understanding of the temporal structure of gene regulatory mechanisms during development (Arbeitman *et al.*, 2002), drug response (Baranzini *et al.*, 2005; Imoto *et al.*, 2006; Tamada *et al.*, 2005) and cell cycle (Orlando *et al.*, 2007; Spellman *et al.*, 1998). In order to explore the dynamics of gene regulation by using the gene expression profiles, it is essential to build statistical models that consider the temporal relationships between abundance of different transcripts (Bansal *et al.*, 2006; Beal *et al.*, 2005; Gardner *et al.*, 2003; Rangel *et al.*, 2004; van Someran *et al.*, 2006; Yamaguchi *et al.*, 2007; Yoshida *et al.*, 2005). If successful, these models may have broad utility, e.g. the discovery of transcripts encoding drug targets and the identification of gene regulatory networks involved in drug responses or biological processes (Imoto *et al.*, 2006; Tamada *et al.*, 2005).

To analyse multivariate time-course data, a wide variety of statistical models have been proposed, e.g. the vector auto-regressive model (VAR). To our knowledge, their applications in time-course gene expression profiling, however, have been limited because the length of the time-course data is fairly short, e.g. typically less than 10, whereas the number of genes involved ranges from 10^2 to 10^4. Obviously, the length of time-course gene expression profiles is not sufficient to infer such a large gene network. For example, the maximum likelihood estimator of VAR does not exist if the number of genes is greater than the length of the time series.

One possible solution to overcome such a difficulty is to explore genetic networks of the transcriptional modules which are sets of genes sharing a common function or involved in the same pathway (Lee *et al.*, 2002; Segal *et al.*, 2003) rather than the use of gene-level networks. In the context of gene expression analysis, the transcriptional modules may be defined by the

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

groups of transcriptionally co-expressed genes. In this article, we provide an approach to identify the potential transcriptional modules and map them onto the gene-level networks, i.e. the module-based gene networks. The proposed method is based on the state space model (Kitagawa and Gersch, 1996) which has the potential to estimate large gene networks from time-course gene expression profiles. Indeed, as will be demonstrated, we succeeded in the identification of cell cycle system by using the time-course gene expression data of *Saccharomyces cerevisiae* (Spellman *et al*., 1998) in which the length of the time-course data and number of genes were 24 and 4382, respectively.

However, when analysing shorter time-course data, e.g. less than 10, the parameter estimations of the state space model still fail due to overfitting. To extend the applicability of the proposed method, we provide a way of using technical or biological replicates of time-course gene expression profiles, which are often measured in duplicate or triplicate in order to assess the reproducibility of data. The use of replicates proves important for achieving highly efficient inference of gene network. The potential of the proposed method will be demonstrated through the time-course analysis of the gene expression profiles of human umbilical vein endothelial cells (HUVECs) during growth factor deprivation-induced apoptosis.

2 STATE SPACE MODEL

2.1 Definition of model

Following the time-course experiments of gene expression, we obtain a series of gene expression vectors $y_n \in \mathcal{R}^p$, $n \in \mathcal{N}_{\text{obs}} \subset \mathcal{N}$, where each vector contains expression profiles of p genes at the n-th time point. The set of entire time points, $\mathcal{N}$, consists of the observed time set $\mathcal{N}_{\text{obs}}$ and the unobserved one $\mathcal{N}^c_{\text{obs}}$. Conventionally, the length of time-course gene expression data, denoted by $N = |\mathcal{N}_{\text{obs}}|$, is fairly short.

One of the most basic models to analyse multivariate time-course data is the first-order VAR

$$y_n = \Gamma y_{n-1} + \epsilon_n. \tag{1}$$

The ϵ_ns follow a zero-mean white noise process with a finite second moment and $\Gamma \in \mathcal{R}^{p\times p}$ is the coefficient matrix which represents the temporal relationships of the p genes. However, in its application to short time-course data in which a large number of genes are involved, the parameter estimations obviously fail due to overfitting because the number of free parameters in Γ becomes larger exponentially as the number of genes p increases. Indeed, it is conventional that the length of time-course expression profiles is usually much shorter than the number of genes, i.e. $N \ll p$.

The main challenge in this study is to address the problem of analysing $N \ll p$ time-course data. A key idea is to impose the parameter constraints on the coefficient matrix Γ of the VAR (1) by exploiting the state space model. Let $x_n \in \mathcal{R}^k$ be the lower-dimensional hidden state vector ($k < p$) which is a blind source for the generation of y_n. As a generative model, the state space model defines the observational equation as

$$y_n = Hx_n + w_n, \quad n \in \mathcal{N}_{\text{obs}}, \tag{2}$$

where $H \in \mathcal{R}^{p\times k}$ is the loading matrix and w_ns follow a white noise process. We assume that w_n is independently distributed according to the normal distribution with mean $\mathrm{E}[w_n] = 0$ and diagonal covariance matrix $\mathrm{E}[w_n w_n^T] = R \equiv \mathrm{diag}(r_1, \ldots, r_p)$. In addition to (2), the evolving time course of x_n is modelled by the first-order Markov process as

$$x_n = Fx_{n-1} + v_n, \quad n \in \mathcal{N}, \tag{3}$$

where $F \in \mathcal{R}^{k\times k}$, and v_ns follow the normal distribution with mean zero and covariance matrix Q. The process of generating y_n and x_n follows (2) and (3) with the initial state vector $x_0 \sim N(\mu_0, \Sigma_0)$. Note that the state vectors evolve at the overall successive time points $\mathcal{N}$ by following the system model (3) whereas the observational equation (2) is defined over its subset $\mathcal{N}_{\text{obs}}$.

The basic assumption of the state space model is that a dynamical behaviour of observed data y_ns is regulated by the time evolution of a few latent factors x_ns. In the context of gene regulations, a latent factor may be considered as an unobserved activity of transcription factors which regulate transcriptions of downstream target genes.

In computational systems biology, several state space models have been proposed for estimating temporal gene networks with successful applications (Beal *et al*., 2005; Li *et al*., 2006; Rangel *et al*., 2004; Wu *et al*., 2004; Yamaguchi *et al*., 2007; Yoshida *et al*., 2005). For example, Rangel *et al*. (2004) and Beal *et al*. (2005) built an input-driven state space model as

$$\begin{aligned} y_n &= Hx_n + Ay_{n-1} + w_n, \\ x_n &= Fx_{n-1} + By_{n-1} + v_n. \end{aligned} \tag{4}$$

The matrix $A \in \mathcal{R}^{p\times p}$ captures the causal relationships of genes and the matrix $B \in \mathcal{R}^{k\times p}$ captures the influences of the previous gene expressions to the current hidden state vectors. In this context, the evolving state vectors represent the latent factors which cannot be measured by gene expression profiles, e.g. genes unobserved by microarray experiments, concentrations of regulatory proteins such as transcription factors, some biological entities present in post-transcriptional modifications. Note that the input-driven model directly describes the temporal gene–gene relationships with the coefficient matrices A and B, whereas the standard state space model, i.e. (2) and (3), does not explicitly describe such direct relationships. Based on this modelling, Rangel *et al*. (2004) derived the gene network model as

$$y_n = (HB + A)y_{n-1} + HFx_{n-1} + w_n + Hv_n.$$

The temporal structure of gene regulations is inferred by the estimated $HB + A$. However, like the conventional VAR (1), with an increase in the number of genes, the coefficient matrices, e.g. A and B, cannot be estimated efficiently due to overfitting, hence, its applications are still limited to analysing very few genes if the length of time-course is short.

In this article, we focus on the standard state space model rather than the input-driven model. Below, we elucidate that the standard state space model implicitly represents a parsimonious parameterization of the first-order VAR in its canonical form. Based on this fact, we can infer the large scale gene

networks by using the standard state space model without the input-driven modifications.

2.2 Module-based gene networks

The parameter vector $\theta \in \Theta$ of the state space model contains all elements in H, F, R, Q and μ_0 where the covariance matrix of the initial state distribution, i.e. Σ_0, is assumed to be given.

We first discuss the lack of identifiability that occurs in the determination of the parameters. Let C be an arbitrary non-singular $k \times k$ matrix. By considering $H^* = HC^{-1}$, $x_n^* = Cx_n$, $F^* = CFC^{-1}$ and $v_n^* = Cv_n \sim N(0, CQC^T)$, the state space model can be transformed into the equivalent form as follows:

$$y_n = H^* x_n^* + w_n,$$
$$x_n^* = F^* x_{n-1}^* + v_n^*.$$

This implies that the parameters of the state space model cannot be uniquely determined by the ordinary estimation procedures, e.g. the maximum likelihood estimation. To rule out such an over-parameterization, we must reduce the degree of freedom of H, F and Q. To avoid it, we establish the following proposition:

Proposition: *The following conditions are sufficient to eliminate transformations of the parameters by any non-singular* $C \in \mathcal{R}^{k\times k}$;

- $Q = I$.
- $H^T R^{-1} H = \Lambda \equiv \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ where $\lambda_1 > \lambda_2 > \cdots > \lambda_k$.
- An arbitrary sign condition is imposed on the elements of the first row of H.

Proof: Due to the first condition $Q = I$, it holds that $CQC^T = CC^T = I$. Hence, the family of C is restricted being orthonormal matrices. Furthermore, according to the second condition $H^T R^{-1} H = \Lambda$, the transformed $CH^T R^{-1} HC^T = C\Lambda C^T$ must be a diagonal matrix $\Lambda^* = \mathrm{diag}(\lambda_1^*, \lambda_2^*, \ldots, \lambda_k^*)$. This implies that C must be a diagonal matrix having 1 or -1 at the diagonal elements. Finally, since the last condition imposes the sign condition on the first row of HC, we obtain $C = I$.

We refer (2) and (3) with these three conditions the canonical state space model.

Here, we derive the parsimonious parameterization of the VAR (1) based on the canonical state space model. By transforming the observational equation (2) under the specified constraints, gene expression vectors can be mapped onto the state space $\mathcal{R}^k$ with the projection matrix $D \in \mathcal{R}^{k\times p}$ as follows:

$$x_n = DR^{-1/2}(y_n - w_n), \quad n \in \mathcal{N}_{\mathrm{obs}}, \tag{5}$$

where the projection matrix is parameterized as

$$D = \Lambda^{-1} H^T R^{-1/2}. \tag{6}$$

If the state dimension k is specified as a value lower than p, the dimensionality of the noise-removed gene expression vectors $R^{-1/2}(y_n - w_n)$ is reduced by the semi-orthogonal projection matrix D. During the parameter estimation process, the reduced-rank data (5) are possibly constructed such that they are likely to follow the first-order Markov process of (3). This process automatically discovers k modules of genes that are relevant to the temporal structure of gene expressions in the following manner: effects of the j-th gene on the i-th module (i-th coordinate system of the state space) are removed by $(D)_{ij}$ lying in the region close to zero and vice versa. The system model (3) represents the temporal relationships between the k modules, i.e. the module networks.

Additionally, Yoshida *et al*. (2005) elucidated that the state space model implicitly represents the gene-level temporal structure with the autoregressive form

$$R^{-1/2}(y_n - w_n) = \Psi R^{-1/2}(y_{n-1} - w_{n-1}) + R^{-1/2} H v_n,$$

where the autoregressive coefficient matrix is given by

$$\Psi \equiv D^T \Lambda F D. \tag{7}$$

In (7), the degree of freedom in the coefficient matrix Ψ is of order $O(p) = p(k+1) + k^2 - k(k-1)/2$, whereas that $O(p^2)$ in the standard VAR (1). From this point of view, the state space model is considered as the parsimonious parameterization of the VAR, and it provides a method for controlling the model complexity by selecting the dimension of state vector k. In our previous work (Yoshida *et al*., 2005), we proposed the state space model with Markov switching in order to identify the structural changes of underlying gene regulatory mechanism. The model of Yoshida *et al*. (2005) assumes the time-dependent autoregressive coefficients in (7) that change smoothly over the successive time points. The autoregressive representation shown in (7) is a special case of that proposed by Yoshida *et al*. (2005).

Note that the coefficient matrix Ψ consists of the product of the projection matrix D, the scaled-system matrix ΛF, and the reconstruction matrix D^T, respectively. It represents the temporal structure of gene-level networks in the following manner: once the k modules $x_{n-1} = DR^{-1/2}(y_{n-1} - w_{n-1})$ are given at time $n-1$, the current modules are generated with the scaled-system matrix ΛF as $\Lambda F x_{n-1} = \Lambda F D R^{-1/2}(y_{n-1} - w_{n-1})$. Furthermore, the current modules regulate expression values of p genes in y_n with the reconstruction matrix D^T as $D^T \Lambda F x_{n-1} = \Phi R^{-1/2}(y_{n-1} - w_{n-1})$. Briefly, the projection matrix, the scaled-system matrix and the reconstruction matrix in (7) correspond to the transcriptional modulation, the module-module interactions and the module-gene interactions, respectively.

2.3 Network construction with permutation test

After estimating the parameters, the temporal relationships between genes can be inferred through the computed $\hat{\Psi}$. Furthermore, the statistical significance for the existence of each gene–gene relationship can be assessed by testing the hypothesis

$$H_0 : (\Psi)_{ij} = 0,$$
$$H_1 : (\Psi)_{ij} \neq 0,$$

for $i, j \in \{1, \ldots, p\}$. Rejection of the null hypothesis suggests that there exists a causal relationship from gene j to gene i across the successive time points.

In this study, the permutation method is used to evaluate the null distribution. Let $Y_N = \{y_n\}_{n\in\mathcal{N}_{\mathrm{obs}}}$ be the observed

data matrix. The method first generates the B permutation samples $Y_N^{(b)}$, $b=1,\ldots,B$, by applying random permutations to all elements in Y_N and then computes the null coefficients $(\hat{\Psi})_{ij}^{(b)}$. By using these estimates, the P-value for each coefficient can be evaluated by $p_{ij}=(2/B)\min\{\beta_g, \beta_l\}$ where $\beta_g = \#\{b : (\hat{\Psi})_{ij}^{(b)} \geq (\hat{\Psi})_{ij}\}$ and $\beta_l = \#\{b : (\hat{\Psi})_{ij}^{(b)} \leq (\hat{\Psi})_{ij}\}$.

3 APPLICATION TO GENE EXPRESSION PROFILES OF BUDDING YEAST

3.1 Proposed method

We demonstrate the potential of the canonical state space model with the application to the gene expression data of *S.cerevisiae*, which were measured by the cDNA microarray experiments conducted during the cell cycle (Spellman *et al.*, 1998). This dataset has been used for evaluating the ability of a wide variety of statistical technologies (Li *et al.*, 2006; Wu *et al.*, 2004; Yamaguchi *et al.*, 2007; Yoshida *et al.*, 2005). Following them, we decided to use this dataset for performing the benchmark comparisons.

Here, we present the gene expression analysis with the time-course data of cdc15-based synchronization. Evolving time-course data were measured at the 24 unequally spaced time points, $\mathcal{N}_{\text{obs}} = \{1, 3, 5, 7, 8, 9, \ldots, 24, 25, 27, 29\}$. For demonstration purposes, we focused on the 4382 genes that contained no missing data points.

After fitting the state space model under $k=4$ by applying the maximum likelihood estimation (for the algorithmic details, see Appendix 1), we captured the underlying temporal relationships of the potential transcriptional modules with the estimated system matrix $\hat{F}$. Figure 1 summarizes the estimated module network $\hat{F}$ and the expression profiles of the 8 ($=2k$) identified modules, where the genes listed at each module were selected in the following way: the j-th gene is assigned to the i-th positive module $\mathcal{M}_{i+}$ or the i-th negative module $\mathcal{M}_{i-}$ if the (i, j)-th element in the estimated projection matrix $(\hat{D})_{ij}$ is ranked in the highest or lowest 100, respectively.

As shown in Figure 1, we observed a very clear aggregation of the expression patterns in each module. For example, most of the time-courses in the first two modules $\mathcal{M}_{1+}$ and $\mathcal{M}_{1-}$ contain high-frequency components in which the up/down-regulations are periodically switched at intervals of 10 min for 80–210 min. Moreover, the genes in $\mathcal{M}_{4\pm}$ exhibit the upward/downward trends across all time points. The genes in $\mathcal{M}_{2\pm}$ and $\mathcal{M}_{3\pm}$ show the cyclic patterns where the periods are approximately 50 min or a slightly longer, but the phases are notably different from each other. These temporal patterns are captured through the estimated four state variables which are also shown in the bottom panel of Figure 1.

The aggregation of temporal expression patterns in the same module suggests the highly relevance of the genes in each module. For example, according to the molecular functions of Gene Ontology (GO), most of the genes in $\mathcal{M}_{1+}$ are related to the transportation of several chemical compounds, including sugar, glucose, carbohydrate and so on. Additionally, $\mathcal{M}_{1-}$ contains a large number of genes involved in the synthesis and modification of rRNAs. The genes relevant to ribosome biogenesis and oxidoreductase activity are over-represented

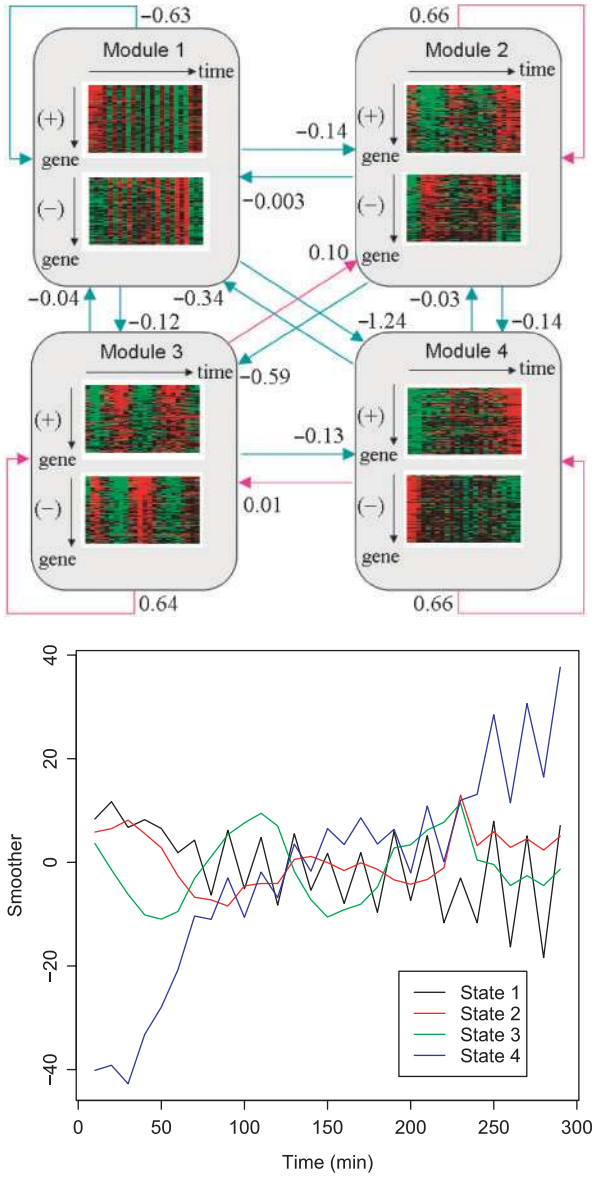

Fig. 1. (Top) Summary of the identified four modules and the estimated temporal structure of cell cycle regulatory network. The estimated $(\hat{F})_{ij}$ is assigned to each edge. Temporal expression patterns of the most representative 100 genes for each module are shown in each node. (Bottom) The estimated four state variables given by the posterior means $\mathrm{E}[x_n|y_n \in \mathcal{N}_{\text{obs}}]$.

in $\mathcal{M}_{4+}$ and $\mathcal{M}_{4-}$, respectively. Finally, a large number of cell cycle-related genes are captured by $\mathcal{M}_{2\pm}$ and $\mathcal{M}_{3\pm}$. Some over-represented GO terms for the identified transcriptional modules are summarized in Supplementary Table 1.

The inferred temporal structures between the identified modules $\hat{F}$ are summarized in Figure 1. First, we focused on the genes in $\mathcal{M}_{1\pm}$ having the self-loop edge to which the negative autocorrelation was assigned $(\hat{F})_{11}=-0.63$. This negative autocorrelation captures the high-frequency contents of $\mathcal{M}_{1\pm}$. Moreover, a strong negative influence from $\mathcal{M}_{1\pm}$ to $\mathcal{M}_{4\pm}$, $(\hat{F})_{41}=-1.24$ was identified while a weak negative

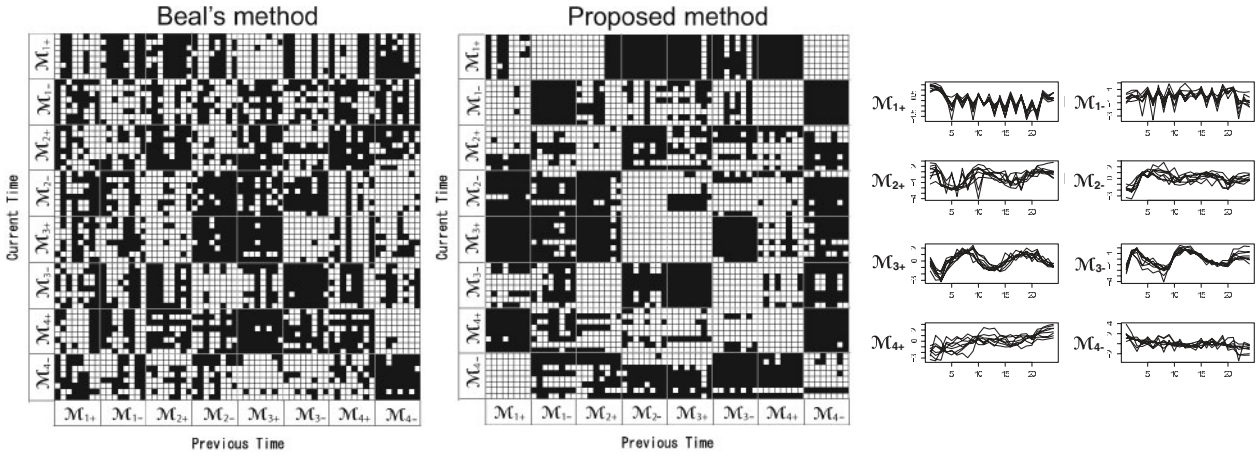

Fig. 2. Heatmap representation of the estimated coefficient matrices $\hat{\Psi}$ corresponding to Beal's method (left) and our proposed method (middle). The 64 genes shown are composed of the eight modules $\mathcal{M}_i$ and $\mathcal{M}_{i-}$, $i = 1, \ldots, 4$ where the expression patterns are shown in the right panel. The positive and negative coefficients are depicted by white and black pixels, respectively.

influence of $\mathcal{M}_{4\pm}$ on $\mathcal{M}_{1\pm}$ was also observed, $(\hat{F})_{41} = -0.34$. Note that the observed expression patterns of $\mathcal{M}_{4\pm}$ are composed of a mixture of trend and periodic components. The estimated coefficient $(\hat{F})_{41} = -1.24$ suggested that the periodic components in $\mathcal{M}_{4\pm}$ are derived from those of $\mathcal{M}_{1\pm}$. Notably, according to the GO cellular component, most of the genes in $\mathcal{M}_{1-}$ and $\mathcal{M}_{4+}$ are annotated by the 'nucleolus'.

3.2 Comparison to input-driven model

Here we present a numerical experiment for comparing the performance of the proposed state space model and the input-driven state space model in (4) developed by Rangel *et al.* (2004) and Beal *et al.* (2005). Originally, Rangel *et al.* (2004) and Beal *et al.* (2005) used the same input-driven model to infer somewhat small-scale gene networks from gene expression data, but they differ in the way of the parameter estimation scheme: The former and latter adopted the maximum likelihood estimation and the Bayesian parameter estimation, respectively. For performing these two methods, we used the distributed MATLAB programs, LDS_ToolBox and VBSSM (Ver.3.4.1) (http://public.kgi.edu/wild/LDS/index.htm and http://www.cse.buffalo.edu/faculty/mbeal/software.html), respectively.

We used the benchmark data set which are comprised of 4321 genes showing no missing time points. By definition, the maximum likelihood estimator of the input-driven model does not exist when the number of genes is larger than the number of time points. Indeed, Rangel's program (LDS_ToolBox) was trapped in ill-posed solutions for all of the 100 trial run with the different starting values of parameters and state dimensions.

Beal's Variational Bayes algorithm (VBSSM) succeeded in easing of over-parameterization of the input-driven model with the Bayesian regularization. For Beal's method and our proposed method, the dimension of state space was preset by $k = 4$. Figure 2 shows the two gene–gene interaction matrices which were estimated by Beal's method (left) and our method (middle). The 64 genes shown are composed of the eight modules where expression patterns of the genes in the same module are clearly aggregated as shown in the right panel of Figure 2. Each interaction matrix was divided into 8×8 blocks corresponding to the eight transcriptional modules observed. It can be seen from Figure 2 that the gene–gene interaction matrix estimated by the proposed method shows mosaic-like patterns across the 8×8 blocks. This observation indicates a statistical nature of our method which implicitly assumes that a set of co-expressed genes belongs to the same regulatory module. On the other hand, for the gene–gene interaction matrix estimated by Beal's method, the mosaic patterns are unclear. Note that for the both methods, we incorporated no external biological information into the network inferences. Thereby, it is hardly possible that any statistical methods can distinguish differences in the regulatory relationships of genes showing the same expression pattern from observed expression profiles only. In this regard, our method provides a natural consequence.

We also point out difference of the internal modules identified by the both methods in terms of biological aspect. For Beal's method, we computed Moore–Penrose inverse H^+ of the observation matrix H in (4), and constructed the 8 ($=2k$) modules, where the genes listed at each module were selected in the following way: the j-th gene is assigned to the i-th positive module M_{i+} or the i-th negative module M_{i-} if the (i, j)-th element in $(H^+)_{ij}$ is ranked in the highest or lowest 100, respectively. For our method, the procedure of selecting genes has been shown in the previous subsection. Then, we proceeded to the significant analysis of the GO terms by using GO::Term Finder (Boyle *et al.*, 2004). Supplementary Table 1 summarizes the over-represented GO categories for the eight identified modules which were identified by Beal's method and our method, respectively. As was discussed in previous, the eight modules identified by our method were converted into the statistically significant molecular functions of GO terms under 1% acceptable level of significance. In contrast, for Beal's method, we could not clarify a clear link between the identified modules and GO terms. This aspect is one of the most distinctive features between the canonical state space model and

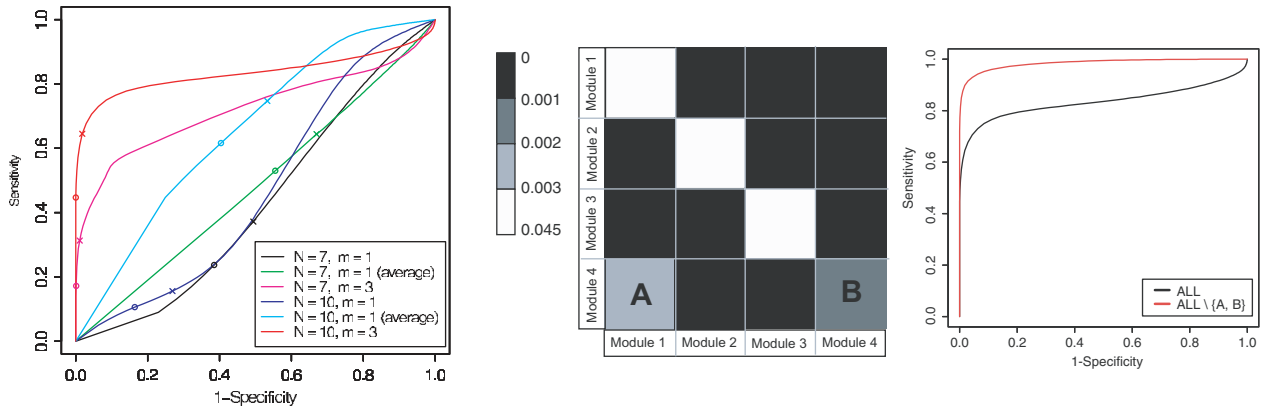

Fig. 3. (Left) ROC curves computed by training the state space model with the six synthetic datasets (a)–(f). (Middle) Heatmap of the autoregressive coefficient matrix Ψ (1000×1000) of the true model. Absolute values of the coefficients are depicted by gray scale image. The coefficients of the two regions, A and B, which correspond to the interactions from the module 1 to 4 and from the module 4 to 4, lie in the regions very close to zero. (Right) The black line shows the ROC curve computed by using the training datasets with $(N, m) = (10, 3)$. Red line is the ROC curve which were computed by removing the small coefficients in A and B from the true interactions.

the input-driven type model. Possibly, the eight modules identified by Beal's method correspond to the unobserved regulatory factors which cannot be measured by gene expression profiles.

4 MODELLING REPLICATED TIME-COURSE DATA

Use of the state space model provides us a way to analyse high-dimensional time-course data by exploring the aggregation of gene expression profiles and the temporal gene networks at the module level, simultaneously. However, although the proposed approach has the potential to construct large scale gene networks, its applicability is limited by when the length of time series is exceedingly short, e.g. less than 10. To overcome such a limitation, one possible solution might be to incorporate the replicates of time-course gene expression profiles into the parameter estimations. Currently, it has become common place to repeat time-course experiments multiple times in order to assess the reproducibility of data. Below, we will discuss the importance of incorporating the replicate data into the parameter estimation and extend the state space model to deal with the replicated measurements.

Let $y_n^{(l)} \in \mathcal{R}^p$ and $x_n^{(l)} \in \mathcal{R}^k$ be the gene expression vector which is measured by the l-th replicate and the corresponding hidden state vector at time n, respectively. The total number of replicates is denoted by m ($l = 1, \ldots, m$). Here, we assume that each of the replicated time-courses is i.i.d. according to

$$y_n^{(l)} = Hx_n^{(l)} + w_n^{(l)}, \quad n \in \mathcal{N}_{\text{obs}},$$
$$x_n^{(l)} = Fx_{n-1}^{(l)} + v_n^{(l)}, \quad n \in \mathcal{N}.$$

Given this generative model, the parameter estimation amounts to maximizing the likelihood function $l(\theta)$ over θ:

$$l(\theta) := \sum_{l=1}^{m} \sum_{n=1}^{N} I(n \in \mathcal{N}_{\text{obs}}) \log p(y_n^{(l)} | Y_{n-1}^{(l)}),$$

where $Y_n^{(l)} \equiv \{y_1^{(l)}, \ldots, y_n^{(l)}\}$ and $p(y_1^{(l)} | Y_0^{(l)}) \equiv p(y_1^{(l)})$. The modified EM algorithm for the maximum likelihood estimation is presented in Appendix 1.

To evaluate the predictive power of the proposed method, we conducted numerical experiments using synthetic data. Under the number of genes $p = 1000$, we generated the synthetic time-course data for $n = 1, \ldots, 10$ with the three replicates ($m = 3$) from the state space model as follows:

$$H = \begin{pmatrix} 1_{250} & & & \\ & 1_{250} & & \\ & & 1_{250} & \\ & & & 1_{250} \end{pmatrix}, F = \begin{pmatrix} 1.03 & & & \\ & 1.0 & & \\ & & 0.8 & \\ 0.5 & & & -0.3 \end{pmatrix},$$

$R = (0.1)^2 I$ and $\mu_0 = (10, 10, 10, 10)^T$ where $1_q \in \mathcal{R}^q$ denotes the vector of one. Supplementary Figure 1 shows the true module network represented by the F and the generated time-courses of the four state variables exhibiting the upward trend, random walk, downward trend and oscillated series, respectively.

Under the above setting, we learned the artificial gene network by using the six training datasets: (a) $(N, m) = (7, 1)$ (one of the three replicates) (b) $(N, m) = (7, 1)$ (averaged time-course of the three replicates) (c) $(N, m) = (7, 3)$ (d) $(N, m) = (10, 1)$ (one of the three replicates) (e) $(N, m) = (10, 1)$ (averaged time-course of the three replicates) and (f) $(N, m) = (10, 3)$. In the creation of the training datasets, we first generated the artificial gene expression profiles of length 10 with three times replication from the true model. Among the total 10 time points, we used the first 7 and 10 time points as the training datasets for (a)–(c) and (d)–(f), respectively. The single time-course data in (a) and (d) were chosen arbitrary from the three replicated time courses in (c) and (f), respectively. The (b) and (e) were created by averaging the gene expression values over the three replicates.

For each synthetic dataset, we performed the maximum likelihood estimations under $k = 4$ and the permutation significance tests for evaluating the gene–gene connectivity. The right panel in Figure 3 shows the ROC (Receiver Operating

Characteristic) curves in which the gene networks were constructed under a variety of acceptable significance levels. For the gene networks constructed using the single time-courses datasets, e.g. (a) and (d), the false-positive and the false-negative rates were greater than 50% (random choice) across most of the significance levels as the ROC curves (black and blue) were mostly present in the area under the 45 line. On the other hand, the predictive accuracy of the network estimations was considerably improved by using the replicates. For example, under the 10% significance level, the true-positive and the true-negative rates of $(N, m) = (10, 3)$ were 64.5% and 98.2%, respectively. This result indicates an importance of using replicated measurements of time-course data.

Here, we should remark the observed fact that the true-positive rates (64.5%) are much smaller than the true-negative rates (98.2%). The middle panel in Figure 3 shows the heatmap display of the true autoregressive coefficient matrix Φ of the 1000 genes which consist of the four regulatory modules. From this, we see that a part of the autoregressive coefficients lie in the region very close to zero (see A and B in the middle panel of Figure 3). Obviously, the power of tests in the identification of such a weak connection turns down, and increase in the false positive rate is caused by the large number of small auto-regressive coefficients. Indeed, if such weak connections were removed from the computation of ROC curve, the asymmetry in the true-positive and the true-negative rates would disappear (right panel in Figure 3).

5 APPLICATION TO GENE EXPRESSION PROFILES OF APOPTOSIS-INDUCED HUVECS

5.1 Data

Affara *et al*. (2007) studied transcriptome of HUVECs with the time-course gene expression data which were created using CodeLink 20k arrays. Based on a previous study that showed endothelial cells undergo a program of transcriptional change as they prepared to die by apoptosis (Johnson *et al*., 2003) they intended to understand the dynamics of gene regulations during apoptosis in the context of development and remodelling of blood vessels. mRNAs were prepared at 0.5, 1.5, 3, 6, 9, 12, 24 h after the induction of apoptosis by growth factor deprivation. The experiments were repeated independently three times ($m = 3$). Among approximately 20 000 genes, we focused on 1048 genes. These genes are comprised of 48 genes known to play important roles in apoptosis and blood vessel development (Carmeliet, 2000; Gerver *et al*., 1999) and 1000 genes giving the highest coefficients of variation. Appendix 2 provides the details for data preprocessing, including gene selection and normalization. After converting the real time set {0.5 h, 1.5 h, 3 h, 6 h, 9 h, 12 h, 24 h} into $\mathcal{N}_{\text{obs}} = \{1, 3, 6, 12, 18, 24, 48\}$ where the entire time points are defined by $\mathcal{N} = \{1, 2, \ldots, 47, 48\}$, we proceeded to the time series analysis.

5.2 Model selection

Unfortunately, we could not determine the state dimension based on the Bayesian information criterion (BIC) since the BIC curves were monotone decreasing as increasing of state dimension. In general, as the number of samples is much smaller than the dimension of data, the BIC curves tend to be monotone decreasing. Possibly, this is due to the fact that as the state dimension is taken to be large, the parameter estimation algorithm tends to be trapped in ill-posed solutions, i.e. occurrence of over-leaning. This aspect will be further discussed in Section 6.

Absence of a reasonable model selection criterion may be a limitation. In order to address such an indeterminacy of state dimension, we present a heuristic approach based on the the singular value decomposition of the projection matrix D. Recall that our method seeks the direction of projection such that the projected data $x_n = DR^{-1/2}(y_n - w_n)$, $n \in \mathcal{N}_{\text{obs}}$, are likely to follow the first-order Markov process in (3). During this process, the gene expression patterns relevant to the temporal structure of data are aggregated into the k coordinate systems of the state space $\mathcal{R}^k$. Let r ($\leq k$) be the rank of D which is equal to the number of non-zero singular values of the projection matrix. If $r < k$, the k rows of D are linearly-dependent, thereby yielding rank deficiency of the row space. To avoid such an over-parameterization, we may chose the state dimension $k = k^*$ such that all singular values of D lie in the region far from zero.

In order to find such a state dimension, we propose a significant test based on the permutations of m replicated time-courses and the minimum singular values of the projection matrices. Let d_{ik} be the i-th singular values of the projection matrix under setting the state dimension to k. Then, the procedure is summarized as follows:

- Generate $2^m - 1$ datasets which consist of all possible combinations of m replicates except for the empty set.
- For each state dimension k, compute the minimum singular value $\min_{i \in \{1,\ldots,k\}} \log d_{ik}^{(j)}$ across $j = 1, \ldots, 2^m - 1$ datasets which are generated in the above step.
- For each state dimension k, perform the kernel density estimation of the minimum singular values by using $\min_{i \in \{1,\ldots,k\}} \log d_{ik}^{(j)}$, $j = 1, \ldots, 2^m - 1$. The estimated density is denoted by f_k.
- Calculate the α% confidential interval I_k based on the f_k.
- Find the minimum $k = k^*$ such that $I_{k+1} \cap I_k = \phi$.

We run this procedure with the expression profiles of HUVECs. The kernel density estimations were performed with the Gaussian kernel where the bandwidth of the kernel function was chosen based on the Silverman's method (Silverman, 1986). In the first step, we created the datasets of size seven which is equal to the number of all possible combinations of three replicates ($m = 3$). For $k \geq 5$, however, we only used the four datasets because we could find no solutions for the likelihood functions corresponding to single time-course data.

In Figure 4, the least singular values of the projection matrices are plotted across the state dimensions ranging from $k = 1$ to 7. Bold segments denote I_k for $k = 1, \ldots, 7$. This plot indicates that the least singular values abruptly turn to be small at $k = 5$. Furthermore, for $k = 5$ or more, the least singular values remain close to zero. Indeed, it held that $I_4 \cap I_5 = \phi$

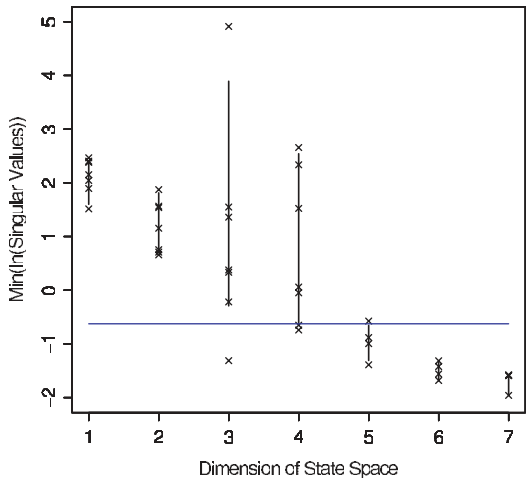

Fig. 4. The result of the rank-deficiency detection for the HUVEC dataset. The singular values were plotted after transformed into logarithmic-scale. Bold segments denote 50% confidential interval I_k for $k = 1, \ldots, 7$. The blue line represents the lower bound of I_4.

under $\alpha = 50\%$. According to this result, we decided to adopt $k = 4$. Besides, Affara *et al.* (2007) identified the eight internal groups of genes by the application of k-means clustering to the same dataset. This result also supports the use of $k = 4$ which implicitly implies the existing eight clusters of genes.

5.3 Result

Figure 5 shows the heatmap of the expression profiles of the most representative 50 genes to each of the identified eight modules. The genes in the same module were mostly co-expressed with each other. Particularly, the genes in the positive and the negative modules, $\mathcal{M}_{i+}$ and $\mathcal{M}_{i-}$, tend to exhibit the opposite expression patterns. For example, many genes listed at $\mathcal{M}_{3+}$ and $\mathcal{M}_{3-}$ are down/up-regulated from 6 to 9 h after the induction of apoptosis. Furthermore, the expression levels of the genes in $\mathcal{M}_{4+}$ and $\mathcal{M}_{4-}$ decrease/increase from 12 to 24 h.

A large number of the cell cycle-related genes were aggregated in $\mathcal{M}_{3\pm}$ and $\mathcal{M}_{4+}$. For example, $\mathcal{M}_{3+}$ and $\mathcal{M}_{4+}$ contained {CCNE1 CDCA7, CDC6, MCM(3, 4, 10), RBL2} and {CCNA2, CCNB1, CDC(2, 20, 25C, A1, A3), KNSL(1, 4, 6, 7), CENP(A, E, F, M)} in the most representative 50 genes, respectively. CDC2 (cyclin dependent kinase 1; CDK1) is known to bind cyclin A (CCNA2) and cyclin B (CCNB1) and regulates the cell cycle progression through G2 to M phase. It was found that many G2-M phase-genes were captured by the $\mathcal{M}_{4+}$. On the other hand, $\mathcal{M}_{3+}$ captured several G1-specific genes, e.g. CCNE1 (cyclin E1) which binds to cyclin dependent kinase 2 and regulates the cell cycle progression during G1 phase. $\mathcal{M}_{3-}$ contained cyclin-dependent kinase inhibitor 1C (CDKN1C; $P57^{KIP2}$) which binds to CDK2-cyclin E complex and inhibits progression from G1 to S phase. Temporal expression patterns of these cell cycle-related genes are shown in Supplementary Figure 2(1). While genes in $\mathcal{M}_{3+}$ and $\mathcal{M}_{4+}$ exhibit downward trend across the entire time intervals, expression levels of CDKN1C in $\mathcal{M}_{3-}$ are monotone increasing after 1.5 h. Possibly, these observations suggest that aberrant overexpression of G1-cyclin-dependent kinase inhibitor, i.e. CDKN1C, causes successive down-regulations of the G1-specific genes in $\mathcal{M}_{3+}$, e.g. cyclin E1, and is implicated in cell cycle arrest in G1.

Besides, $\mathcal{M}_{4-}$ mainly included the genes involved in immune and inflammatory responses, e.g. IFI35, IFI78(MX1), IFIT4, IFIT5, IFITM1, IRF7, STAT1, BAL, CRLF1, and also, ubiquitin-proteasome system, PSMB9, UBE2L6, USP18. The ubiquitin-proteasome system is essential at several stages during NF-κB-inducible inflammatory responses. NF-κB is a transcription factor which resides in the cytoplasm in inactive form, complexed to members of a family of inhibitory proteins referred to as IκB. Activating signals (e.g. binding of TNF-α to its receptor) cause phosphorylation of IκB kinase. This triggers the degradation of IκB through the ubiquitin-proteasome system, and then the free NF-κB can translocate to the nucleus and activate transcription of many genes involved in the inflammatory responses. As shown in Supplementary Figure 2(4), these inflammatory response-inducible genes in $\mathcal{M}_{4-}$ are abruptly activated from 12 to 24 h after consecutive up-regulation of cell-cycle inhibitory factors which possibly implicate arrest of cell cycle progression.

To understand the estimated temporal structure of the gene regulations involved in the arrest of cell cycle and the activation of immune and inflammatory responses, we constructed the module-based gene networks from the estimated coefficient matrix $\hat{\Psi}$ by setting the acceptable level of significance to 5%. Figure 5 highlights the identified module-based gene network where the genes shown were ranked in the most representative 50 genes for each module, and selected arbitrary.

Here, we focus on the estimated regulatory role of TRAF1 (tumor necrosis factor-associated factor 1) which is an important adapter protein involved in TNF (tumor necrosis factor)-mediated signalling pathway, leading to the activation of NF-κB, apoptosis, mitogen activated protein kinases (MAPK) cascades (Aggarwal, 2000). Under 5% significance level, TRAF1 was identified as the most highly connected gene (hub gene). The estimated network of TRAF1 is summarized as follows: As shown in Supplementary Figure 2(3), TRAF1 is consecutively up-regulated by 12 h after the apoptosis induction, and then, down-regulated through 12–24 h. According to the estimated network in Figure 5, TRAF1 and interleukin 8 (IL8) which is also an important inflammatory mediator are connected by the positive edges, indicating these two genes are involved in the same pathway. Indeed, Keifer *et al.* (2001) reported that transcriptions of either TRAF1 and IL8 are regulated by the cytokine-induced expression of NFκB. Interestingly, while IL8 shows the ongoing upward-trend during 1.5–24 h, TRAF1 is down-regulated after 12 h. Schwenzer (1999) discovered the positive feed-forward regulation of TRAF1 in the TNF-inducible NFκB pathway due to the observation of binding of NFκB to three putative binding sites within the human TRAF genes. The observed expression pattern indicates presence of inhibitory factors of TRAF1 that suppress this positive feedback loop between 12 and 24 h.

$\mathcal{M}_{1+}$ captured several important inhibitors of MAPK signalling pathways, e.g. dual specificity phosphatase (DUSP2, DUSP6), regulator of G protein signalling (RGS4)

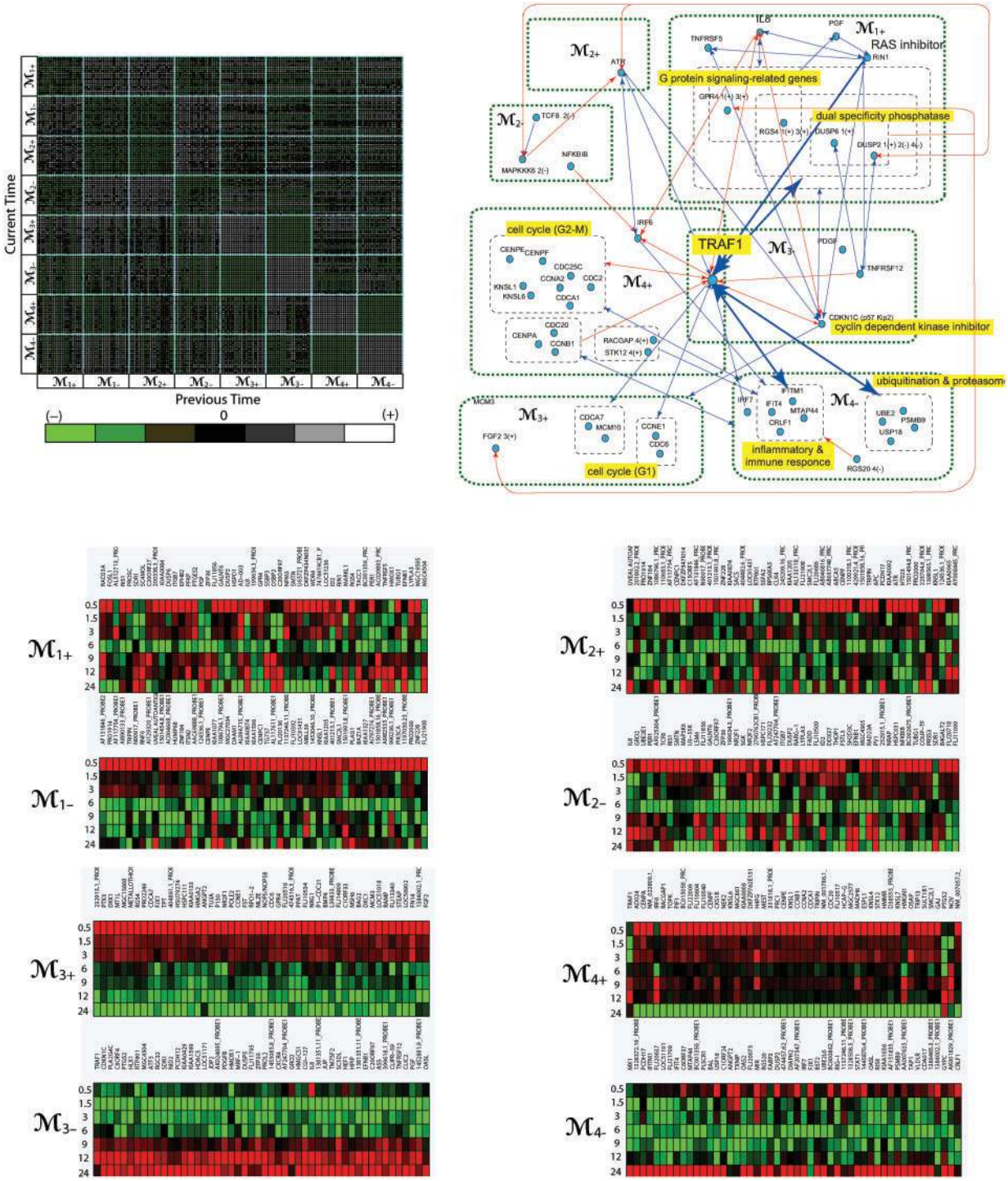

Fig. 5. Summary of the gene expression analysis of HUVECs undergoing growth factor deprivation-induced apoptosis: (Upper left) Heatmap representation of the estimated coefficient matrix $\hat{\Psi}$ which was divided into 8×8 blocks corresponding to the eight transcriptional modules identified. The genes in the *i*-th positive and negative modules were selected in the following way; the *j*-th gene was assigned to the *i*-th positive or the *i*-th negative module if the (i, j)-th element of the estimated projection matrix $(D)_{ij}$ was ranked in the highest or lowest 20, respectively. (Upper right) Gene network constructed under the acceptable level of significance 5% (right). The genes are classified and surrounded by the green-dashed lines according to the attributed modules. Furthermore, the genes involved in the common function are also classified, e.g. dual specificity phosphatase, G protein signalling-related genes, cell cycle. Directed edges with the positive or negative value are colored by red or blue. (Bottom) Gene expression patterns of the transcriptional modules identified.

and Ras inhibitor (RIN1). Here, we focus on the identified negative edges between TRAF1 and these inhibitory factors. Supplementary Figure 2 shows inversely correlated expression patterns between expression levels of TRAF1 and these inhibitory factors. According to Schwenzer (1999), the members of TRAF family are involved in activation of NFκB and c-Jun N-terminal kinase (JNK). In addition, several papers have suggested presence of cross-talk between MAPK and TNF-mediated signal cascades (Aggarwal, 2000; Han *et al.*, 1999). DUSPs are the inhibitors of MAPK signaling pathways by the dephosphorylation of MAPK molecules, e.g. p38, JNK. Besides, Ras is a small GTPase which activates p38 MAPK cascade as a MAPKKK, and possibly, this pathway is suppressed by the expression of its inhibitory factor RIN1. Regulator of G protein signaling (RGS4) also inhibits activity of Ras through the GTPase-acceleration that rapidly switches off G protein-coupled receptor signaling pathways. The constructed gene network model suggests that these inhibitory factors possibly affect the suppression of TRAF1 during 12–24 h through the TRAF1-related cross-talk between MAPK and TNF-mediated signal pathways.

Note also that TRAF1 also connected to cyclin dependent kinase inhibitor CDKN1C with the positive edge. This indicates that TRAF1-mediated pathway is involved in the cell cycle arrest through the aberrant expression of CDKN1C. Recently, involvement of cell cycle arrest and TNF-mediated signaling pathway has been reported by several papers, e.g. Mukherji *et al.* (2006). Suppression of G1 cyclin (CCNE1) by CDKN1C was captured by the negative significant edge.

6 MODEL SELECTION

The use of the state space model is a promising approach for estimating large scale gene networks from the short time-courses of gene expression profiles that contain few data points. However, the problem of selecting the state dimension with the information criterion is unresolved. For instance, as mentioned in Section 5, when we analysed the time-course profiles of apoptosis-induced HUVECs, the BIC curve across $k = 1, \ldots, 10$ did not exhibit folds. Such a tendency becomes prominent when the length of the time-course is short. In Section 5, for a guidance of the determination of a reasonable state dimension, we suggested a statistical testing procedure based on the singular value decomposition of the computed projection matrices. However, in this context, using the information criterion may be more acceptable for some users of the proposed method. Therefore, we discuss here the applicability of the information criterion-based approach.

We show the performance of model selection using the information criterion along with the simulation studies. We generated a number of synthetic data according to the state space model under $k = 4$ with the parameter values described in Appendix 3. The number of genes was set to 1000. Appendix 3 shows the change in shape of the BIC curves across several pairs of the number of time points N and the replicates m; $(N, m) \in \{(7, 1), (7, 2), (7, 3), (10, 1), (10, 2), (15, 1)\}$. For each experiment, we generated the simulated time-course data five times and then depicted the box plots for the computed BICs. Among these combinations, the BICs corresponding to $(N, m) = (7, 1)$ and $(10, 1)$ were monotone decreasing. In the other cases, we could correctly identify the number of the internal modules. These observations suggested that at least more than two time-course experiments are required for selecting the state dimension based on the BIC when relatively few time points are available and the number of genes involved is of order 10^3.

7 CONCLUDING REMARKS

A major difficulty in time-course analysis of gene expression is caused by the large number of genes involved in the pathways. Due to technical and financial limitations, it is unlikely that gene expression data containing sufficient data points to infer large scale gene networks will be available in the near future. Therefore, the development of statistical technologies for analyzing exceedingly short time-course data are an important challenges that need to be taken up.

In this article, we have presented some promising approaches towards the statistical inference of large gene networks. One practical solution to overcome such a difficulty is to identify module-based gene networks by exploring existing transcriptional modules. To this end, the state space model was used. The method automatically identifies the temporal aggregations of the gene expression profiles and assembles them into large scale gene networks. We demonstrated the potential of the state space model.

We also remarked on its limitation ascribed to small numbers of time-course data points. According to numerical experiments, we found that the applicability of the state space model was limited to analysis of time-course data where 15–20 data points were collected and the number of genes is set to $p \approx 10^3$. One way to overcome such a limitation is to incorporate the replicates of time-course data into the parameter estimation process. To our knowledge, the importance of using the replicates in time-course gene expression analyses has been relatively overlooked so far. However, as we demonstrated in this article, the use of replicated time-course data enables us to achieve highly efficient estimation of gene networks. This is especially important for the majority of currently available time series microarray data in which the number of time points is relatively small.

ACKNOWLEDGEMENT

We really appreciate all of the comments and the suggestions from the two anonymous referees that improve the quality of our article considerably.

Conflict of Interest: none declared.

REFERENCES

Aggarwal,B.B. (2000) Tumor necrosis factors receptor associated signaling molecules and their role in activation of apoptosis, JNK and NF-κB. *Ann. Rheum. Dis.*, **59** (Suppl. I), i6–i16.

Affara,M. *et al.* (2007) Understanding endothelial cell apoptosis: What can the transcriptome glycome and proteome reveal? *Phil. Trans. Roy. Soc.*, **362**, 1469–1487.

Arbeitman,M. *et al*. (2002) Gene expression during the life cycle of Drosophila melanogaster. *Science*, **298**, 2270–2275.

Bansal,M. *et al*. (2006) Inference of gene regulatory networks and compound mode of action from time-course gene expression profiles. *Bioinformatics*, **22**, 815–822.

Baranzini,S.E. *et al*. (2005) Transcription-based prediction of response to IFNb using supervised computational methods. *PLoS Biology*, **3**, 166–176.

Beal,M.J. *et al*. (2005) A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics*, **21**, 349–356.

Boyle,E.I. *et al*. (2004) GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.

Carmeliet,P. (2000) Mechanisms of angiogenesis and arteriogenesis. *Nat. Med.*, **6** (1), 389–395.

Gardner,T.S. *et al*. (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, **102**, 102–105.

Gerver,H.P. *et al*. (1999) VEGF is required for growth and survival in neonatal mice. *Development*, **126**, 1149–1159.

Han,Y. *et al*. (1999) Tumor necrosis factor-alpha-inducible IkappaBalpha proteolysis mediated by cytosolic m-calpain. A mechanism parallel to the ubiquitin-proteasome pathway for nuclear factor-kappab activation. *J. Biol. Chem*., **274**, 787–794.

Imoto,S. *et al*. (2006) Computational strategy for discovering druggable gene networks from genome-wide RNA expression profiles. *Pacific Symp. Biocomput*., **11**, 559–571.

Johnson,N. *et al*. (2003) Endothelial cells preparing to die by apoptosis initiate a program of transcriptome and glycome regulation. *FASEB J*., **18**, 188–190.

Keifer,J.A. *et al*. (2001) Inhibition of NF-kappa B activity by thalidomide through suppression of IkappaB kinase activity. *J. Biol. Chem*., **276**, 22382–22387.

Kitagawa,G. and Gersch,W. (1996) *Smoothness priors analysis of time series*. Springer-Verlag, New York.

Lee,T. *et al*. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **798**, 799–804.

Li,Z. *et al*. (2006) Using a state-space model with hidden variables to infer transcription factor activities. *Bioinformatics*, **22**, 747–754.

Mukherji,M. *et al*. (2006) Genome-wide functional analysis of human cell-cycle regulators. *Proc. Natl Acad. Sci. USA*, **103**, 14819–14824.

Orlando,D.A. *et al*. (2007) A probabilistic model for cell cycle distributions in synchrony experiments. *Cell Cycle*, **6**, 478–488.

Rangel,C. *et al*. (2004) Modeling T-cell activation using gene expression profiling and state-space models. *Bioinformatics*, **20**, 1361–1372.

Schwenzer,R. (1999) The human tumor necrosis factor (TNF) receptor-associated factor 1 gene (TRAF1) is up-regulated by cytokines of the TNF ligand family and modulates TNF-induced activation of NF-kappaB and c-Jun N-terminal kinase. *J. Biol. Chem*., **274**, 19368–19374.

Segal,E. *et al*. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet*., **34**, 166–176.

Silverman,B.W. (1986) *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York.

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

Tamada,Y. *et al*. (2005) Identifying drug active pathways from gene networks estimated by gene expression data. *Genome Inform*., **16**, 182–191.

van Someran,E.P. *et al*. (2006) Least absolute regression network analysis of the murine osteoblast differentiation network. *Bioinformatics*, **22**, 477–484.

Wu,F.X. *et al*. (2004) Modeling gene expression from microarray expression data with state-space equations. *Pacific Symp. Biocomput*., **9**, 581–592.

Yamaguchi,R. *et al*. (2007) Finding module-based gene networks in time-course gene expression data with state space models. *IEEE Signal Processing Magazine*, **24**, 37–46.

Yoshida,R. *et al*. (2005) Estimating time-dependent gene networks from time series microarray data by dynamic linear models with Markov switching. *Proc. IEEE Comput. Syst. Bioinform. Conf*, pp. 289–298.