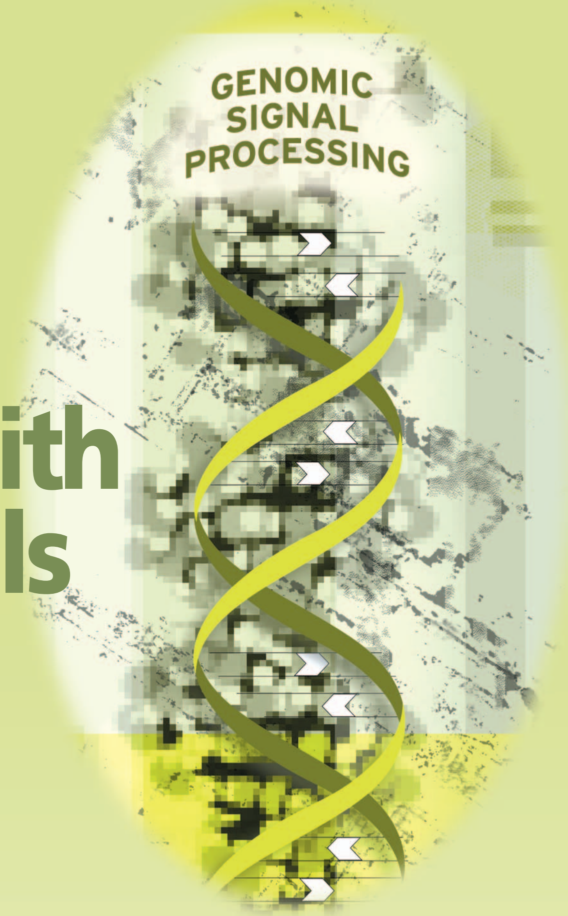
Rui Yamaguchi, Ryo Yoshida, Seiya Imoto, Tomoyuki Higuchi, and Satoru Miyano

Finding Module-Based Gene Networks with State-Space Models

Mining high-dimensional and short time-course gene expression data



© EYEWIRE

DNA microarray experiments allow us to examine expression levels of a large number of genes simultaneously. When experiments are sequentially carried out in a particular period (e.g., during cell cycles), a time course of which dimension is equal to the number of genes is achieved. To analyze such time-course gene expression data may lead us to further understanding of the mechanism to regulate expressions of many genes and responses of gene expressions to drugs, etc. Let y_n be an l-dimensional vector containing observed expression levels of l genes at the nth time step ($n = 1, \ldots, N$). A notable feature of time-course gene expression data is that the number of the time points N is usually much smaller than that of the genes l. In our previous experience [1], N is only 19, while l is about 800. It is not an unusual situation but a typical case in time-course microarray data analysis. One of the most significant challenges in bioinformatics is to establish a statistical method that can analyze such a high-dimensional and short-length time-course data.

A simple idea to model a series of time-course experiments $\{y_1, \ldots, y_N\}$ is to use a multivariate autoregressive model. However, in its application to a short time-course gene expression data, the conventional method of the parameter estimation might fail due to the overfitting. Linear Gaussian state-space models (SSMs, e.g., [2], [3]) provide us with a way to overcome such difficulties. SSMs have been used in a wide variety of applications with great success [2], [4]. There are several research studies using SSMs for analyzing time-course gene expression data with successful applications, e.g., [5], [6]. In SSMs, a sequence of the observation vectors $\{y_1, \ldots, y_N\}$ is modeled by assuming that at each time step y_n was generated from k-dimensional hidden-state variable vector denoted by x_n. Given a data set, the tasks to be addressed is the estimation of the parameters and sequence of the internal state vectors and also the determination of the state dimension. In the context of signal processing, the state vector x_n is often regarded as some unknown signals in y_n. On the other hand, in our problem, we can expect that x_n represents expression level of biological entities, i.e., gene modules, which are groups of the transcriptionally coexpressed genes having similar biological functions. In fact, it is a novel feature of this study to estimate the module networks by using SSMs.

In this study, we explore the following problems to analyze time-course gene expression data by SSMs. One is regarding methods for parameter estimation and determination of the dimension of the internal state variable. Although several methods have been applied, e.g., [1], [6], [7], there are few literature studies which with to compare them. Thus we give a brief review of the existing literature that use the SSM to analyze the gene expression time-course data. Another one is about identifiability of the model. If we simply estimate the

1053-5888/07/$25.00©2007IEEE

parameters of SSMs without any constraints for parameter space, they lack identifiability. To identify a system uniquely, it requires a specific algorithm to estimate the parameters with some constraints. For that purpose, we derive an identifiable form of SSMs and an algorithm for estimating parameters. The last one is to extract biological information by interpreting the estimated parameters, such as mechanism of gene regulations at the module level. For that one, we explore methods to extract further information using the estimated parameters, that is, we reconstruct a module network from time-course gene expression data.

[ALTHOUGH WE SHOWED SOME VARIANTS OF PARAMETER ESTIMATION METHODS FOR SSMS IN THE PREVIOUS SECTION, THERE IS A SUBSTANTIAL PROBLEM FOR SYSTEM IDENTIFICATION BY USING SSMS.]

SSMs

Let y_n be an l-dimensional vector containing observed expression levels of l genes at the nth time step where $n = 1, \ldots, N$. To model such time-course data, we use linear Gaussian state-space models that are often simply called SSMs. In SSMs, a sequence of the observation vectors $\{y_1, \ldots, y_N\}$ is modeled by assuming that at each time step y_n was generated from k-dimensional hidden state variable denoted by x_n. A basic model of state-space model is shown as follows:

$$x_n = Fx_{n-1} + v_n \quad \text{(system model)} \tag{1}$$

$$y_n = Hx_n + w_n \quad \text{(observation model)} \tag{2}$$

where F is the state transition matrix ($k \times k$ matrix), H is the observation matrix ($l \times k$ matrix), and $v_n \sim N_k(0_k, Q)$ and $w_n \sim N_l(0_l, R)$ are the system noise and the observation noise, respectively. The initial state vector x_0 is assumed to be a Gaussian random vector with mean vector μ_0 and covariance matrix Σ_0, i.e., $x_0 \sim N_k(\mu_0, \Sigma_0)$. In this problem, we need to estimate unknown parameters and state vector in the model. The dimension of state vector (k) is also unknown and thus needs to be determined for the optimal one.

According to existing literature, the number of modules is suggested to be much smaller than that of genes, e.g., [8], [9]. If x_n could represent state of modules, the estimated dimension k becomes smaller than the number of genes l. In that case, this method can be seen as a dimension reduction. Estimating the parameter vector $\theta = \{H, F, R, Q, \mu_0\}$ and the state vector x_n, we can expect to obtain insightful information of the biological system, for example magnitude of the effect of the modules on genes (gene-module interaction) from H matrix. We can also estimate networks between the modules by investigating F matrix.

SYSTEM IDENTIFICATION

In this section, we briefly review several parameter estimation methods and determination of the dimension of the state vector (i.e., number of modules) used in bioinformatics. Then we remark the lack of identifiability of SSM and obtain a form of SSM which retains identifiability of the system. We show an expectation-maximization (EM) algorithm [10] for estimating parameters of the model.

METHODS FOR PARAMETER ESTIMATION AND DIMENSION DETERMINATION FOR STATE VECTOR

From the existing literature using SSMs for time-course gene expression data, there are some variations: 1) how to estimate the model parameters and 2) how to determine the dimension of the internal state variable. To give a clear scope of how to use SSMs for such time-course data, we review these variants and compare results from them. Here we refer three papers [1], [6], [7] in which they used the same data and the same SSM (1) and (2) but different methods to estimate the parameters and to determine the dimension of x_n. The data that they used is a well-known public domain data, that is, gene expression data of the yeast (*Saccharomyces cerevisiae*) obtained during cell cycles [11].

In [6], the authors proposed a two-step manner; at the first step, the state variable x_n and the H is estimated by factor analysis and the dimension of x_n is determined by minimizing Bayesian information criterion (BIC) [12]. Then at the second step F is estimated using a least squares method.

In [7], considering a Bayesian estimation, the authors estimated F and H simultaneously using the variational Bayes method assuming prior distributions for the parameters [13]–[15]. The dimension of x_n is determined by maximizing the variational free energy.

In the last paper [1], the authors estimated F and H simultaneously as [7] but by using a maximum likelihood estimation with the EM algorithm [3]. They determined an optimal dimension of x_n based on the minimum BIC.

The resultant optimal dimensions of the state variable were different for these papers, in spite of the same data set and the same model, that is, five for [1] and [6] and two for [7]. We do not determine here which method is better for this kind of data. More systematic comparison would be needed by using both artificial and real data examples. In the following analysis, the maximum likelihood estimation by EM algorithm and BIC to determine the dimension of x_n [1] is used to estimate module networks.

IDENTIFIABLITY OF SSMs

Although we showed some variants of parameter estimation methods for SSMs in the previous section, there is a substantial problem for system identification by using SSMs. If we simply estimate parameters of an SSM without any constraints for the parameter space (e.g., [1], [6], [7]), it lacks the identifiability, that is, there exist infinite number of parameterizations yielding the same likelihood.

The lack of identifiability of an SSM is remarked as follows: Let Ψ ($k \times k$ matrix) be an arbitrary nonsingular matrix. The SSM can be replaced by

$$\Psi x_n = \Psi F \Psi^{-1} \Psi x_{n-1} + \Psi v_n \,, \tag{3}$$

$$y_n = H \Psi^{-1} \Psi x_n + w_n \,. \tag{4}$$

This implies the SSM is equivalent under arbitrary transformations, that is, $x_n \to \Psi x_n$, $H \to H\Psi^{-1}$, $F \to \Psi F \Psi^{-1}$, and $Q \to \Psi Q \Psi'$, where Q is the variance-covariance matrix of the system noise v_n. To overcome such overparameterization, we state the following proposition.

PROPOSITION 1

To avoid the lack of identifiability of SSM, it is sufficient to impose

- $Q = I_k$
- $H' R^{-1} H = \Lambda \equiv \mathrm{diag}\{\lambda_1, \ldots, \lambda_k\}$
- an arbitrary signed conditions for all elements in a particular $\eta_i = (\eta_{i1}, \ldots, \eta_{ik})'$ is assumed to be given on the parameter space, where $H' = (\eta_1, \ldots, \eta_l)$.

PROOF

Due to $Q = I_k$, it holds that $\Psi Q \Psi' = \Psi \Psi' = I_k$. Hence, the family of Ψ is restricted to be that of orthonormal matrices. Furthermore, since $H' R^{-1} H = \Lambda$, the transformed $\Psi H' R^{-1} H \Psi' = \Psi \Lambda \Psi'$ also must be a diagonal matrix. This implies that the Ψ must have 1 or -1 in its diagonal elements. Finally the third condition restricts $\Psi = I_k$ because the sign of all elements in η_i is fixed. We call the SSM with the above constraints SSM (CSSM). An EM algorithm for CSSMs is derived in the next section.

MAXIMUM LIKELIHOOD ESTIMATION WITH EM ALGORITHM

To obtain the maximum likelihood estimator of the parameters in CSSMs, we derive an EM algorithm. The EM algorithm for SSMs, which have no constraints, is already formulated by [3] and [16]. Proposition 1 suggests that the constraints for CSSMs relate only for Q, H, and R. Therefore, the modification points of EM algorithm for CSSMs from that for SSMs are only required in the parts relating to these parameters. More specifically, the update equations in M-step for those parameters only need to be modified. Thus, at first, we show the EM algorithm for SSMs as a reference. Then we describe the modification points for CSSMs.

Let $\{Y_N, X_N\}$ be the complete data, where $Y_N = \{y_1, \ldots, y_N\}$ is the set of observation data and $X_N = \{x_0, \ldots, x_N\}$ is the set of state variables (unobserved data). Then the joint likelihood for the complete data is given by

$$P(Y_N, X_N; \theta) = P(x_0) \prod_{n=1}^{N} P(x_n | x_{n-1}) P(y_n | x_n) \,, \tag{5}$$

where $\theta = \{H, F, R, Q, \mu_0\}$ is the parameter vector in the model. Note that Σ_0 is assumed to be known [3]. The probability densities $P(x_0)$, $P(x_n | x_{n-1})$, and $P(y_n | x_n)$ are given by the Gaussian distributions $N_k(\mu_0, \Sigma_0)$, $N_k(F x_{n-1}, Q)$, and $N_l(H x_n, R)$, respectively. Thus the joint log-likelihood of the complete data becomes

$$\begin{aligned} \log P(Y_N, X_N; \theta) = &- \frac{1}{2} \log |\Sigma_0| \\ &- \frac{1}{2} (x_0 - \mu_0)' \, \Sigma_0^{-1} \, (x_0 - \mu_0) \\ &- \frac{N}{2} \log |Q| - \frac{1}{2} \sum_{n=1}^{N} \left(x_n - F x_{n-1}\right)' \\ &\times Q^{-1} \left(x_n - F x_{n-1}\right) - \frac{N}{2} \log |R| \\ &- \frac{1}{2} \sum_{n=1}^{N} (y_n - H x_n)' \, R^{-1} \, (y_n - H x_n) \\ &- \frac{k + N\,(k + l)}{2} \log 2\pi \,. \end{aligned} \tag{6}$$

In the EM algorithm, the conditional expectation of the joint log-likelihood of the complete data

$$q(\theta \mid \theta^{\dagger}) = E\left[\log P(Y_N, X_N; \theta) \mid Y_N, \theta^{\dagger}\right] \tag{7}$$

is iteratively maximized with respect to θ until convergence, where $\theta^{\dagger}$ is the parameter vector obtained in the previous iteration. It is well known that the log-likelihood calculated with the $(i+1)$th iterative estimated parameters is larger than that with the ith iterative estimated parameters. An iteration of EM algorithm consists of two steps called the expectation step (E-step) and the maximization step (M-step), respectively. Each step in the $i + 1$th iteration is shown as follows: In E-step, $q(\theta \mid \theta_i)$ is calculated by

$$\begin{aligned} q(\theta \mid \theta_i) &= E\left[\log P(Y_N, X_N \mid \theta) \mid Y_N, \theta_i\right] \\ &= - \frac{1}{2} \log |\Sigma_0| \\ &\quad - \frac{1}{2} \mathrm{trace} \Big\{ \Sigma_0^{-1} \Big(V_{0|N} \\ &\quad + \left(x_{0|N} - \mu_0\right) \left(x_{0|N} - \mu_0\right)' \Big) \Big\} \\ &\quad - \frac{N}{2} \log |Q| \\ &\quad - \frac{1}{2} \mathrm{trace} \left\{ Q^{-1} \left(C - BF' - FB' + FAF' \right) \right\} \\ &\quad - \frac{N}{2} \log |R| \\ &\quad - \frac{1}{2} \mathrm{trace} \Big\{ R^{-1} \sum_{n=1}^{N} \Big[\left(y_n - H x_{n|N}\right) \\ &\quad \times \left(y_n - H x_{n|N}\right)' + H V_{n|N} H' \Big] \Big\} \\ &\quad - \frac{k + N\,(k + l)}{2} \log 2\pi \,, \end{aligned} \tag{8}$$

where $\theta_i = \{H(i), F(i), R(i), Q(i), \mu_0(i)\}$ is the parameter vector estimated in the ith iteration, and

$$A = \sum_{n=1}^{N} \left(V_{n-1|N} + x_{n-1|N} x'_{n-1|N} \right) , \tag{9}$$

$$B = \sum_{n=1}^{N} \left(V_{n,n-1|N} + x_{n|N} x'_{n-1|N} \right) , \tag{10}$$

$$C = \sum_{n=1}^{N} \left(V_{n|N} + x_{n|N} x'_{n|N} \right) . \tag{11}$$

In the above equation, the conventional Kalman smoothing estimators $x_{n|N}$, $V_{n|N}$, and $V_{n,n-1|N}$,

$$x_{n|N} = E\{x_n|Y_N\} \tag{12}$$

$$V_{n|N} = E\left\{(x_n - x_{n|N})(x_n - x_{n|N})'|Y_N\right\} \tag{13}$$

$$V_{n,n-1|N} = E\left\{(x_n - x_{n|N})(x_{n-1} - x_{n-1|N})'|Y_N\right\} \tag{14}$$

can be calculated by using the Kalman filter [17] and the fixed-interval smoother algorithm [4].

The log-likelihood

$$\log L(Y_N|\theta_i) = \log \int P(X_N, Y_N|\theta_i) dX_N \tag{15}$$

is also obtained as a byproduct of the Kalman filter.

In M-step, θ_i is updated to θ_{i+1} to be $\theta_{i+1} = \arg\max_\theta q(\theta|\theta_i)$ by $\partial_H q(\theta \mid \theta_i) = 0$, $\partial_F q(\theta \mid \theta_i) = 0$, $\partial_R q(\theta \mid \theta_i) = 0$, $\partial_Q q(\theta \mid \theta_i) = 0$, and $\partial_{\mu_0} q(\theta \mid \theta_i) = 0$. Thus $\theta_{i+1} = \{H(i+1), F(i+1), R(i+1), \ Q(i+1), \mu_0(i+1)\}$ is obtained by

$$H(i+1) = \left(\sum_{n=1}^{N} E\left\{ y_n x'_n \mid Y_N \right\} \right) C^{-1} \tag{16}$$

$$F(i+1) = BA^{-1} , \tag{17}$$

$$R(i+1) = N^{-1} \sum_{n=1}^{N} \Big[\left(y_n - Hx_{n|N} \right) \times \left(y_n - Hx_{n|N} \right)' + HV_{n|N}H' \Big] , \tag{18}$$

$$Q(i+1) = N^{-1}(C - BA^{-1}B') , \tag{19}$$

$$\mu_0(i+1) = x_{0|N} . \tag{20}$$

The procedure to obtain the maximum likelihood estimator of the parameter vector $\hat{\theta}$ is summarized as:

1) Select the initial values of $\theta_0 = \{H(0), F(0), R(0), Q(0), \mu_0(0)\}$ and some reasonable baseline levels of Σ_0. The conventional Kalman smoothing estimators $x_{n|N}$, $V_{n|N}$, $V_{n,n-1|N}$ (12)–(14) can be recursively calculated by Kalman filter and the fixed-interval smoother with the upper initial parameters.

2) Calculate the conditional expectation of the log likelihood with (8) (E-step).

3) Calculate (16)–(20) and obtain the next iterative estimated parameters that maximize conditional expectation of the log likelihood (M-step).

4) Insert estimated parameters to the state-space equations (1) and (2), and calculate the conventional Kalman smoothing estimators.

5) Repeat the upper procedures steps 2–4 until the log likelihood is converged.

The EM algorithm for the maximum-likelihood estimation may fail into a local maximum. Therefore the global maximum must be chosen by comparing results from several sets of the initial values. We note that this algorithm can be extended to deal with missing values in observation values naturally [3], [16].

MODIFICATION OF EM ALGORITHM FOR CSSM

In this section, we explain a modification of EM algorithm for estimating CSSMs. Regarding Q, (19) is removed due to the constraints of $Q = I_k$. Thus parameter vector for CSSMs becomes $\theta = \{H, F, R, \mu_0\}$. The modification for H and R are as follows: If we assume $R = rI_l$ $(r > 0)$, (18) is replaced by

$$r(i+1) = (Nl)^{-1} \sum_{n=1}^{N} \operatorname{trace} \Big[\left(y_n - Hx_{n|N} \right) \times \left(y_n - Hx_{n|N} \right)' + HV_{n|N}H' \Big] . \tag{21}$$

Then the condition $H'R^{-1}H = \Lambda$ can be written by $H'H/r = \Lambda$, which means $h'_p h_q = 0$ $(p \neq q, \ 1 \leq p, q \leq k)$, where h_p and h_q are the pth and the qth columns of H, respectively. Given column vectors $\{h_p | 1 \leq p, q \leq k, h'_p h_q = 0$ if $p \neq q\}$ of $H(i)$, a column-wise recursive update equation for h_p of $H(i+1)$ is derived by maximizing $q(\theta|\theta_i)$ subject to constraints $h'_p h_q = 0$ $(p \neq q)$ as follows:

$$h_p = \frac{1}{\sum_{n=1}^{N} \langle x_{pn}^2 \rangle} \left(\sum_{n=1}^{N} \langle y_n x_{pn} \rangle - \sum_{q \neq p} h_q \sum_{n=1}^{N} \langle x_{qn} x_{pn} \rangle - \sum_{q \neq p} \frac{1}{\|h_q\|^2} \sum_{n=1}^{N} h'_q \langle y_n x_{pn} \rangle h_q + \sum_{q \neq p} h_q \sum_{n=1}^{N} \langle x_{qn} x_{pn} \rangle \right) , \tag{22}$$

where $\langle \alpha \rangle = E(\alpha | Y_N, \theta_i)$ is a conditional expectation of α. If we set $p \leftarrow 1$, we can obtain $H(i+1)$ by recursing the calculation of (22) for k times with replacing h_p of $H(i)$ by h_p calculated by the equation and then setting $p \leftarrow p+1$ for each recursion. Thus for CSSMs, (16) needs to be replaced by the resultant $H(i+1)$.

CRITERION TO DETERMINE THE DIMENSION OF THE STATE VARIABLE

To determine an optimal dimension of the state vector (k), we use the BIC, which is given by

$$\mathrm{BIC}(k) = -2 \log L(Y_N|\hat{\theta}^{(k)}) + \lambda_p \log \nu_s , \tag{23}$$

where $\log L(Y_N|\hat{\theta}^{(k)})$ is the maximum marginal log-likelihood with the parameter vector $\hat{\theta}^{(k)}$ estimated in EM algorithm (15). The number of parameters to be estimated is denoted by λ_p. The number of samples is represented by ν_s. In this case, $\nu_s = N$: the number of time points. We determine the dimension of the state vectors that has the minimum BIC, i.e., $\hat{k} = \arg\min_k \mathrm{BIC}(k)$, as the optimal one.

IN THIS STUDY, TO EXTRACT INSIGHTFUL INFORMATION FROM TIME-COURSE GENE EXPRESSION DATA WE USED STATE-SPACE MODELS.

ESTIMATING MODULE NETWORKS

In this section, at first we explain a view of SSM for a biological system, that is, a representation of regulatory relationship of genes and gene-modules. Then we explain an algorithm to estimate a module network from the estimated parameter of SSM. Finally, it is applied to the real data set.

We note that in the following section we use CSSMs and assume $R = rI_l$ and $r > 0$.

TRANSCRIPTIONAL MODULES

We can expect that the internal state variable x_n may represent expression level of biological entities, i.e., gene "modules," which are groups of the coexpressed genes having similar biological functions, because of the form of observation model (2).

If we assume that the state vector, $x_n = [x_{1n}, \ldots, x_{kn}]'$, represents expression levels of k modules at time n, the observation model describes how expression levels of modules contribute to the expression levels of each gene. On the other hand, the observation model leads to another representation

$$x_n = Zy_n + \tilde{w}_n, \tag{24}$$

where $Z = UD^{-1}V'$, $\tilde{w}_n = -Zw_n$, and V, D, and U are matrices obtained from the singular value decomposition of H: $H = VDU'$. From this equation, we can measure the magnitude of contributions of genes to the expression levels of modules.

For the ith module, the equation can be written by $x_{in} = \sum_j z_{ij}y_{jn} + \tilde{w}_{in}$, where z_{ij} is the (i, j)th element of Z. If we gather genes which have large $|z_{ij}|$ for each module, we can obtain group of genes. In fact by considering signs of z_{ij}, we obtain two groups of genes for each modules, that is, group with positive z_{ij} and that for negative z_{ij}. That can be seen as a kind of clustering, however, it can allow a gene to belong to multiple modules. This property is much suitable for existing biological knowledge.

On the other hand, by the form of system model (1), we can see it as a model representing dynamic interactions between modules. Because it describes effect from x_{n-1} to x_n, the relationship can be seen as causal relationship. Thus we can expect to obtain module-regulatory networks by using the estimated parameter, that is, F.

SEARCHING METHODS OF THE MODULE NETWORKS

By considering the system model (1), the module network can be drawn by using $F = \{f_{ij}\}$. Here we denote the ith module by MDL_i regardless of time index. In that case, it can be done by connecting MDL_i and MDL_j with an arc (a directed edge) from MDL_j to MDL_i when $f_{ij} \neq 0$. Following a convention, we use a normal arc (→)when $f_{ij} > 0$, and an inhibitory arc (⊣)when $f_{ij} < 0$.

The problem to make a module network is to find nonzero elements in the estimated F.However, if we simply estimate $\theta = \{H, F, R, \mu_0\}$ in CSSM by the EM algorithm, it is highly possible that all elements in F become nonzero. It means that we obtain a graph in which any pair of nodes is connected regardless of the significance of every arc. Such a graph is not meaningful. Therefore we need to find significant arcs. To find them in a network, we put constraints on F keeping some elements zero and estimate parameters and obtain BIC for the constrained model. Then we compare BICs obtained for models having different constraints and select an optimal model of which BIC become the smallest. Finally, we make a network using F of the optimal model.

Regarding the parameter estimation method for SSMs with such constraints on the parameters, Wu et al. [18] derived an EM algorithm. Following them, the constraints on F are represented as

$$\Gamma \mathrm{vec}\, F = g \tag{25}$$

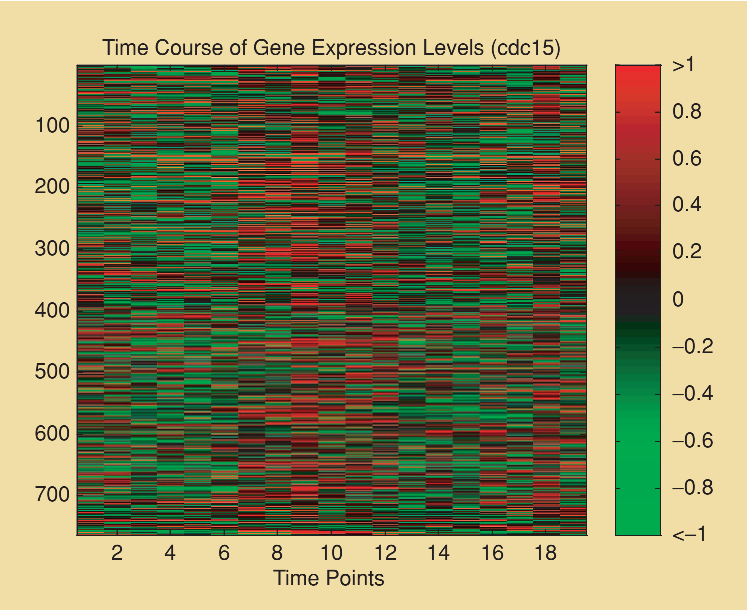
for known constant matrix Γ and vector g, where $\mathrm{vec}\, F$ denotes the vector formed from matrix F by stacking the columns of F beneath the one another. For example, let $F = \{f_{ij}\}$ be a 2×2 matrix and put constraints that $f_{12} = f_{21} = 0$.In that case, Γ and g become $\Gamma = [0_{2\times1} I_2 0_{2\times1}]$ and $g = [0_{2\times1}]$.For M-step in EM algorithm, an updating equation of F under (25) is obtained by

$$\begin{aligned}\mathrm{vec}F = &\,\mathrm{vec}(BA^{-1}) + (A^{-1} \otimes Q)\Gamma' \\ &\times \{\Gamma(A^{-1} \otimes Q)\Gamma'\}^{-1}\{g - \Gamma \mathrm{vec}(BA^{-1})\},\end{aligned} \tag{26}$$

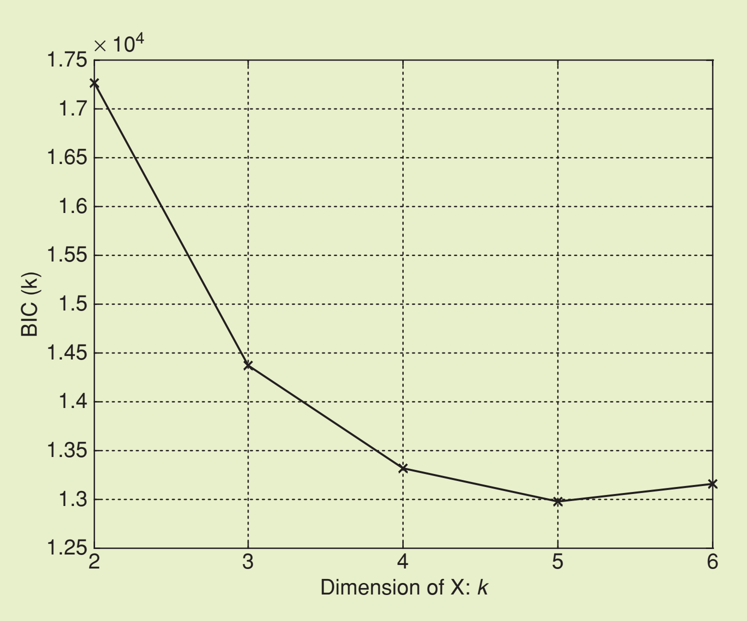
where $\otimes$ is the tensor product, A and B are given by (9) and (10) [18]. In our settings, $Q = I_k$ and $g = 0_{N_c\times1}$, where N_c is the number of constraints on F, that is, the number of elements to be zero in F. Thus (17) must be replaced with (26) for the EM algorithm.

Although we can obtain the best network by comparing BICs for models using the all possible constraints on F, such an exhaustive search is almost impossible especially when the dimension of x_n (k) becomes large. Hence, we search an optimal network by a greedy search algorithm to find an optimal constraints on F, in which we increase the number of constraints N_c one by one starting from $Nc = 0$ until $\widehat{\mathrm{BIC}}^{N_c} < \widehat{\mathrm{BIC}}^{N_c+1}$, where $\widehat{\mathrm{BIC}}^{N_c}$ is the smallest BIC among those from candidate models in which the number of zero elements in F is N_c. More specifically $\widehat{\mathrm{BIC}}^{N_c} = \min \mathrm{BIC}_i^{N_c}$, where $\mathrm{BIC}_i^{N_c}$ is BIC from the ith candidate model having N_c constraints $(i = 1, \ldots, i_{N_c})$. We denote F having N_c constraints by F^{N_c}. In the ith candidate model having $N_c + 1$ constraints, positions of N_c elements set to be zero in $F_i^{N_c+1}$ are the same as $\hat{F}^{N_c}$ for $\widehat{\text{BIC}}^{N_c}$. Then a position of the N_c+1th zero in $F_i^{N_c+1}$for the candidate model is selected from the part of elements having no constraints in $\hat{F}^{N_c}$. We note that to avoid to obtain meaningless solution, we do not put a constraint on an element which is the only element having no constraints in the row. Using the EM algorithm, we calculate $\text{BIC}_i^{N_c+1}$ for the ith candidate model having N_c+1 constraints. As the initial parameter of the algorithm we use $\theta_{0,i}^{N_c+1} = \{F_{i,N_c}^{N_c+1}, \hat{H}^{N_c}, \hat{R}^{N_c}, \hat{\mu}_0^{N_c}\}$, where $F_{i,N_c}^{N_c+1}$ is a matrix of which elements are the same as those of $\hat{F}^{N_c}$ but an additional zero is set as the N_c+1th constraint; and $\hat{F}^{N_c},\hat{H}^{N_c}$, $\hat{R}^{N_c}$, and $\hat{\mu}_0^{N_c}$ are those in $\hat{\theta}^{N_c}$ for $\widehat{\text{BIC}}^{N_c}$. We note that for CSSM having N_c constraints, $\widehat{\text{BIC}}^{N_c}$ is calculated by (23) with the number of parameters $\lambda_p = k\{l+(k+3)/2\}+1-N_c$, where l is the number of the genes, k is the number of modules.

For the greedy search, at first step ($N_c = 0$) we need to prepare the initial parameters (θ_0^0) for the EM algorithm and then calculate $\widehat{\text{BIC}}^0$. After that, we calculate $\widehat{\text{BIC}}^{N_c+1}$ increasing N_c one by one until the condition ($\widehat{\text{BIC}}^{N_c} < \widehat{\text{BIC}}^{N_c+1}$) is satisfied. If it is satisfied, we stop the search and denote the N_c by $\hat{N}_c$. Then we obtain the optimal parameter for the optimal constrained model for $\widehat{\text{BIC}}^{\hat{N}_c}$.



[FIG1] The time-course gene patterns of 800 cell cycle related genes in the cdc15 data.



[FIG2] $\widehat{\text{BIC}}^{\hat{N}_c}(k)$ versus the dimension of x_n: k.

TIME-COURSE MICROARRAY GENE EXPRESSION DATA

To estimate the optimal dimension of state vectors for real dynamical biological systems, we applied the above method to a publicly available cDNA time-course microarray gene expression data obtained for studying the cell-cycle regulated genes of budding yeast (*Saccaromyces cerevisiae*) [11]. The data set is available at http://cellcycle-www.stanford.edu.

Spellman et al. [11] identified 800 genes as the cell cycle regulated genes based on cluster analysis. They obtained the microarray data using samples from yeast cultures synchronized by three independent methods: α factor arrest, elutriation, and arrest of a *cdc15* temperature-sensitive mutant. Here we used the *cdc15*'s time-course data of the 800 genes. The time course of a gene includes evenly spaced 19 time points. The observation interval is 10 min. Figure 1 shows variations of expression levels of genes. For the analysis, we selected genes in which the number of missing points is less than 10. As a result, the number of selected genes is $l = 763$.

ANALYSIS AND RESULTS

To determine the optimal number of modules and estimate the module network from the real time-course gene expression data, at first, we applied the identifiable models (CSSMs) having different number of modules (k). We denote $\widehat{\text{BIC}}^{\hat{N}_c}$ for the model having k modules by $\widehat{\text{BIC}}^{\hat{N}_c}(k)$. Then we determined the optimal number of modules $\hat{k}$ by $\hat{k} = \arg\min_k \widehat{\text{BIC}}^{\hat{N}_c}(k)$.

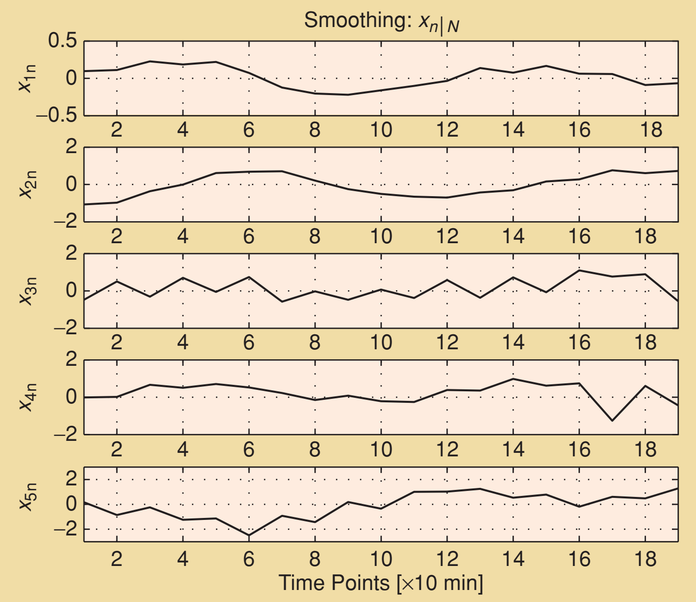
Figure 2 shows the profile of $\widehat{\text{BIC}}^{\hat{N}_c}(k)$ versus k, $(k = 1, \ldots, 6)$. To obtain $\widehat{\text{BIC}}^{\hat{N}_c}(k)$ for each k, 20 different initial parameters (θ_0^0) were used and the minimum one were accepted as $\widehat{\text{BIC}}^{\hat{N}_c}(k)$. The profile takes the minimum at $k = 5$. Thus we can decide that $\hat{k} = 5$ is the optimal number of dimension of the internal variable. This number is the same as the result of [1].

As a result, we obtained the optimal parameter vector $\hat{\theta} = \{\hat{H}, \hat{F}, \hat{R}, \hat{\mu}\}$corresponding to $\hat{k} = 5$ and the estimator of x_n. Figure 3 shows the time course of the smoothing estimator $\hat{x}_{n|N}$. We can observe different pattern of time series for each module. Figure 4 shows the time course of observed gene expression data y_n and its prediction by the optimal model $\hat{y}_{n|n-1}$ of cyclins having different peaks that exemplify typical cell-cycle related genes. It shows that the optimal model could make fairly good predictions for such different patterns of gene expressions. It may supports the validity of the estimation.

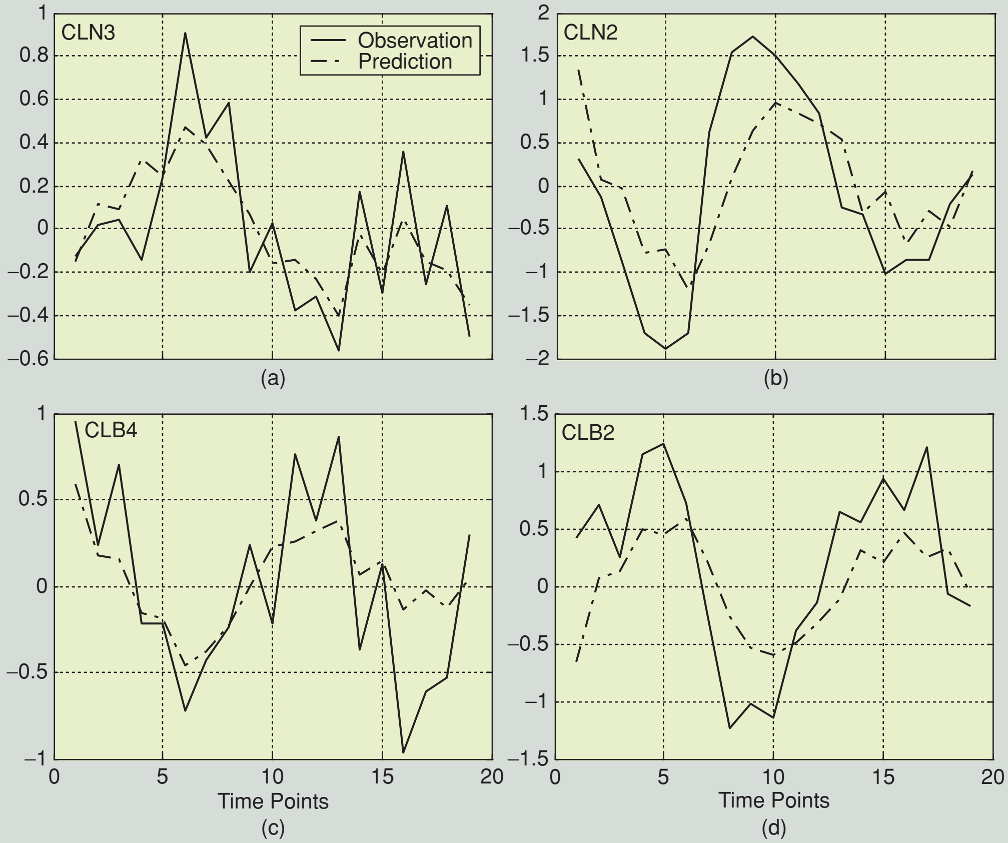
The estimated $\hat{F}$ is given by

$$\hat{F} = \begin{bmatrix} 0.478 & 0 & 0 & 0 & 0 \\ 0 & 0.579 & 0 & 0 & 0 \\ 1.561 & 0 & 0 & 0 & 0 \\ 1.408 & 0 & 0 & 0 & 0 \\ -3.666 & 0 & 0.869 & 0 & 0.667 \end{bmatrix}. \quad (27)$$

Using $\hat{F}$ and the system model (2), we can obtain a network representing the module-module interactions (Figure 5). The captions in colored boxes are explained in the next section.



[FIG3] Time course of the smoothing $x_{n|N}$.



[FIG4] The time-course of observed gene-expression y_n (solid line) and prediction $\hat{y}_{n|n-1}$ (dashed line) for cyclins with different peaks: (a) CLN3, (b) CLN2, (c) CLB4, and (d) CLB2.

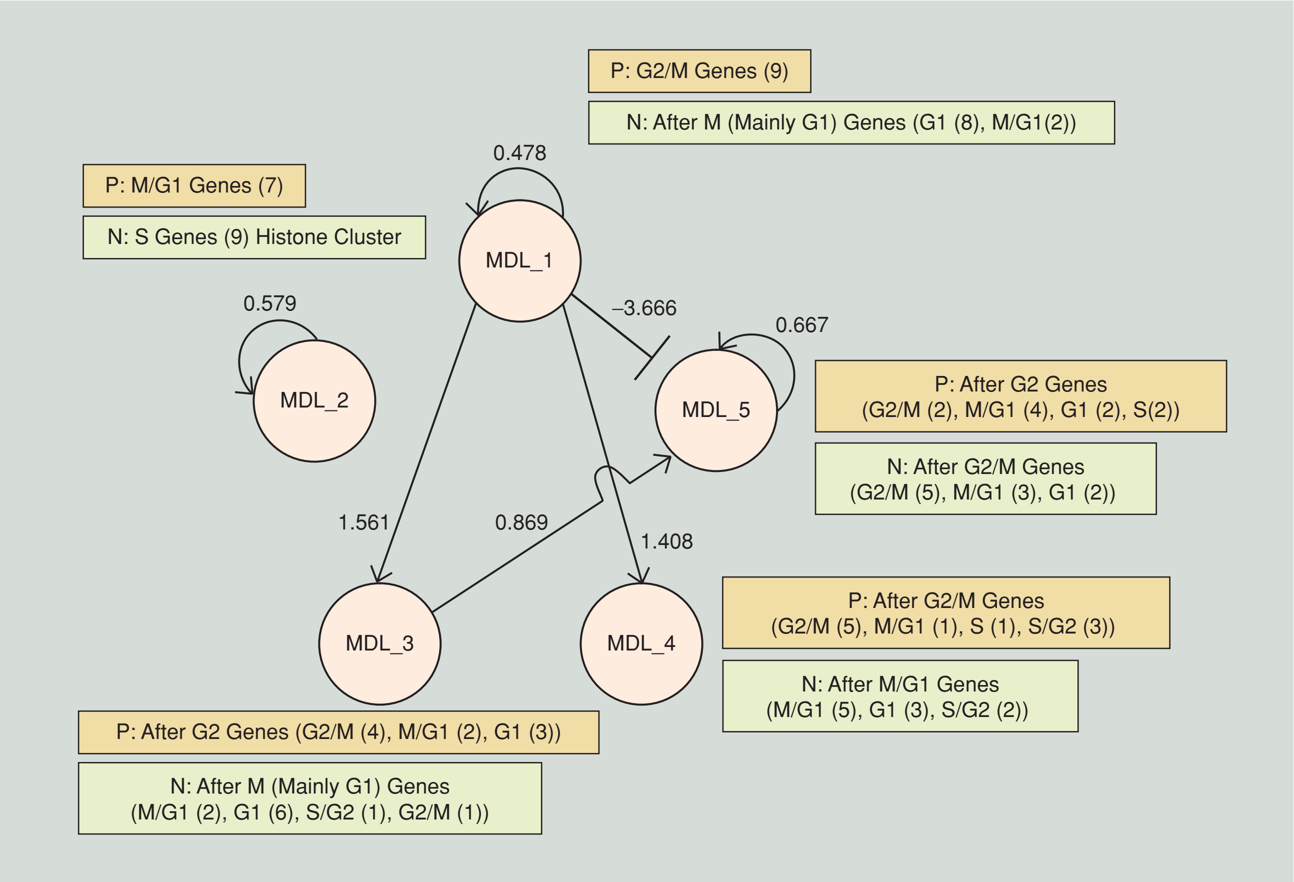
Figure 6 shows profiles of genes that are assigned as the positive member (shown in the left panels) or the negative member (shown in the right panels) for each module. Genes which have the largest (the smallest) ten contributions are shown as the positive (negative) member for each module. The gene profiles in the same panel looks similar. The profiles of the positive and negative member of genes in the same module show an antiphase relationship. Thus those genes might be successfully classified into modules.

WE SHOWED THAT THE STATE-SPACE MODELS ALLOW US TO EXTRACT USEFUL INFORMATION FROM HIGH-DIMENSIONAL TIME-COURSE GENE EXPRESSION DATA.

DISCUSSION AND CONCLUSION

In this study, to extract insightful information from time-course gene expression data we used SSMs. We first gave a brief survey of existing literature using SSMs applied to a data set with different parameter estimation methods and different criteria to determine the number of modules. We then derived SSMs with constraints (CSSMs) to overcome the lack of identifiability of unconstrained SSMs and showed an EM algorithm for CSSMs. We then explored methods to extract further information using estimated parameters, that is, a module network. To search an optimal network, we developed a greedy search algorithm based on a framework of statistical model selection using BIC. We applied the method to a real gene expression time-course data. As a result, we could determine the optimal number of modules and obtain the module network.

Here we discuss the resultant module network (Figure 5) and the genes classified into modules (Figure 6). Spellman et al. [11] assigned attributes (called peaks) for each genes in the data which represent the time when the gene expressions levels take the peak during cell cycle. According to the four phases ($G_1 \to S \to G_2 \to M$) in a cell cycle, for which the M phase is followed by the G_1 phase to start the next cycle, they gave one of the five peaks G_1, S, S/G_2, G_2/M, and M/G_1 for each gene. Using this information for genes in modules, we can characterize the modules. There are two colored boxes for each module in Figure 5. which summarize the feature of the modules. The yellow (blue) one with P (N) is for positive (negative) member of modules: phrases after the colon express the feature of the module. The number in parentheses after a peak name is the number of genes assigned in it. As shown in them, there is a tendency that genes with similar peaks accumulate in the same module

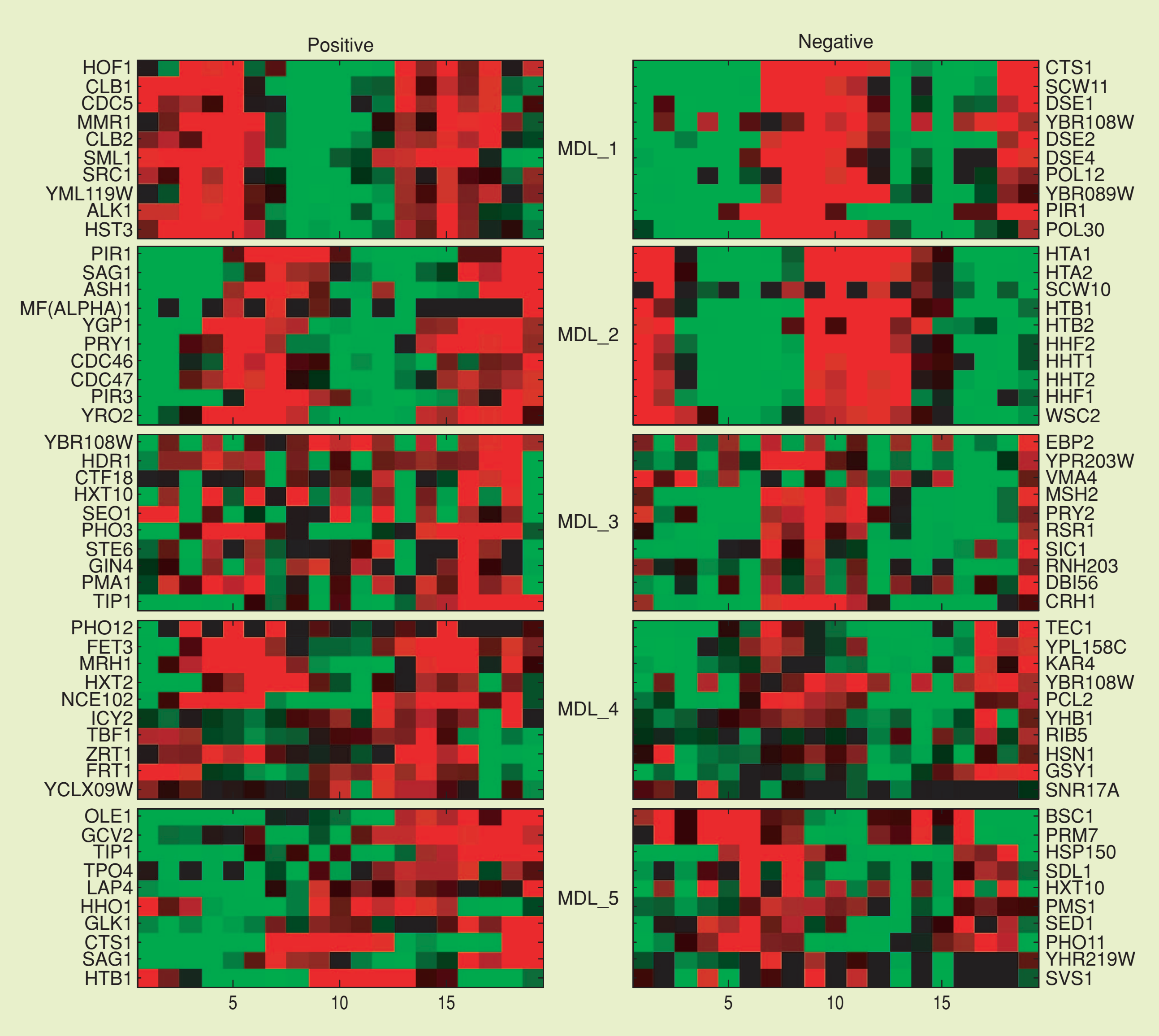

P: G2/M Genes (9)
N: After M (Mainly G1) Genes (G1 (8), M/G1(2))
0.478
P: M/G1 Genes (7)
N: S Genes (9) Histone Cluster
MDL_1
0.579
–3.666
0.667
MDL_2
MDL_5
P: After G2 Genes (G2/M (2), M/G1 (4), G1 (2), S(2))
N: After G2/M Genes (G2/M (5), M/G1 (3), G1 (2))
1.561
0.869
1.408
P: After G2/M Genes (G2/M (5), M/G1 (1), S (1), S/G2 (3))
MDL_3
MDL_4
N: After M/G1 Genes (M/G1 (5), G1 (3), S/G2 (2))
P: After G2 Genes (G2/M (4), M/G1 (2), G1 (3))
N: After M (Mainly G1) Genes (M/G1 (2), G1 (6), S/G2 (1), G2/M (1))

[FIG5] The estimated module network from $\hat{F}$.

as the same member. Considering such characteristics of the modules and directions and types of arcs between modules, we can advocate that the obtained network codes a partially consistent regulatory relationship between modules recalled from the time sequence of the phases in cell cycles. Furthermore, we found that the same, i.e., positive or negative, member in a module have similar type of genes or functionary related genes. An example is that almost all of the genes of the negative members in MDL_2 are histon. Another interesting finding is that the genes in the same module tend to form protein complexes, such as CLB1 and CLB2 in MDL_1, CDC46 and CDC47 in MDL_2 and so on. This feature is helpful to identify cellular functions of genes that are not biologically determined.

In this study, we showed that the SSMs allow us to extract useful information from high-dimensional time-course gene expression data. We succeeded in estimating biologically plausible module network of *Saccharomyces cerevisiae* cell cycle genes. Based on this information, the next research is focused on the estimation of gene-gene interaction as gene regulatory networks [19]–[21], that is a big challenge in bioinformatics.

[ONE OF THE MOST SIGNIFICANT CHALLENGES IN BIOINFORMATICS IS TO ESTABLISH A STATISTICAL METHOD THAT CAN ANALYZE SUCH A HIGH-DIMENSIONAL AND SHORT-LENGTH TIME-COURSE DATA.]



[FIG6] The time course of genes belonging to the modules. The left (right) panels show genes which have the positive (negative) contribution for each module. Each panel contains ten genes having larger (for the positive group) or smaller (for the negative group) contribution to a module.

ACKNOWLEDGMENTS

This research is partially supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan, Grant-in-Aid for Scientific Research (A), 17200020, 2005. This work is also partially supported by Kyushu University 21st Century COE Program, Development of Dynamic Mathematics with High Functionality, of the MEXT of Japan.

AUTHORS

Rui Yamaguchi (ruiy@ims.u-tokyo.ac.jp) is an assistant professor of the Laboratory of Biostatistics, Human Genome Center, Institute of Medical Science, University of Tokyo. He received his Ph.D. in science at the Kyushu University in 2003. His main research interest is study and applications of state-space models and Bayesian networks. He is especially interested in gene networks.

Ryo Yoshida (yoshidar@ims.u-tokyo.ac.jp) is an assistant professor of the Laboratory of DNA Information Analysis, Human Genome Center, Institute of Medical Science, University of Tokyo. He received the Ph.D. in statistical science at the Graduate University for Advanced Studies in 2004. Central in his research is to establish statistical methodology for cluster analysis, time series analysis, and machine learning theory.

Seiya Imoto (imoto@ims.u-tokyo.ac.jp)is an assistant professor with the Laboratory of DNA Information Analysis, Human Genome Center, Institute of Medical Science, University of Tokyo. He received the B.S., M.S., and Ph.D. degrees in mathematics from Kyushu University in 1996, 1998, and 2001, respectively. His current research interests cover statistical analysis of high dimensional data by Bayesian approach, DNA microarray gene expression data analysis, gene regulatory network analysis, and computational drug target discovery.

Tomoyuki Higuchi (higuchi@ism.ac.jp) iscurrently a vice director-general at the Institute of Statistical Mathematics (ISM), Research Organization of Information and Systems. He is also a professor of ISM and of the Graduate University for Advanced Studies. He received his Ph.D. in science at the University of Tokyo in 1989. His primary research interests are in Bayesian modeling of space-time data and data mining.

Satoru Miyano (miyano@ims.u-tokyo.ac.jp) is a professor of Human Genome Center, Institute of Medical Science, University of Tokyo. He received the B.S., M.S., and Ph.D. degrees in mathematics from Kyushu University, Japan, in 1977, 1979, and 1984, respectively. His research group is developing computational methods for inferring gene networks from microarray gene expression data and other biological data. Currently, his research group is intensively working for developing the gene network of human endothelial cell by knocking down hundreds of genes. With these technical achievements, his research direction is now heading toward a creation of systems pharmacology.

REFERENCES

[1] R. Yamaguchi and T. Higuchi, "State-space approach with the maximum likelihood principle to identify the system generating time-course gene expression data of yeast," *Int. J. Data Mining Bioinformatics*, vol. 1, no. 1, pp. 77–87, 2006.

[2] A.C. Harvey, *Forecasting, Structural Time Series Models and the Kalman Filter*. New York: Cambridge Univ. Press, 1989.

[3] R.H. Shumway and D.S. Stoffer, "An approach to time series smoothing and forecasting using the EM algorithm," *J. Time Series Anal.*, vol. 3, no. 4, pp. 253–264, 1982.

[4] G. Kitagawa and W. Gersch, *Smoothness Priors Analysis of Time Series*. New York: Springer-Verlag, 1996.

[5] C. Rangel, J. Angus, Z. Ghahramani, M. Lioumi, E. Sotheran, A. Gaiba, D.L. Wild, and F. Falciani, "Modelling T-cell activation using gene expression profiling and state space models," *Bioinformatics*, vol. 20, no. 9, pp. 1361–1372, 2004.

[6] F.X. Wu, W.J. Zhang, and A.J. Kusalic, "Modeling gene expression from microarray expression data with state-space equations," in *Proc. Pacific Symp. Biocomputing*, City, State, vol. 9, 2004, pp. 581–592.

[7] N. Yukinawa, J. Yoshimoto, S. Oba, and S. Ishii, "System identification of gene expression time-series based on a linear dynamical system model with variational Bayesian estimation," (in Japanese), *Inf. Process. Soc. Japan, Trans. Math. Modeling and Its Applicat.*, vol. 46, no. 10, pp. 57–65, 2005.

[8] Z. Bar-Joseph, G.K. Gerber, T.I. Lee, N.J. Rinaldi, J.Y. Yoo, F. Robert, D.B. Gordon, E. Fraenke, T.S. Jaakkola, R.A. Young, and D.K. Gifford, "Computational discovery of gene modules and regulatory networks," *Nat. Biotechnol.*, vol. 21, no. 11, pp. 1337–1342, 2003.

[9] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman, "Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data," *Nat. Genetics*, vol. 34, no. 2, pp. 166–176, 2003.

[10] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statistical Soc.,* ser. B, vol. 39, no. 1, pp. 1–38, 1977.

[11] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher, "Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization," *Mol. Biol. Cell*, vol. 9, no. 12. pp. 3273– 3297, 1998.

[12] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.

[13] H. Attias, "Inferring parameters and structure of latent variable models by variational Bayes," in *Proc. 15th Conf. Uncertainty Artificial Intelligence*, 1999, pp. 21–30.

[14] Z. Ghahramani and M.J. Beal, "Propagation algorithms for variational Bayesian learning," *Adv. Neural Inform. Process. Syst.*, vol. 13, no. 1, pp. 507–513, 2001.

[15] J. Yoshimoto, S. Ishii, and M. Sato, "System identification based on on-line variational Bayes method and its application to reinforcement learning," in *Proc. Artificial Neural Networks and Neural Information Processing (ICANN/ICNIP 2003)*, pp. 123–131, 2003.

[16] R.H. Shumway, "Dynamic mixed models for irregularly observed time series," *Resenhas-Reviews of the Institute of Mathematics and Statistics*, Univ. Sao Paulo, Brazil: Univ. San Paulo Press, vol. 4, no. 4, pp. 433–456, 2000.

17] R.E. Kalman, "A new approach to linear filtering and prediction problems," *Trans. Amer. Soc. Mech. Eng., J. Basic Eng.*, vol. 82, no. 1, pp. 35–45, 1960.

[18] L.S.-Y. Wu, J.S. Pai, and J.R.M. Hosking, "An algorithm for estimating parameters of state-space models," *Stat. Prob. Lett.*, vol. 28, no. 2, pp. 99–106, 1996.

[19] N. Friedman, M. Linial, I. Nachman, and D. Pe'er., "Using Bayesian network to analyze expression data," *J. Comp. Biol.*, vol. 7, no. 3-4, pp. 601–620, 2000.

[20] I. Shmulevich, E.R. Dougherty, S. Kim, and W. Zhang, "Probabilistic Boolean networks: A rule-based uncertainty model for gene regulatory networks," *Bioinformatics*, vol. 18, no. 2, pp. 261–274, 2002.

[21] S. Imoto, T. Higuchi, T. Goto, K. Tashiro, S. Kuhara, and S. Miyano, "Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks," *J. Bioinform. Comp. Biol.*, vol. 2, no. 1, pp. 77–98, 2004.