Front. Electr. Electron. Eng. 2012, 7(4): 357–366
DOI 10.1007/s11460-012-0207-x

RESEARCH ARTICLE

Wanwan TANG, Rui LI, Shao LI, Yanda LI

Co-regulated gene module detection for time series gene expression data

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2012

Abstract It is important to detect interaction effect of multiple genes during certain biological process. In this paper, we proposed, from systems biology perspective, the concept of co-regulated gene module, which consists of genes that are regulated by the same regulator(s). Given a time series gene expression data, a hidden Markov model-based Bayesian model was developed to calculate the likelihood of the observed data, assuming the co-regulated gene modules are known. We further developed a Gibbs sampling strategy that is integrated with reversible jump Markov chain Monte Carlo to obtain the posterior probabilities of the co-regulated gene modules. Simulation study validated the proposed method. When compared with two existing methods, the proposed approach significantly outperformed the conventional methods.

Keywords co-regulated gene module, Bayesian, hidden Markov model, Markov chain Monte Carlo

1 Introduction

Understanding cell responses to environmental stimuli is one of the central tasks for systems biology. Large scale gene expression profiling with microarray platforms is widely used to identify the responsive genes with significantly changed expressions in the induced process. But the gene-based method does not consider the interactions of multiple genes, which is generally believed important in most of the biological functions [1–4]. Identification of gene modules rather than independent genes can provide better understanding of the underlying regulation mechanisms.

Received July 2, 2012; accepted August 13, 2012

Wanwan TANG (✉), Rui LI, Shao LI, Yanda LI

MOE Key Laboratory of Bioinformatics, Bioinformatics Division, Tsinghua National Laboratory of Information Science and Technology / Department of Automation, Tsinghua University, Beijing 100084, China
E-mail: tww05@mails.tsinghua.edu.cn

Conventional methods for analyzing gene interactions focus on investigating genes that co-express in certain biological processes, such as hierarchical clustering [5], *K*-means [6], and self-organizing maps [7]. Graph-based approaches are widely used for the identification of co-expression modules [8–10]. In these methods, pairs of gene expression similarity are calculated to construct a network, in which nodes represent genes and two genes are connected if their similarity is above a threshold. The module identification is carried based on the network. The detected genes in the module tend to have high/low expressions simultaneously.

However, the statistics obtained from stability assumption that ignores the dynamic process of genes may result in inaccurate conclusion, as genes generally express in a highly dynamic way. Furthermore, which may be more fundamental, the genes that work together in certain process may not co-express with each other, thus the effort that is focused on co-expression pattern of genes may lose much power in detecting interactive genes, especially in time series data analysis in which the expression level of a certain gene could vary strongly as it may be regulated by certain factors that could change their statuses through the process.

To overcome these limitations, we propose in this paper an algorithm named CrgMODE (Co-regulated gene MOdule DEtection) to investigate the gene interactions in a certain biological process. Co-regulated Gene Module (CrGM) consists of genes that are regulated by the same factor(s). We first develop a Bayesian model that partitions the genes into different modules in which genes are supposed to be regulated by the same factor(s), which may be unknown. The expression process inside each module is modeled based on hidden Markov model (HMM) in which the regulator is represented by a hidden variable whose status may be changing during the biological process, while the expressions of random genes are modeled by Gaussian mixture model (GMM). To obtain the posterior probabilities of the modules (gene partition), we developed a Markov chain Monte Carlo (MCMC) sampling strategy

that integrates both Gibbs sampling and reversible jump Markov chain Monte Carlo (RJ-MCMC), in order to address the uncertainty in both marker partition and module number determination.

We systematically compare the proposed approach with two existing algorithms on simulated time series microarray data sets from four models. The results show that our method outperforms the other two algorithms.

2 Methods

2.1 Co-regulated Gene Module (CrGM)

Under specific stimuli, the transcription of downstream genes will be activated or repressed with certain patterns, which can be possibly identified in the time series microarray data. The variations of gene expressions provide important information of this dynamic biological process. Here, we propose the concept of CrGM in order to get better understanding of interaction effect of genes in this process. Genes in the same CrGM are regulated by the same regulator in the process, while those in different CrGMs are regulated by different regulators.

The relationship between genes and their regulators is shown in Fig. 1. Regulated genes are contained in CrGMs, and the unregulated ones are outside the modules. Each CrGM has a hidden regulator (which could be interpreted as the combinative effect of certain factors) and different regulators affect their own downstream genes during the dynamic process independently. The states of these hidden regulators are varying during the biological process with certain patterns. For the genes being regulated, we further define being singly regulated and being co-regulated. Being singly regulated means that the gene has a unique regulator only for the gene itself. In other words, the CrGM is composed of only one gene (Fig. 1, CrGM 2). Being co-regulated means that several genes share the same hidden regulator, which means the CrGM is composed of multiple genes (Fig. 1, CrGM 1). It is obvious that being singly regulated is a special case of being co-regulated where the number of regulated genes is one. We use the term "co-regulated" to denote the cases of both being singly regulated and being co-regulated.

The dynamic process inside each CrGM is modeled based on HMM. We use kth-order Markov chain to model the dynamic process of the hidden regulator:

$$P\Big(s(t)|s(t-1),s(t-2),\ldots,s(0)\Big) = P\Big(s(t)|s(t-1),s(t-2),\ldots,s(t-k)\Big), \quad (1)$$

where $s(t)$ is the state of the regulator at time point t.

The relationship between increments of the expression levels of genes and the state of the regulator in a CrGM is shown in Fig. 2. The state of the regulator in the present is determined by its former k states. Once the regulator state is fixed, the increments of the expression levels of these genes are generated by emission probabilities independently. Here emission probability refers to the probability of the increment of the expression level of a gene conditional on the state of its regulator. Generally the emission probabilities for different genes in the same CrGM could be quite different.

For genes that are not-regulated, we assume that there is no regulator for them or the regulator is a constant during the process we are observing. Therefore the increment is random and is assumed following i.i.d. GMM.

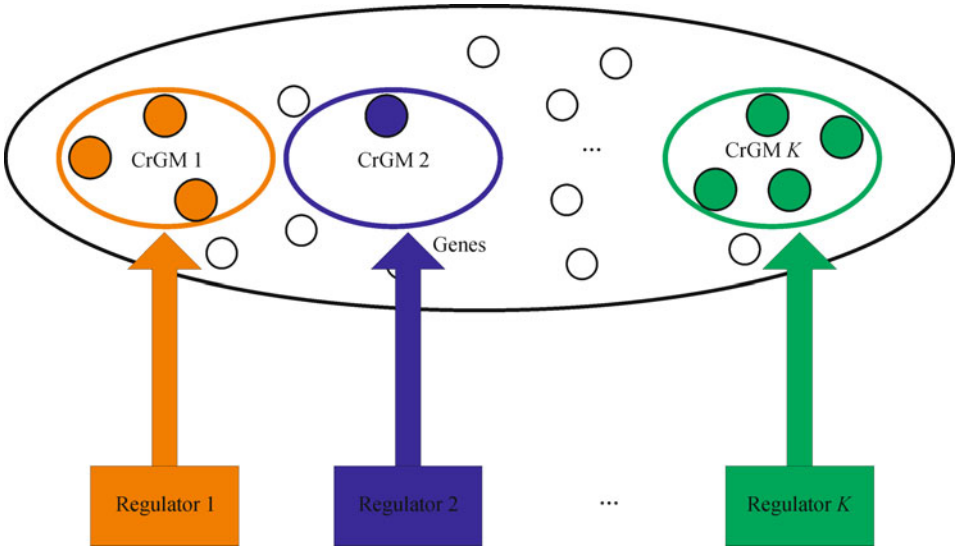

Fig. 1 Relationship between genes and their regulators. Regulated genes are contained in CrGMs and the unregulated ones are outside the modules. Each CrGM has a hidden regulator.

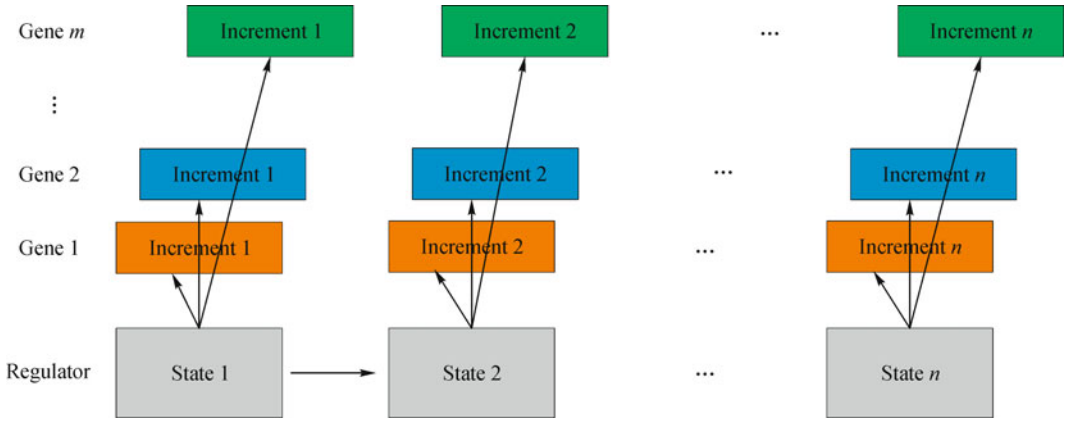

Fig. 2 Relationship between increments of genes and states of the regulator in a CrGM in a certain biological process.

2.2 Bayesian marker partition model

Suppose in a time series microarray data set, there are L genes with expression levels at $T+1$ time points. The expression of gene i at time t is represented as $G_i(t)$, $i=1,2,...,L$ and $t=0,1,...,T$. Let $d_i(t)$ be the increment of the expression level of a gene,

$$d_i(t)=\log\frac{G_i(t)}{G_i(t-1)}, \tag{2}$$

where $t=1,2,...,T$. Therefore, $d_i(t)>0$ means that the gene expression level increases during the period from time point $t-1$ to time point t, while $d_i(t)<0$ means that the expression level decreases. Let $D_i=\left(d_i(1),d_i(2),...,d_i(T)\right)$ record all the increments from time point 1 to time point T for the ith gene and $\boldsymbol{D}=(D_1,D_2,...,D_L)'$ for the whole observed data.

As we already defined in the last section, the L genes can be partitioned into $K+1$ modules $M_0,M_1,...,M_K$, where M_0 contains genes unregulated and M_1 to M_K are CrGMs.

Let I_i, $i=1,2,...,L$, be the indicator for the partition of the ith gene, and $\boldsymbol{I}=(I_1,I_2,...,I_L)'$ represent the partition for all the L genes. It is natural that I_i has $K+1$ possible values $0,1,...,K$. Let l_m be the number of genes partitioned into the mth module, $m=0,1,...,K$. Now we have $l_0+l_1+\cdots+l_K=L$. Let $\boldsymbol{D}_m=(D_{m1},D_{m2},...,D_{ml_m})'$ be the increments of the expression levels of genes that belong to the mth module, and $D_{mi}=\left(d_{mi}(1),d_{mi}(2),...,d_{mi}(T)\right)$ the increments for the ith gene in the mth module for time point 1 to time point T, $m=0,1,...,K$ and $i=1,2,...,l_m$. Let $S_m=\left(s_m(1),s_m(2),...,s_m(T)\right)'$ be the states of the hidden regulator in the mth module from time point 1 to time point T, and assume $s_m(t)$ has R possible values $1,2,...,R$, $m=1,2,...,K$. With these concepts, the problem of detecting groups of genes that are co-regulated in the dynamic biological process is equivalent to partitioning the genes into different CrGMs.

Module M_0 consists of genes that are not regulated. As discussed in the last section, we assume the increments of the expression levels of genes in M_0 follow i.i.d. GMM with R sub-populations. Therefore, the likelihood of the observed increments $\boldsymbol{D}_0$, given the partition $\boldsymbol{I}$ and the parameter $\boldsymbol{\Theta}$ of GMM can then be written as

$$p(\boldsymbol{D}_0|\boldsymbol{\Theta},\boldsymbol{I})=\prod_{i=1}^{l_0}\prod_{t=1}^{T}p\left(d_{0i}(t)|\Theta_i\right), \tag{3}$$

where $d_{0i}(t)$ denotes the increment of expression level of the ith gene in module M_0 at time point t and $\boldsymbol{\Theta}=(\Theta_1,\Theta_2,...,\Theta_{l_0})'$ the parameter matrix for Gaussian mixture distribution for the ith gene in M_0. $\Theta_i=(\boldsymbol{\alpha}_i,\boldsymbol{\mu}_i,\boldsymbol{\sigma}_i)$, where $\boldsymbol{\alpha}_i=(\alpha_{i1},\alpha_{i2},...,\alpha_{iR})$, $\boldsymbol{\mu}_i=(\mu_{i1},\mu_{i2},...,\mu_{iR})$, and $\boldsymbol{\sigma}_i=(\sigma_{i1},\sigma_{i2},...,\sigma_{iR})$, is used as the parameter for generating increment at every time point for the ith gene in M_0, from the GMM that is illustrated with Eq. (4), $i=1,2,...,l_0$:

$$p\left(d_{0i}(t)|\Theta_i\right)=\sum_{j=1}^{R}\alpha_{ij}\left(\frac{1}{\sqrt{2\pi\sigma_{ij}^2}}e^{-\frac{(d_{0i}(t)-\mu_{ij})^2}{2\sigma_{ij}^2}}\right), \tag{4}$$

where $0\leqslant\alpha_{ij}\leqslant 1$, $j=1,2,...,R$, and $\sum_{j=1}^{R}\alpha_{ij}=1$, $i=1,2,...,l_0$.

We use expectation-maximization (EM) algorithm to get $\tilde{\Theta}_i=\arg\max_{\Theta_i}\left(\prod_{t=1}^{T}p(d_{0i}(t)|\Theta_i)\right)$, $i=1,2,...,l_0$. Let $\tilde{\boldsymbol{\Theta}}=\left(\tilde{\Theta}_1,\tilde{\Theta}_2,...,\tilde{\Theta}_{l_0}\right)'$. We assume that the likelihood calculated based on the maximum likelihood solution of the parameters is a good representation for the unconditional one which could have been calculated by integrating out prior distributions of the parameters, in order to reduce the computation burden. Now we have

$$p(\boldsymbol{D}_0|\boldsymbol{I})=p(\boldsymbol{D}_0|\tilde{\boldsymbol{\Theta}},\boldsymbol{I}). \tag{5}$$

For a CrGM M_m, $m = 1,2,...,K$, containing l_m co-regulated genes, the state of regulator $s_m(t)$ follows kth-order Markov model with R possible state values 1,2,...,R over time. Let $\boldsymbol{\pi}_m$ be an R^k -dimensional vector that denotes the joint probability of all R^k possible values for the original k states. $\boldsymbol{T}_m$ is an R^k by R transition matrix, where the element in the ith row and the jth column $\boldsymbol{T}_m(i,j)$ denotes the probability that the current state has the jth value conditional on the fact that the vector of the last k states has the ith value. Let $\boldsymbol{E}_{mi}$ be the emission matrix for the ith gene in M_m where the element in the jth row and the first column $\boldsymbol{E}_{mi}(j,1)$ and the element in the jth row and the second column $\boldsymbol{E}_{mi}(j,2)$ are the mean and standard deviation parameters of the Gaussian distribution that the ith gene in M_m follows, given the current state of the hidden regulator has its jth value.

Let $\lambda_m = \{\boldsymbol{T}_m,\boldsymbol{\pi}_m,\boldsymbol{E}_{m1},\boldsymbol{E}_{m2},...,\boldsymbol{E}_{ml_m}\}$ denote all the parameters for the HMM we use here. The likelihood of the observed increment $\boldsymbol{D}_m$, $m = 1,2,...,K$, given the partition $\boldsymbol{I}$ can be written as

$$\begin{aligned} &p(\boldsymbol{D}_m|\lambda_m,\boldsymbol{I}) \\ &= \sum_{\text{all } S_m} p(D_{m1},D_{m2},...,D_{ml_m}|S_m,\lambda_m)p(S_m|\lambda_m) \\ &= \sum_{\text{all } S_m} \left(\prod_{i=1}^{l_m}\prod_{t=1}^{T} p\left(d_{mi}(t)|S_m,\lambda_m\right)\right)p(S_m|\lambda_m). \end{aligned} \quad (6)$$

The summation over all state sequences could be calculated efficiently using forward algorithm [11].

The maximum likelihood solution for λ_m, $\tilde{\lambda}_m = \underset{\lambda_m}{\text{argmax}}\, p(\boldsymbol{D}_m|\lambda_m,\boldsymbol{I})$ is obtained with Baum-Welch algorithm [12]. Now we also use the likelihood calculated with $\tilde{\lambda}_m$ to represent the likelihood of the observed increments of the expression levels of genes in M_m in order to reduce the computational burden which could have been quite heavy if the prior distributions of the parameters were integrated out. We have

$$p(\boldsymbol{D}_m|\boldsymbol{I}) = p\left(\boldsymbol{D}_m|\tilde{\lambda}_m,\boldsymbol{I}\right). \quad (7)$$

Considering the definition of CrGM, the distribution of increments of the expression levels of genes in different CrGMs are independent. Putting the above likelihood functions together, we have the posterior distribution of $\boldsymbol{I}$, given the observed increments $\boldsymbol{D}$, as

$$p(\boldsymbol{I}|\boldsymbol{D}) \propto p(\boldsymbol{I})\prod_{m=0}^{K} p(\boldsymbol{D}_m|\boldsymbol{I}). \quad (8)$$

Here we need to set the prior distribution for the marker partition first. Different kinds of informative prior knowledge could be integrated into the proposed Bayesian framework. For simplicity in comparing different methods in this paper, we assume that the prior distributions of the partitions of genes are independent, and for each gene, without prior knowledge, the probability that it belongs to the mth module is ρ_m, $0 \leqslant \rho_m \leqslant 1$, $\sum_{m=0}^{K}\rho_m = 1$. We set $\rho_m = 1/L$, $m = 1,2,...,K$, unless otherwise specified.

2.3 Gibbs sampling strategy for gene partitioning

As we usually have tens of thousands of genes that are under investigation, it is not easy to make statistical inference directly from Eq. (8). In order to obtain the posterior distribution of $\boldsymbol{I}$ and detect the CrGMs, we introduce Gibbs sampling strategy here. Given the posterior distribution of $\boldsymbol{I}$ in Eq. (8), we have the following Gibbs sampler:

$$p(I_i = m|\boldsymbol{I}_{[-i]},\boldsymbol{D}) = \frac{p(\boldsymbol{I}_i = m,\boldsymbol{I}_{[-i]},\boldsymbol{D})}{\sum_{m'=0}^{K} p(I_i = m',\boldsymbol{I}_{[-i]},\boldsymbol{D})}, \quad (9)$$

where $m = 0,1,...,K$ and $\boldsymbol{I}_{[-i]} = (I_1,...,I_{i-1},I_{i+1},...,I_L)'$. While the likelihoods share some common components, we are able to do the calculation in the following way. Let

$$\begin{aligned} \mu_m &= \frac{p(I_i = m|\boldsymbol{I}_{[-i]},\boldsymbol{D})}{p(I_i = 0|\boldsymbol{I}_{[-i]},\boldsymbol{D})} \\ &= \frac{p(I_i = m,\boldsymbol{I}_{[-i]})\prod_{m'=0}^{K} p(\boldsymbol{D}_{m'}|I_i = m,\boldsymbol{I}_{[-i]})}{p(I_i = 0,\boldsymbol{I}_{[-i]})\prod_{m'=0}^{K} p(\boldsymbol{D}_{m'}|I_i = 0,\boldsymbol{I}_{[-i]})} \\ &= \frac{p(I_i = m)p(\boldsymbol{D}_0|I_i = m,\boldsymbol{I}_{[-i]})p(\boldsymbol{D}_m|I_i = m,\boldsymbol{I}_{[-i]})}{p(I_i = 0)p(\boldsymbol{D}_0|I_i = 0,\boldsymbol{I}_{[-i]})p(\boldsymbol{D}_m|I_i = 0,\boldsymbol{I}_{[-i]})}, \end{aligned} \quad (10)$$

for $m = 0,1,...,K$, and then we have

$$p(I_i = m|\boldsymbol{I}_{[-i]},\boldsymbol{D}) = \frac{\mu_m}{\sum_{m'=0}^{K}\mu_{m'}}. \quad (11)$$

Now the Gibbs sampling procedure is performed as follows:

Step 1 Generate initial value for indicator I_i for $i = 1,2,...,L$ from its prior distribution ρ_m, $m = 0,1,...,K$.

Step 2 Randomly choose an indicator I_i, $i = 1,2,...,L$, calculate its posterior probabilities $p(I_i = m|\boldsymbol{I}_{[-i]},\boldsymbol{D})$, $m = 0,1,...,K$, and then update the indicator accordingly.

Step 3 Repeat Step 2 until convergence or a pre-defined maximum number of iterations.

Here we assume the number of CrGMs is known (K) and use it in the Gibbs sampling procedure. However, in real data analysis, K is generally unknown. The handling of the uncertainty of K will be discussed in the following section, together with the integration of Gibbs sampling and RJ-MCMC.

2.4 Sampling the number of modules

RJ-MCMC procedure [13] is introduced here to address the uncertainty of K in the application of the proposed method in analyzing real time series gene expression data. The following procedure illustrates the way that RJ-MCMC is integrated with Gibbs sampling to give the posterior distribution of CrGMs.

Step 1 Set K as a positive integer, e.g., $K = 1$.

Step 2 Implement the Gibbs sampling procedure for $n_1 L$ iterations with the current value of K. Record $P = p(\boldsymbol{I})\prod_{m=0}^{K} p(\boldsymbol{D}_m|\boldsymbol{I})$.

Step 3 K is increased ($K' \leftarrow K+1$) with probability p_i or decreased ($K' \leftarrow K-1$) with probability p_d, where we have $p_i \geqslant 0$, $p_d \geqslant 0$, and $p_i + p_d = 1$. There are two special cases: when K equals to 1, we do not decrease K; when there are empty modules, the increase of K is not conducted.

Step 4 Implement the Gibbs sampling procedure for $n_2 L$ iterations with the current value of K. Record $P' = p(\boldsymbol{I})\prod_{m=0}^{K'} p(\boldsymbol{D}_m|\boldsymbol{I})$.

Step 5 The new parameter value K' is accepted with probability $\alpha(K \leftarrow K') = \min\left\{1, \frac{P' p_d}{P p_i}\right\}$ if K is increased, or $\alpha(K \leftarrow K') = \min\left\{1, \frac{P' p_i}{P p_d}\right\}$ if K is decreased.

Step 6 Repeat Steps 2 to 5.

The above procedure is used to get a Markov chain that gives the posterior distribution of the CrGMs in which the number of modules (K) follows its stationary distribution. In this paper we use $p_i = p_d = 0.5$, $n_1 = 10$, and $n_2 = 5$.

2.5 Statistical significance of CrGMs

Starting from modeling the posterior distributions of co-regulated gene modules and random genes with its time series expression level data, the maximum-likelihood solutions are obtained with HMM and GMM, respectively. Then, we are able to partition the markers into different modules (including the module that consists of all the random genes). To obtain the posterior probability of this partition, we develop an MCMC sampling method that integrates both Gibbs sampling and RJ-MCMC. While RJ-MCMC addresses the uncertainty in the number of CrGMs K, we record the samples obtained from Gibbs sampling, when RJ-MCMC reaches its stationary distribution after a sufficient number of burn-in iterations. This framework is basically developed from Bayesian perspective and the sampled posterior probability is quite informative for further statistical inference, e.g., researchers may focus more on genes from CrGMs that have higher posterior probabilities to study their interactions and regulation relationships. However, it may be also useful to obtain the statistical significance for the CrGMs under the frequentist hypothesis testing framework.

For a detected CrGM, the null hypothesis H_0 is that the genes in this module are not co-regulated and the alternative hypothesis H_1 is that the genes are co-regulated. We propose here a permutation test procedure in which the posterior probabilities of CrGMs are used as test statistic so that we can give the experiment-wised significant level (EWSL) for the detected CrGMs. The procedure is summarized as follows:

Step 1 Implement the sampling procedure to the observed time series microarray data. Record all detected CrGMs, their module sizes (the number of genes included in the module), and their posterior probabilities. Also record the parameter setting used here.

Step 2 Generate a permuted data set by shuffling the time labels of each gene from the observed time series microarray data.

Step 3 Apply the sampling procedure to the permuted data, with the same parameter setting as used in the original data. Record the maximum posterior probability of the detected CrGMs for each module size. Record zero if there is no CrGM detected for a certain module size.

Step 4 Repeat Steps 2 and 3 N times and record N maximum posterior probabilities for each module size.

Step 5 For each CrGM (suppose its size is s) detected from the observed time series microarray data, calculate the number of times that the N maximum posterior probabilities for module size s obtained from permuted data sets are greater than or equal to the posterior probability of this CrGM and divide this number by N to obtain a p-value for this CrGM.

3 Results

3.1 Time series microarray models

To test the performance of the proposed CrgMODE, we design four time series microarray models with different characteristics, as illustrated in Table 1. All the four models have 2 CrGMs containing 5 and 10 genes, respectively. The orders of Markov chains of the hidden regulators of Model l and Model 2 are both 1, and the numbers of genes of the two models are 20 and 100, respectively. The orders of Markov chains of the hidden regulators of Model 3 and Model 4 are both 2, and the numbers of genes of the two models are 20 and 100, respectively.

Based on the above models, we generate increments of the expression levels of regulated genes following HMM and unregulated ones following GMM. The transition matrices of these four models are shown in Table 2. We set the value of the regulator state 1, 2 and 3. The regulator changes to one of the three states following the corresponding transition matrix. When the state of the regulator is fixed, the value of the increment of the expression level of one gene in a CrGM is generated by a Gaussian distribution for which the mean is in Table 3 and

Table 1 Four time series microarray models with different characteristics

model	order	number of genes	number of CrGMs	number of genes in CrGM 1	number of genes in CrGM 2
Model 1	1	20	2	5	10
Model 2	1	100	2	5	10
Model 3	2	20	2	5	10
Model 4	2	100	2	5	10

Table 2 Transition matrices

model	CrGM	order	transition matrix
Model 1 / 2	CrGM 1	1	$\begin{bmatrix} 0 & 1 & 0 \\ 0.05 & 0.05 & 0.9 \\ 0.9 & 0.05 & 0.05 \end{bmatrix}$
	CrGM 2	1	$\begin{bmatrix} 0.05 & 0.9 & 0.05 \\ 1/3 & 1/3 & 1/3 \\ 0.9 & 0.05 & 0.05 \end{bmatrix}$
Model 3 / 4	CrGM 1	2	$\begin{bmatrix} 0.9 & 0.05 & 0.05 \\ 1/3 & 1/3 & 1/3 \\ 0.05 & 0.05 & 0.9 \\ 1/3 & 1/3 & 1/3 \\ 0.05 & 0.9 & 0.05 \\ 1/3 & 1/3 & 1/3 \\ 0.9 & 0.05 & 0.05 \\ 1/3 & 1/3 & 1/3 \\ 0.05 & 0.05 & 0.9 \end{bmatrix}$
	CrGM 2	2	$\begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 0.05 & 0.9 & 0.05 \\ 1/3 & 1/3 & 1/3 \\ 0.05 & 0.05 & 0.9 \\ 1/3 & 1/3 & 1/3 \\ 0.9 & 0.05 & 0.05 \\ 1/3 & 1/3 & 1/3 \\ 0.05 & 0.9 & 0.05 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$

the standard deviation is 0.05. For all random genes, in increments of their expression levels follow Gaussian distribution for which the mean parameter is 0 and the standard deviation is 0.75.

With these models, we are able to generate increments for simulation studies, in which positive and negative real numbers represent the activated and repressed transcription of downstream genes, respectively. Then we obtain the simulated gene expression data according to Eq. (2).

3.2 Performance of the proposed CrgMODE

We generate one data set for each of the four models where there are 1000 time points for each gene and the genes in CrGMs are the last ones, and implement CrgMODE on the four data sets. The results of the posterior probability that each gene belongs to each module are shown in Fig. 3. For each of the four data sets, we run $100L$ iterations of the Gibbs sampling procedure, in which the first half is taken as burn-in, and the second half is used to obtain the posterior probabilities of CrGMs, by recording detected CrGMs once in every L iterations. After we get the posterior probabilities for all the detected CrGMs from Gibbs sampling procedure, we choose those with posterior probability larger than 0.8 and p-value smaller than 0.05 as the identified CrGMs. The p-values of CrGMs are calculated as follows:

We generate 100 permuted data sets for each simulated data set by shuffling the time labels of each gene. For each permuted data set, we also run the Gibbs sampling procedure ($100L$ iterations) with the same parameter setting as used in the original observed (simulated) data

Table 3 Mean of the Gaussian distribution for emission matrices

state value		1	2	3
CrGM 1	Gene 1	−0.4	0.8	0
	Gene 2	0.8	0.4	0
	Gene 3	0.4	−0.4	−0.8
	Gene 4	0	−0.4	0.8
	Gene 5	0.8	0.6	−0.8
CrGM 2	Gene 1	1.4	0.2	−1.4
	Gene 2	0.6	1.2	0.6
	Gene 3	1.2	−1	−1.4
	Gene 4	−1	−1.2	−0.2
	Gene 5	−0.2	−0.6	1.2
	Gene 6	−0.2	−1	−1.4
	Gene 7	−1	−1.4	1
	Gene 8	−1.4	0.6	−0.2
	Gene 9	−1	0.6	−0.6
	Gene 10	−1.2	0.2	1.2

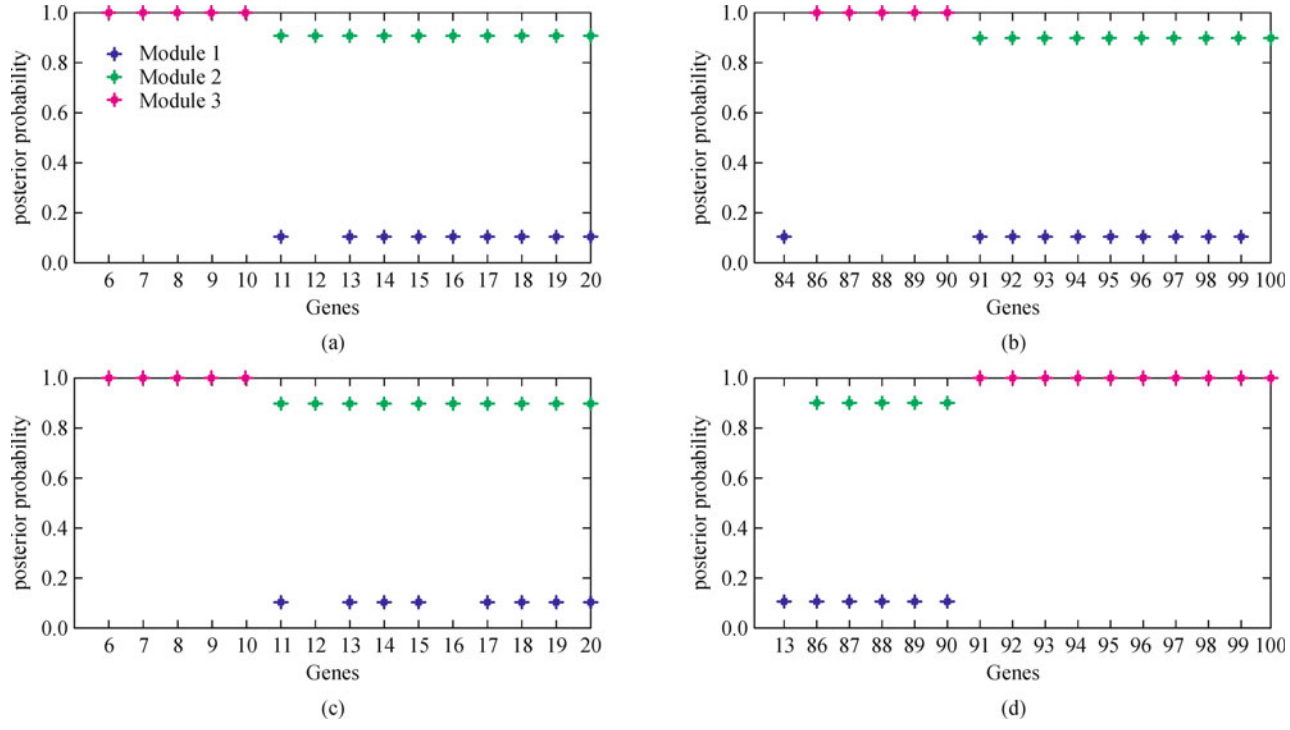

Fig. 3 Posterior probabilities of genes that belong to certain modules. For each time series microarray models, the posterior probabilities of genes that belong to CrGMs are plotted. Posterior probabilities of all the other genes are all zero. For all the four models, counting backward in the gene list, the 1st to the 10th genes are in the same CrGM and the 11th to the 15th genes are in the same CrGM. (a) Model 1; (b) Model 2; (c) Model 3; (d) Model 4.

set and record the maximum posterior probability for each module size. Finally, for each CrGM (suppose its size is s), we calculate the proportion that the recorded maximum probabilities for module size s are greater than or equal to the posterior probability of the candidate module to obtain its p-value (not shown here).

We can see that although random genes could be partitioned into CrGMs, it is quite often that they will be kicked out of CrGMs in the following sampling procedure with the detection of new true co-regulated genes in the same CrGM. As a result, all the true CrGMs are identified (with posterior probability larger than 0.8 and p-value smaller than

0.05) and these identified CrGMs contain and only contain all the genes that belong to them as designed in the models. No random genes are identified in any CrGM.

3.3 Comparison with existing methods

To further illustrate the performance of the proposed method, we also implement the popular partitioning *K*-means [14] and the co-expression-based module identification algorithm WGCNA [15–17] to identify modules from the four models.

For the *K*-means analysis, the number of clusters is set to 3 to match the actual number of modules ($K + 1$). As *K*-means is not able to determine directly which cluster contains mainly co-regulated genes in the same CrGM and which contains random genes, it is not easy to choose clusters as detected co-regulated ones for comparison. Here we select two clusters out of the three that give the highest ACC as co-regulated ones to be used in the comparison with other methods.

WGCNA was proposed by Horvath et al. to detect modules of co-expressed genes [15–17]. To implement WGCNA, one should first construct a weighted gene co-expression network based on pair-wised Pearson correlations between the expression profiles and then use hierarchical clustering to detect modules [16]. We use the code provided by Horvath et al. to implement WGCNA [15–17]. When implementing hierarchical clustering, we set the height parameter where the tree should be cut as 0.6 and the number of genes a module should at least contain as 2. The co-expression modules are treated as the detected modules.

For each of the four models, we generate 100 independent data sets, each of which contains 1000 time points. Then, we implement CrgMODE, *K*-means, and WGCNA on each data set. *K*-means and WGCNA are also implemented on the increment space, where increments are obtained with Eq. (2) first to replace the original expression data. Taking genes in CrGM as the positive ones, and genes unregulated as the negative ones, we calculate the expressions for specificity (SP), sensitivity (SE), accuracy (ACC), and Fscore as follows:

$$\begin{aligned} SP &= \frac{TN}{TN + FP}, \\ SE &= \frac{TP}{TP + FN}, \\ ACC &= \frac{TP + TN}{TP + TN + FP + FN}, \\ PR &= \frac{TP}{TP + FP}, \\ RE &= \frac{TP}{TP + FN}, \\ Fscore &= 2 \times \frac{PR \times RE}{PR + RE}, \end{aligned} \tag{12}$$

where *TP*, *TN*, *FP*, and *FN* represent true-positives, true-negatives, false-positives, and false-negatives, respectively. *PR* and *RE* are short for precision and recall, respectively.

After implementing CrgMODE, *K*-means, and WGCNA on the data sets in each model, we get the mean of SP, SE, ACC, and Fscore.

Figure 4 shows the comparison of the four statistics of the three methods. The means of SP, SE, ACC, and Fscore obtained by CrgMODE are all higher than or equal to that of *K*-means and WGCNA. The observation that *K*-means tends to have higher SP but lower SE, implying that *K*-means is not quite capable in clustering co-regulated gene into the same group. WGCNA has higher SE than *K*-means, which is quite subtle, but still significantly lower than the proposed method. This may be due to the fact that WGCNA has no power in capturing dynamic relationships in gene regulations.

4 Discussion

The posterior probability of the indicator vector is proportional to the product of prior probability and the likelihood of observed data, where the likelihood of each CrGM are obtained from HMM and the likelihood of random genes are calculated based on GMM. The uncertainty of the number of CrGMs is addressed through RJ-MCMC procedure and the posterior probabilities of CrGMs are obtained by Gibbs sampling. A permutation-based procedure is developed to give the experiment-wised significance level for each detected CrGM. The comparison between the proposed approach and two existing methods shows the advantage of CrgMODE in detecting genes that are co-regulated.

The proposed approach focuses on the interaction between gene expressions in dynamic biological process. This interaction effect could be linear correlations based on which some existing methods are developed. However, the introduction of the concept of co-regulation, together with the application of Bayesian model, gives CrgMODE the fundamental advantage in detecting genes that express in some more complex nonlinear ways. It should also be noticed that with modeling CrGMs in a Bayesian framework, CrgMODE works on the multiple genes that are co-regulated simultaneously, rather than investigating only pair-wised relationships one by one. The deriving of higher order information could also be a possible reason that the proposed method may be able to get more insights compared with conventional methods.

The proposed method describes the time series gene expression data as an HMM model, and an MCMC-based method is developed in order to obtain the posterior probabilities. Besides the consideration of interaction effects and the idea of partitioning genes into modules whose concepts are well defined, the most significant

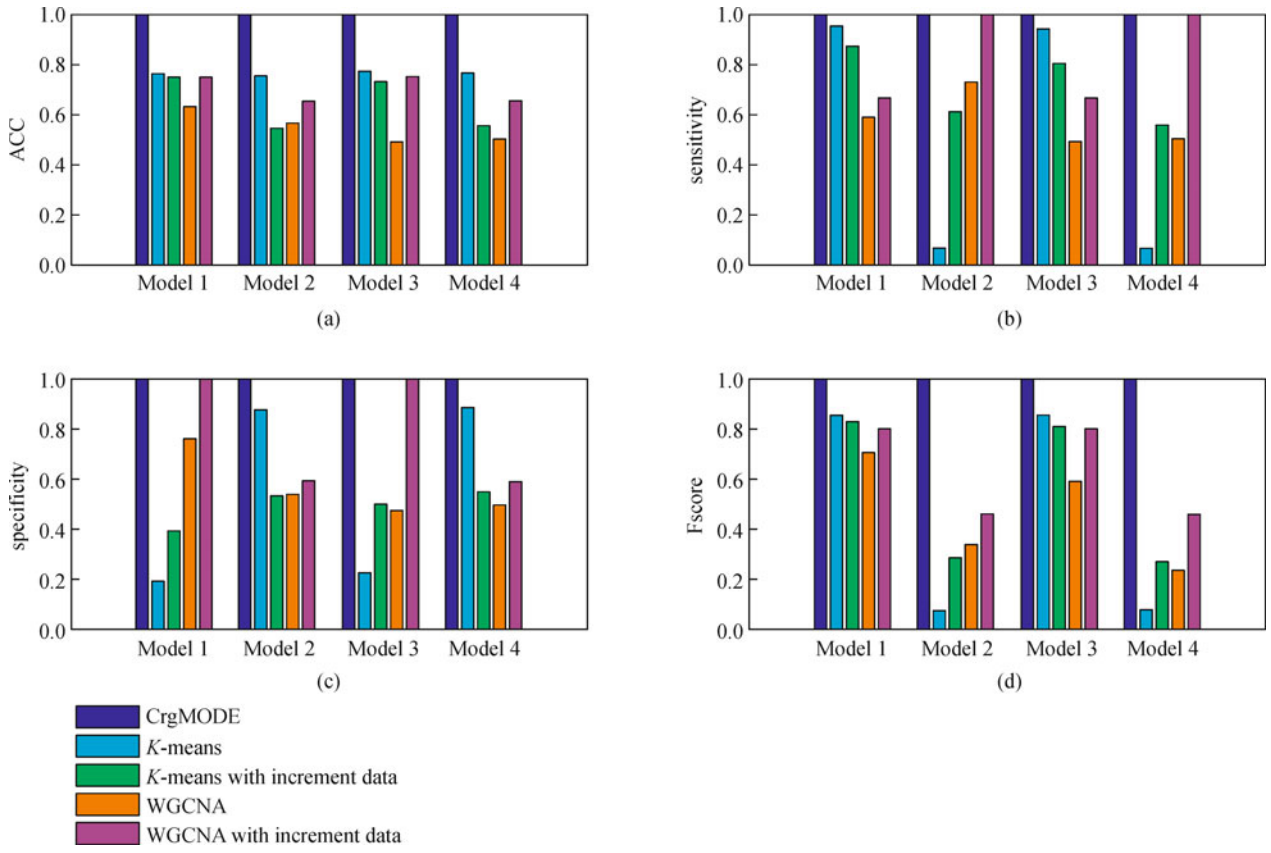

Fig. 4 Comparison of CrgMODE, *K*-means, and WGCNA.

advantage of the proposed method may be due to the fact that it uses the time series information that existing methods hardly use. While the information in dynamic biological process is important to certain biological functions, it is reasonable that we could get more insights by properly analyzing this information.

A natural advantage of Bayesian methods is its flexibility in integrating prior knowledge, which could be done in both parametric way and non-parametric way. The prior probabilities of indicators for certain genes could be used to reflect existing biological knowledge or results from other independent works, e.g., network-based analyzing results [4]. Comparing with the use of likelihood calculated based on maximum likelihood solution of parameter, e.g., $\tilde{\boldsymbol{\Theta}}$ and λ_m, the integration over parameter spaces with proper prior distributions could reflect more understanding of the problem and be more robust. This will be one direction of our future work.

Although CrgMODE is proposed for the analyzing of time series gene expression data, it is quite natural to extend this approach to the application on other types of large scale time series data. While the module-based methods have already showed their success in detecting interactions from population-based data [18], it will be quite interesting to extend CrgMODE to the 3-dimensional situation (samples, markers, and time points).

Acknowledgements This work was partly supported by the National Natural Science Foundation of China (Grant Nos. 60934004 and 61021063) and the Beijing excellent PhD thesis project.

References

1. Mootha V K, Lindgren C M, Eriksson K F, Subramanian A, Sihag S, Lehar J, Puigserver P, Carlsson E, Ridderstråle M, Laurila E, Houstis N, Daly M J, Patterson N, Mesirov J P, Golub T R, Tamayo P, Spiegelman B, Lander E S, Hirschhorn J N, Altshuler D, Groop L C. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics, 2003, 34(3): 267–273
2. Hartwell L H, Hopfield J J, Leibler S, Murray A W. From molecular to modular cell biology. Nature, 1999, 402(6761 Suppl): C47–C52
3. Wang L, Zhang B, Wolfinger R D, Chen X. An integrated approach for the analysis of biological pathways using mixed models. PLoS Genetics, 2008, 4(7): e1000115
4. Gu J, Chen Y, Li S, Li Y. Identification of responsive gene modules by network-based gene clustering and extending: Application to inflammation and angiogenesis. BMC Systems Biology, 2010, 4(1): 47
5. Eisen M B, Spellman P T, Brown P O, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America,

1998, 95(25): 14863–14868

6. Tavazoie S, Hughes J D, Campbell M J, Cho R J, Church G M. Systematic determination of genetic network architecture. Nature Genetics, 1999, 22(3): 281–285
7. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander E S, Golub T R. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proceedings of the National Academy of Sciences of the United States of America, 1999, 96(6): 2907–2912
8. Carter S L, Brechbühler C M, Griffin M, Bond A T. Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics, 2004, 20(14): 2242–2250
9. Davidson G S, Wylie B N, Boyack K W. Cluster stability and the use of noise in interpretation of clustering. In: Proceedings of IEEE Symposium on Information Visualization 2001. 2001, 23–30
10. Elo L L, Järvenpää H, Oresic M, Lahesmaa R, Aittokallio T. Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process. Bioinformatics, 2007, 23(16): 2096–2103
11. Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 1989, 77(2): 257–286
12. Baum L E, Petrie T, Soules G, Weiss N. A maximization technique occurring in statistical analysis of probabilistic functions of Markov chains. Annals of Mathematical Statistics, 1970, 41(1): 164–171
13. Green P J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika, 1995, 82(4): 711–732
14. Strauss D J. Clustering algorithms — Hartigan, JA. Biometrics, 1975, 31(3): 793
15. Oldham M C, Horvath S, Geschwind D H. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. Proceedings of the National Academy of Sciences of the United States of America, 2006, 103(47): 17973–17978
16. Horvath S, Zhang B, Carlson M, Lu K V, Zhu S, Felciano R M, Laurance M F, Zhao W, Qi S, Chen Z, Lee Y, Scheck A C, Liau L M, Wu H, Geschwind D H, Febbo P G, Kornblum H I, Cloughesy T F, Nelson S F, Mischel P S. Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. Proceedings of the National Academy of Sciences of the United States of America, 2006, 103(46): 17402–17407
17. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology, 2005, 4: Article17
18. Tang W, Wu X, Jiang R, Li Y. Epistatic module detection for case-control studies: A Bayesian model with a Gibbs sampling strategy. PLoS Genetics, 2009, 5(5): e1000464