



Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data

Y. Luan and H. Li*

Rowe Program in Human Genetics, School of Medicine, University of California, Davis, CA 95616, USA

Received on April 10, 2003; revised on June 27, 2003; accepted on July 29, 2003

ABSTRACT

Motivation: The expressions of many genes associated with certain periodic biological and cell cycle processes such as circadian rhythm regulation are known to be rhythmic. Identification of the genes whose time course expressions are synchronized to certain periodic biological process may help to elucidate the molecular basis of many diseases, and these gene products may in turn represent drug targets relevant to those diseases.

Results: We propose in this paper a statistical framework based on a shape-invariant model together with a false discovery rate (FDR) procedure for identifying periodically expressed genes based on microarray time-course gene expression data and a set of known periodically expressed guide genes. We applied the proposed methods to the α -factor, *cdc15* and *cdc28* synchronized yeast cell cycle data sets and identified a total of 1010 cell-cycle-regulated genes at a FDR of 0.5% in at least one of the three data sets analyzed, including 89 (86%) of 104 known periodic transcripts. We also identified 344 and 201 circadian rhythmic genes *in vivo* in mouse heart and liver tissues with FDR of 10 and 2.5%, respectively. Our results also indicate that the shape-invariant model fits the data well and provides estimate of the common shape function and the relative phases for these periodically regulated genes.

Availability: Matlab programs are available on request from the authors.

Contact: hli@ucdavis.edu

Supplementary information: <http://dna.ucdavis.edu/~hli/period.html>

INTRODUCTION

The expressions of many genes associated with certain periodic biological processes, such as circadian rhythmic regulation and cell cycle regulation are known to be rhythmic. In contrast, the gene expression profiles of genes associated with aperiodic biological processes, such as tissue repair and response to serum stimulus are not rhythmic. Such

transcriptional rhythms can be very important for daily timing of physiological processes (Storch *et al.*, 2002). It is therefore important to identify those genes whose expressions are synchronized to some ongoing biological processes (Langmead *et al.*, 2002). Identification of these genes may help in studying the molecular basis of many diseases and in turn provides potential drug targets for treating those diseases. For example, in *Drosophila melanogaster*, mutations affecting any of the seven known clock genes have corresponding effects on behavioral and molecular rhythms. Indeed, human orthologs of three of these *Drosophila* 'clock' genes have been associated with disorders of sleep (Claridge-Chang *et al.*, 2001) and most of these genes are periodically regulated.

DNA microarray experiments allow for genome-wide identification of periodically expressed genes synchronized to biological processes. Typically, time course gene expression data are collected by microarray experiments in which gene expression levels of thousands of genes are measured across a number of time points during the biological process. For example, Cho *et al.* (1998) and Spellman *et al.* (1998) performed genome-wide transcriptional analysis of mitotic cell cycle of yeast using microarrays and identified about 800 cell-cycle-regulated genes. By using microarray technology, Claridge-Chang *et al.* (2001) identified about 400 transcripts that showed significant oscillation in the head of *Drosophila*. Storch *et al.* (2002) studied circadian gene expression in mouse liver and heart *in vivo*. The methods employed in identifying these genes in these papers range from Fourier analysis (Spellman *et al.*, 1998; Claridge-Chang *et al.*, 2001) to methods using some threshold criteria (Storch *et al.*, 2002). However, most methods used are *ad hoc*, and none of these methods attempted to model the observed noisy microarray data. Johansson *et al.* (2003) proposed to use the partial least square regression to identify genes with periodic fluctuations in expression levels coupled to the cell cycle in the budding yeast, where they used *sine* and *cosine* curves for fitting the observed expression profile for each gene. However, due to possible lack of synchronization of the cells at later times during the time course and due to the fact that

*To whom correspondence should be addressed.

the cells spend different durations over different cell cycle phases, the simple *sine* or *cosine* curves may not fit the data well. In addition, for short time course gene expression data, which are very typical for microarray time course studies, the Fourier transformation or general time-series spectra analysis may not work well (Langmead *et al.*, 2002).

The aim of this paper is to develop methods for identifying genes that show periodic expression patterns during the time course of a biological process. From previous experiments, biologists are often certain about a set of genes of known function which show certain patterns of expression during a given biological process. In this paper, we call these genes the guide genes, the term that was used in Storch *et al.* (2002). Based on the time course gene expression profiles of these guide genes, we propose to develop statistical models to estimate the gene expression patterns and to identify other genes that follow similar expression patterns. The challenge is that genes with similar expression profiles may have same patterns but different phases and/or amplitudes and overall expression levels. To accommodate these differences among the guide genes and other potential periodically expressed genes, we propose to use a shape-invariant model (Lawton *et al.*, 1972) with a cubic B-spline based periodic function for modeling the common curve. This model explicitly models the gene expression profile as a function of time. Given the estimated common expression profile, for a given test gene, we propose to perform a likelihood ratio test for testing the amplitude of the gene being zero, and to employ the false discovery rate (FDR) procedure (Benjamini and Hochberg, 1995) for identifying genes with periodic patterns similar to the guide genes and for controlling the percentage of the falsely identified genes.

The rest of the paper is organized as follows: we first present the shape-invariant model and a two-stage procedure for estimating the parameters based on the time course gene expression data of the guide genes. We then present a procedure for testing the amplitude and for identifying periodically expressed genes using FDR. After the methods section, we present applications of the methods to the yeast cell cycle data for identifying cell-cycle-regulated genes and to the circadian gene expression data sets in mouse liver and heart to identify circadian rhythmic genes. We conclude with a brief discussion of the methods and results.

STATISTICAL METHODS

A shape-invariant model for the guide genes

Assume that we have known m genes which show common periodic expression patterns during a biological process, such as cell cycle or circadian rhythm regulation. Let Y_{ij} be the log-gene expression level measured by cDNA or Affymetrix arrays at time t_{ij} , for $i = 1, \dots, m$ and $j = 1, \dots, n_i$, where n_i is the number of gene expression data available for the i -th gene, allowing for possible missing gene expression data at

some time points. Without loss of generality, we assume that the period is 1, and the time points are measured in $[0, 1]$ interval. We assume that these genes follow the same expression pattern, but their individual profiles may differ in phases and/or amplitudes, and propose to adopt the shape-invariant model developed in Lawton *et al.* (1972) and Wang and Brown (1996) for modeling the gene expression profiles of the guide genes. This model assumes

$$Y_{ij} = \mu_i + \beta_i f(t_{ij} - \tau_i) + \epsilon_{ij}, \quad (1)$$

for $i = 1, \dots, m$ and $j = 1, \dots, n_i$. Here, μ_i is the mean gene expression levels over time $[0, 1]$ for the i -th gene, f is the common curve, which is periodic with period equal to 1 and $\sup_{t \in [0, 1]} |f(t)| = 1$, β_i is the maximum deviation from the mean for the i -th gene, and $0 \leq \tau_i \leq 1$ is the phase of the i -th gene. Finally, ϵ_{ij} 's are the error terms. In this model, $\{\mu_i, \beta_i, \tau_i\}$ is the vector of gene-specific parameters and f is the common periodic function shared by all the periodically expressed genes.

In this paper, we model the function f in model (1) by the linear combination of cubic B-spline basis (De Boor, 1978),

$$f(t) = \sum_l^p \gamma_l B_l(t), \quad (2)$$

where p is the dimension of the B-spline basis and $\gamma = \{\gamma_1, \dots, \gamma_p\}$ is the coefficient of the B-splines basis. The cubic B-spline provides quite flexible functional form for modeling the curves (De Boor, 1978; Rice and Wu, 2001), and were demonstrated to fit the short time course gene expression data well (Li *et al.*, 2002; Luan and Li, 2003). We further restrict the coefficients $\gamma_l, l = 1, \dots, p$, so that the function $f(t)$ satisfies the period conditions:

$$f(0) = f(1), \quad f'(0) = f'(1). \quad (3)$$

This induces two linear constraints on parameter vector γ .

A two-step procedure for estimating the parameters

Under models (1) and (2), we propose to use a similar two-step procedure as in Wang and Brown (1996) for estimating the parameters, including the common function $f(t)$. In step 1, for a given f , for each gene, the gene-specific parameters μ_i, β_i and τ_i can be estimated by solving the non-linear least-square problem:

$$\min \frac{1}{n_i} \sum_{j=1}^{n_i} [y_{ij} - \mu_i - \beta_i f(t_{ij} - \tau)]^2,$$

which can be done simply by grid search. For fixed $\mu = \{\mu_1, \dots, \mu_m\}$, $\beta = \{\beta_1, \dots, \beta_m\}$ and $\tau = \{\tau_1, \dots, \tau_m\}$, step 2

involves only the minimization of

$$\sum_{i=1}^n \frac{1}{n_i} \sum_{j=1}^{n_i} [y_{ij} - \mu_i - \beta_i \sum_{l=1}^p \gamma_l B_l(t_{ij} - \tau_i)]^2,$$

over the B-spline coefficients γ_l for $l = 1, \dots, p$, subjects to linear constraints (3), which can be easily done by Newton–Ralphson procedure. Note that this two-step estimation procedure does not make any distributional assumption on the error term in model (1).

A FDR-based procedure for identifying periodically expressed genes

Based on the available data for guide genes, we first estimate the common function $f(t)$ by the two-step procedure discussed in last section, denoted by $\hat{f}(t)$ here. Our goal is then to identify other genes which follow the same periodic expression pattern, subject to horizontal shift (phase) and/or different amplitude. Under model (1), expression levels are periodically regulated if and only if $\beta \neq 0$. For a given gene with time-course gene expression profile $y = \{y_1, \dots, y_T\}$ measured at t_1, \dots, t_T , conditioning on the estimated common function based on data of the guide genes, we want to test whether $\beta = 0$ in model (1). This is equivalent to test how well the observed gene expression profile of a test gene fits the common curve, allowing for change of overall expression level, amplitude of the expression and phase of the common curve.

To facilitate such test, we assume that the error term ϵ_{ij} in model (1) follows a multivariate normal distribution with mean zero and first-order autoregressive correlation, i.e. the variance–covariance matrix is given by

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \dots & \rho^{n_i-1} \\ \rho & 1 & \rho & \dots & \rho^{n_i-2} \\ & & \vdots & & \\ \rho^{n_i-2} & \dots & \rho & 1 & \rho \\ \rho^{n_i-1} & \dots & & \rho & 1 \end{pmatrix},$$

where σ^2 is the error variance and ρ is the first-order correlation of errors between two nearby time points. We can then obtain the maximum likelihood estimate of the model parameters and perform a likelihood ratio test for $H_0: \beta = 0$.

Suppose that we have n -test genes for which we want to test whether they show similar expression patterns as the guide genes, and denote the p -values for testing $H_i: \beta_i = 0, i = 1, \dots, n$, as p_1, \dots, p_n . Usually n is quite large, and the standard Bonferroni adjustment for the type 1 error rate is too conservative and cannot be applied. Instead, we propose to employ the FDR procedure of Benjamini and Hochberg (1995) for choosing the cutoff point of the p -value in order to control for overall FDR. The FDR procedure provides an alternative to the multiple significance testing problem by calling for controlling the expected proportion of falsely rejected hypotheses, the FDR. Let $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(n)}$ be the

ordered p -values for the n -test genes, and denoted by $H_{(i)}$ the null hypothesis corresponding to $p_{(i)}$. Let k be the largest i for which $p_{(i)} \leq (i/n)q^*$, then reject all $H_{(0i)}, i = 1, \dots, k$. This procedure controls the FDR at q^* . In the context of this paper, the FDR can be interpreted as the proportion of genes that do not have periodic expression but are identified as periodically expressed by our methods.

APPLICATIONS TO REAL DATA SETS

We apply and demonstrate the proposed methods for identifying the cell-cycle-regulated genes in yeast and circadian rhythmic genes in liver and heart tissues in mouse. In all the analyses, we used cubic B-spline with two equally spaced knots for modeling the common function $f(t)$, therefore, $p = 6$ in function (2).

Identifying cell-cycle-regulated genes in yeast

Cell cycle is one of life’s most important processes, and identification of cell-cycle-regulated genes will greatly facilitate the understanding of this important process. Spellman *et al.* (1998) monitored genome-wide mRNA levels for 6178 yeast ORFs simultaneously using several different methods of synchronization including an α -factor-mediated G_1 arrest which covers approximately two cell-cycle periods with measurements at 7 min intervals for 119 min with a total of 18 time points, a temperature-sensitive *cdc15* mutation to induce a reversible M-phase arrest, and a temperature-sensitive *cdc28* mutation to arrest cells in G_1 phase reversibly (<http://genome-www.stanford.edu/cellcycle/data/rawdata/>). For the *cdc15* experiment, gene expression data were measured every 10 min for 290 min, lacking observations for the 0, 20, 40, 60, 260 and 280 min time point. For the *cdc28* experiment, samples were taken every 10 min from 0 to 160 min for a total of 17 time points. In the following analysis, we used the periods of 58, 115 and 85 min for the α -factor synchronized cells, *cdc15* cells and *cdc28* cells, respectively. These numbers were estimated by Zhao *et al.* (2001) by minimizing a weighted sum of squares and were also used by Johansson *et al.* (2003). Therefore, the data sets cover approximately two cell-cycle periods. We estimated the missing gene expression levels by using a nearest-neighbor estimation procedure where the average values of eight nearest genes with no missing data are used to estimate the missing data (Hastie *et al.*, 1999).

There are a total of 104 genes that were determined to be cell-cycle-regulated by traditional genetic analysis methods (Spellman *et al.*, 1998), but one gene had no data in the Spellman gene expression database. We therefore used 103 genes as our guide gene to build the models. For these synchronized microarray experiments, one would expect that as time goes, the cells become less synchronized and, therefore, the expression profiles are expected to be different between the first complete cell-cycle period and the second complete cell-cycle period. This fact can be clearly observed in our data

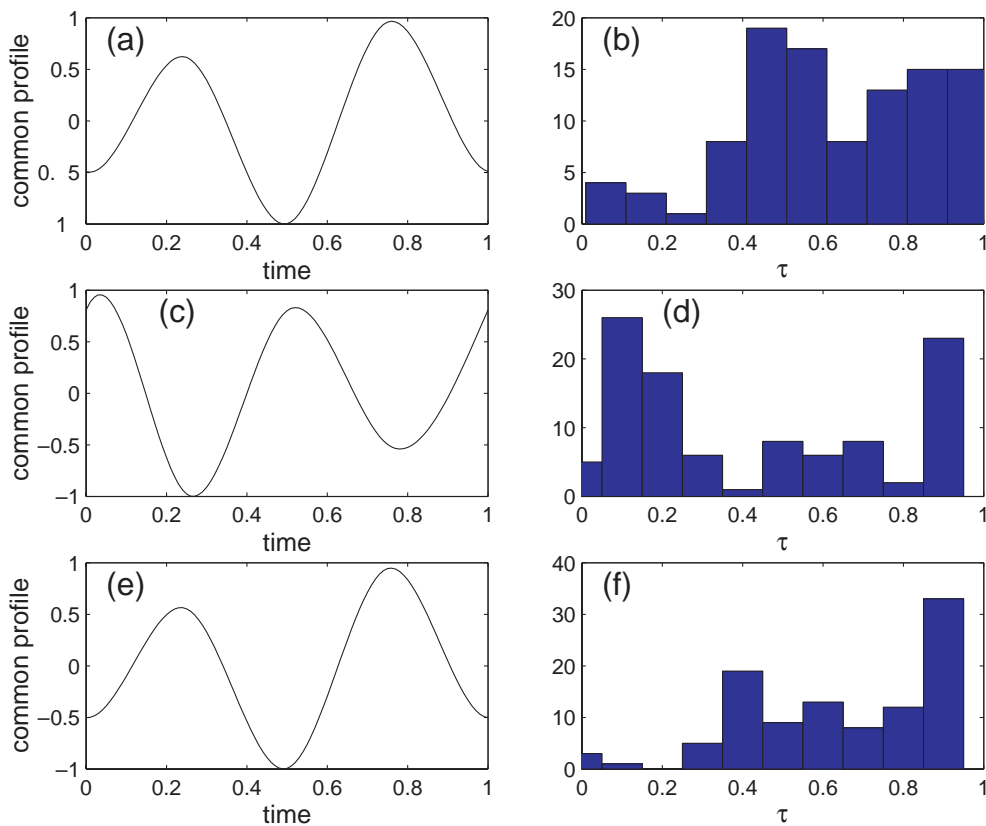


Fig. 1. Results for the yeast cell cycle gene expression data. Plots (a), (c) and (e) are the estimated common function based on the guide genes, for α -factor, *cdc15* and *cdc28* experiments, respectively. Plots (b), (d) and (f) are histograms the estimates of the phase for the 103 guide genes, for α -factor, *cdc15* and *cdc28* experiments, respectively. Note that the phase estimates are relative to the common curves and the combination of the common curve and the phase determines the shape of the expression profile for a given gene.

and was also observed by Zhao *et al.* (2001). To avoid explicitly modeling the lack of synchronization or attenuation in gene expression over time, we treat the data in two cell-cycle periods as one ‘combined period’ and use the guide genes to tell us what gene expression pattern we expect to observe in this ‘combined period’.

Figure 1a–f shows the estimates of the common function $f(t)$ based on these 103 guide genes and the histograms of the estimates of the phase parameter τ for the three different cell cycle experiments. It is interesting to note that although there are two clear peaks corresponding two cell cycle, the curves are not quite symmetric. This might be due to the lack of synchronization of the yeast cells as cell cycle unfolds. The shapes of the common function and the estimated phases indicate the gene expression levels are in general lower in the second cell-cycle periods. This is also expected due to attenuation in gene expression over time.

Based on the estimated common curves as shown in Figure 1, for a FDR of 0.5%, we identified 297, 482 and 623 periodically regulated transcripts during yeast cell cycle using the data set of α -factor, *cdc15* and *cdc28*, respectively.

The total number of transcripts that were identified in at least one experiment is 1010, including 89 or 86% of the known cell-cycle-regulated genes. Complete lists of the periodic transcripts identified by our methods are available in the web supplementary materials. Figure 2 shows the number of transcripts that were identified by each of the three data sets and the number of transcripts identified by two or three data sets. A total of 307 transcripts showed periodic expression pattern in at least two experiments and a total of 703 genes appeared periodic in expression in only one data set. Figure 3 shows the image plots of the normalized gene expression levels for the genes identified by each of the three data sets, sorted by the estimates of the phases τ . These genes show a very clear periodic expression patterns over time. Finally, we identified 47, 67 and 53 out of 103 known cell-cycle-regulated using the data set of α -factor, *cdc15* and *cdc28*, respectively. Plots of the observed gene expression data of the 14 known guide genes, which were missed by our methods indicate that these genes either did not show any clear periodic patterns or had very low level of gene expression (see web supplement Figure S1).

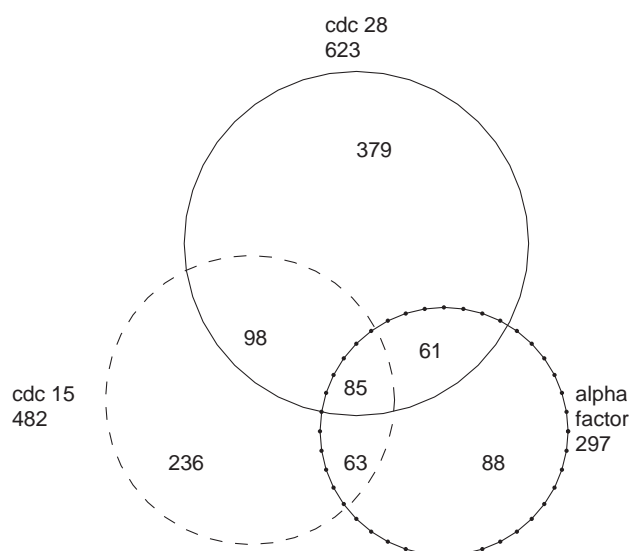


Fig. 2. Venn diagram of the number of genes identified based on three different cell cycle experiments for $FDR = 0.5\%$. The total number of periodic genes identified in each data set is shown and is represented by a circle.

Comparing results from different analyses for the yeast cell cycle data sets

As a comparison, Spellman *et al.* (1998) used Fourier analysis of combined data of the three experiments and identified 799 genes that are periodic, out of which 798 transcripts have gene expression data available. Our methods picked up 596 (75%) of these genes as being periodic in at least one data set. Nearly all the 307 genes that were identified to be periodically expressed by our methods in at least two data sets are also recognized by the methods of Spellman *et al.* (1998). However, there are 404 genes that were identified by our methods in at least one experiment, but were missed by Spellman *et al.* (1998). The gene expression plots of these genes (Fig. 3) show clear periodic expression pattern in at least one experiment. In contrast, there are 202 genes that were identified by Spellman *et al.* (1998) but were not identified as such by our methods.

Zhao *et al.* (2001) developed an interesting single-pulse mode (SPM) for the mean expression of each gene as cell cycle proceeds and applied this model to the three data sets of the yeast cell cycle experiments. Different from our B-spline model for the gene expression profile, they used a special parametric model for the mean gene expression. Under the mean model, they first test whether the observed data of a given gene significantly depart from SPM and for those genes for which the expression pattern does not deviate from SPM, they further test whether the elevation in gene expression is zero. By selecting appropriate cut-points so that the genome-wide significance level of about 0.3%, they identified a total of 1088 transcripts as periodically regulated, including 254 genes identified in at least two data sets and 834 genes

identified in only one data set. These numbers are quite comparable to what we identified. There are a total of 712 genes that were identified by both our methods and the methods in Zhao *et al.* (2001). However, due to different criteria/models used, there are 376 genes which were identified by Zhao *et al.* (2001) but were not classified as such by our methods. In contrast, there are 298 genes which were identified by our methods but were missed by the methods in Zhao *et al.* (2001). The image plots of the observed gene expression profiles of these genes are shown in Figure S2 in our web supplementary materials. A close examination of these genes cannot reach a conclusion on which methods work better.

As a final comparison of these two methods, Figure S3 in our web supplementary materials shows the observed gene expression data and the fitted B-spline curves to these profiles for the five periodic genes shown in Figure 3 of Zhao *et al.* (2001). Clearly, the B-spline model approximates the profiles of the data very well.

Identifying circadian rhythmic genes in liver and heart of mouse

Many mammalian peripheral tissues have endogenous circadian clock oscillators that generate transcriptional rhythms. Such transcriptional rhythms can be important for daily timing of physiological processes (Storch *et al.*, 2002). Storch *et al.* (2002) reported gene expression analysis *in vivo* in mouse liver and heart using oligonucleotide arrays representing 12 488 genes. In their experiment, mice were entrained to a 12 h light/dark cycle for more than 2 weeks, and then placed in constant dim light for ≥ 42 h. Gene expression levels were measured at 4 h interval over two circadian cycles, for a total of 12 time points. Storch *et al.* (2002) identified 575 genes in liver and 462 genes in heart with circadian expression patterns based on the gene expression profiles of nine guide genes that are known to exhibit circadian regulation in both liver and heart, including *Per1*, *Per2*, *Per3*, *Bmal1*, *Tef*, *Dpb*, *E4bp4*, *Cry1* and *Cry2*. Based on a probability of detection of $p < 0.05$ on at least 7 of the 12 arrays, Storch *et al.* (2002) classified as expressed a total of 4805 genes in liver and 5120 genes in heart; among these, 4773 genes in liver and 5101 genes in heart have no missing gene expression value over all 12 time points, including 330 and 423 circadian rhythmic genes identified in heart and liver. In the following analysis, we will only concentrate on these gene that are expressed and have no missing gene expression data. Our aim is to identify among these genes those which show circadian expression patterns.

We first fit the shape-invariant model based on the data of the gene expression levels in the mouse heart tissue of the nine guide genes. Preliminary analysis of the data indicates possible attenuation in gene expression in the second day. We therefore combined data from two days into one combined periodic process and assume the period of this combined process to be 48 h. The likelihood ratio test for $\beta = 0$ resulted

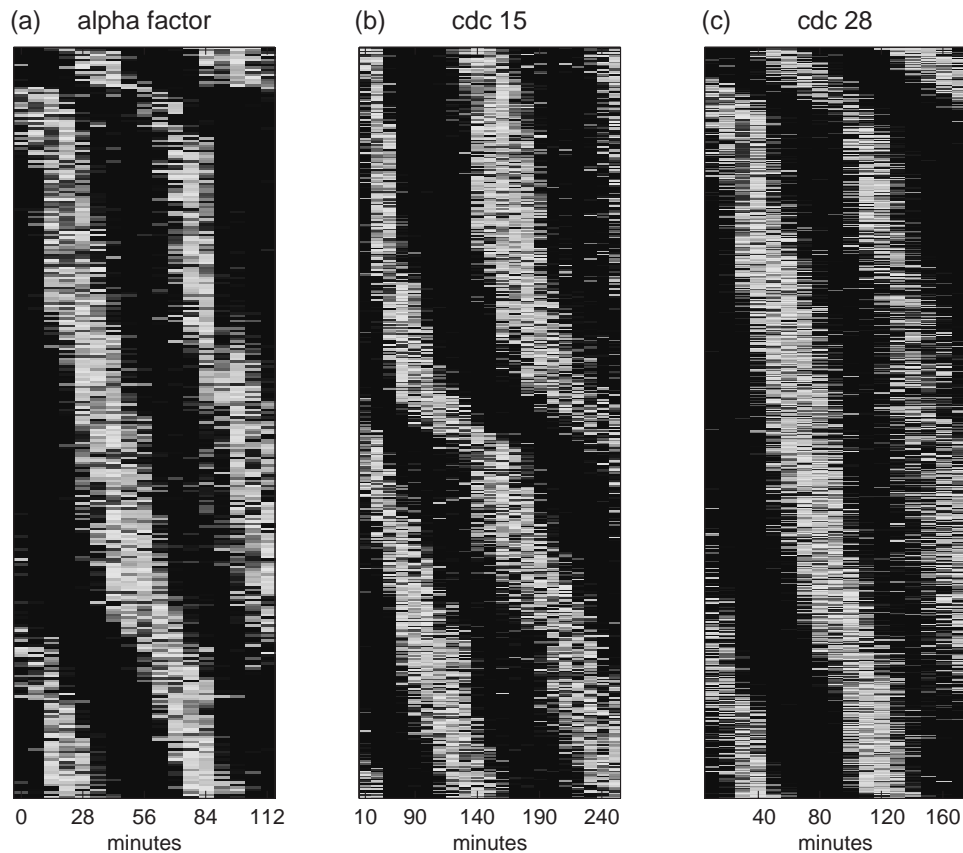


Fig. 3. Temporal profiles of the cell-cycle-regulated genes identified by the proposed methods, sorted by the estimated phases. **(a)** 297 genes identified based on the data of α -factor experiment; **(b)** 482 genes identified based on the data of *cdc15* experiment; **(c)** 623 genes identified by *cdc28* experiment. Each column represents a time point during cell cycle and each row a gene. Dark shades represent lower expression levels and light shades represent higher expression levels. Gene expression levels are standardized across the time points.

in removing two guide genes *Cry1* and *Per3* from the guide gene set by not rejecting the null hypothesis. This gave us seven guide genes. Figure S4a–g in our web supplementary materials shows the observed time-course expression data and the estimated smooth gene expression profile for the seven guide genes. Clearly, the estimated curves fit the data reasonably well, indicating that the shape-invariant model we used gave a reasonably good approximation to the observed data. It is also evident that these seven genes have different phase and different amplitude of their two-day gene expression levels. Figure S4h shows that the estimated common curve $f(t)$. It is interesting to note that although there are two clear peaks corresponding two different days, the curves are not quite symmetric. This might be due to lack of synchronization of the cells as time goes on.

Based on the estimated p -values for the 5101 test genes and the FDR procedure, for an FDR of 2.5%, we identified only very small number of genes that circadian rhythmic pattern. For a FDR of 5%, we identified 30 such genes. For a FDR of 10%, we identified 344 genes as circadian rhythmic genes

in the mouse heart, including all seven guide genes which our model is based on. The maximum p -value among these identified genes is 0.0030. As a comparison, Storch *et al.* (2002) identified 330 genes among the 5101 genes that are expressed in the heart. The image plots of the gene expression levels for the genes identified by both methods show clear periodic expression patterns (see web supplement Figure S5). However, there were only 90 genes which were identified by both methods, and it is difficult to conclude which set of genes show better periodic patterns simply based on the image plots. There were 254 genes that were identified by our methods but missed by Storch *et al.* (2002) and 240 genes that were identified by Storch *et al.* (2002) but missed by our methods (see web supplement Figure S5 for the plots of the gene expression profiles of these genes). Using a permutation procedure, Storch *et al.* (2002) estimated that 12% of their selected genes for heart can be ascribed to noise. In addition, Storch *et al.* (2002) also mentioned ‘a considerable greater prevalence of particular circadian regulation at low amplitudes, particularly in heart data’. This may partially explain the discrepancies of

the genes identified by the two methods. Further examination of the 240 genes that we missed indicate that they include many genes with very low transcriptional levels, which corresponds to very small β coefficient in our model. These genes were therefore excluded as circadian rhythmic genes by our methods. However, since the FDR = 10%, many of the genes identified by our methods could also be those with low amplitude of gene expressions. As a comparison, for FDR = 5%, our methods identified only 30 genes, including all the seven guide genes.

We performed a similar analysis for the mouse liver data set. For FDR of 2.5%, our methods identified 201 circadian rhythmic genes from a total of 4773 test genes with no missing data, including six of the seven guide genes from which our model is based on. The maximum p -value among these genes identified is 0.0010. As a comparison, Storch *et al.* (2002) identified 423 genes as circadian rhythmic genes in liver among those 4774 test genes. The image plots of these genes identified are shown in Figure S6 in the web supplement. There are 128 genes which were identified by both methods, with 73 genes identified by our methods but not by Storch *et al.* (2002) and 295 genes identified by Storch *et al.* (2002) but not by our methods (see Figure S6 in web supplement for the image plots of the observed gene expression profiles). Again for FDR of 2.5%, our methods identified a substantially smaller number of genes. For FDR of 3%, we identified 255 genes with 153 overlapping with those identified in Storch *et al.* (2002). For FDR = 4%, we identified 430 genes with 218 overlapping with those identified in Storch *et al.* (2002). Given the estimated rate of 16% of the selected genes by Storch *et al.* (2002) for liver which can be ascribed to noise and the high prevalence of circadian regulation at low amplitudes, it is not surprising to see that our procedure selected a smaller number of circadian rhythmic genes in liver.

In summary, for both heart and liver data, the discrepancies of the genes identified can be partially explained by different ways of dealing with those genes with low amplitudes of transcription. Given the noisy nature of the microarray data, formal statistical tests used in our analysis can be more advantageous in separating true signals from the underlying noises than those of other methods. Finally, the lists of the genes identified are available in the web supplement.

DISCUSSION

We have proposed a model-based method for identifying periodic expressed genes based on microarray time-course gene expression profiles. This procedure estimates the common gene expression shape based on a set of known periodically expressed genes and uses a FDR-based procedure for identifying other periodically expressed genes. We demonstrate our methods by analyzing three synchronized yeast cell cycle time-course experiments, and circadian gene expression

profiles *in vivo* of mouse heart and liver. For all five data sets, the shape-invariant model with cubic B-splines fits the gene expression data of the guide genes very well. For a FDR of 0.5%, the proposed procedure identified 86% of the known periodically expressed genes using the yeast cell cycle data set and almost all the known circadian rhythmic genes using the mouse heart and liver data sets. For all five data sets, we identified many periodically expressed genes that were not identified by previous analyses. These genes show clear periodical expression patterns in their observed data. We also observed that the cubic B-spline functions are quite flexible in modeling the observed time-course gene expression profiles.

In our analyses of the real data sets, in order to deal with the possible lack of synchronization of cells during a biological process, we combined data measured in two biological periods into one combined period and used data of guide genes to estimate the common curve in the combined period. Zhao *et al.* (2001) proposed to explicitly model the attenuation of gene expression over time by assuming a specific parametric model. Our results of yeast cell cycle data analysis are quite comparable to those obtained by Zhao *et al.* (2001). An alternative approach is to model the common expression profiles in two biological periods by using two different B-spline functions. In addition, for simplicity, in our analysis, we assumed a first-order auto-correlation structure for the error terms. This proved to be enough since no significantly serial correlations were observed in the data sets after the mean expression profiles are adjusted.

The methods presented in this paper make several assumptions and also have some limitations. First, it requires that a set of guide genes is available in order to build the model. For some biological processes, we may not have these genes. In this case, it requires data to cover several periods of the processes, and the standard spectra analysis from time series literature might be applied. Second, we assume that all the guides genes follow the same shape in their expression profile. This assumption works well for all three data sets that we analyzed. However, it may not hold for all biological processes. An alternative is to assume a mixture of several curves for the guides genes. Third, in typical microarray time-course experiments, the number of time points are usually small. In order to model the gene expression trajectory, we used simple cubic B-spline with pre-determined number and locations of knots to model the common shape function. We would expect that the B-spline function approximates most of the gene expression curves. However, in practice, one always needs to check how well the model fits the data. If we assume that the knots are equally spaced, we can use AIC or BIC technique for selecting the number of knots. An alternative to the B-spline, we can use the non-parametric spline-based approach as proposed by Wang and Brown (1996) for modeling the common function. However, the non-parametric approach of Wang and Brown (1996) is more computer-intensive in estimating the parameters. Lastly, although strictly speaking, the

FDR procedure requires that the tests are independent, both Benjamini and Hochberg (1995) and Storey and Tibshirani (2001) indicated that the procedure is still valid under dependence and can be used for identifying differentially expressed genes in the context of micorarray gene expression data.

In conclusion, we have proposed a model-based procedure for identifying genes whose expressions over time are synchronized to certain periodic biological process based on time-course microarray experiments. The methods are able to identify a set of cell-cycle-regulated genes in yeast and genes that show circadian rhythmic patterns in the mouse heart and liver tissues. The methods are quite general and can potentially be used for identifying genes which are synchronized to other important biological processes.

ACKNOWLEDGEMENTS

The authors wish to thank Dr Storch for providing the mouse gene expression data sets and the referees for many helpful comments. This research is supported in part by NIH grant ES09911.

REFERENCES

- Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.
- Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J. and Davis, R.W. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.
- Claridge-Chang, A., Wijnen, H., Naef, F., Boothroyd, C., Rajewsky, N. and Young, M.W. (2001) Circadian regulation of gene expression systems in the *Drosophila* head. *Neuron*, **32**, 657–671.
- De Boor, C. (1978) *A Practical Guide to Splines*. Springer-Verlag.
- Hastie, T., Tibshirani, B., Sherlock, G., Eisen, M., Brown, P. and Botstein, D. (1999) Imputing missing data for gene expression arrays. *Technical report*, Division of Biostatistics, Stanford University.
- Johansson, D., Lindgren, P. and Berglund, A. (2003) A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription. *Bioinformatics*, **19**, 467–473.
- Langmead, C.J., Yan, A.K., McClung, C.R. and Donald, B.Y. (2002) Phase-independent rhythmic analysis of genome-wide expression patterns. *Proceedings of the Sixth Annual International Conference on Computational Molecular Biology*. Washington, DC.
- Lawton, W.H., Sylvestre, E.A. and Maggio, M.S. (1972) Self-modeling nonlinear regression. *Technometrics*, **13**, 513–532.
- Li, H., Luan, Y., Hong, F. and Li, Y. (2002) Statistical methods for analysis of time course gene expression data. *Front. Biosci.*, **7**, a90–a98.
- Luan, Y. and Li, H. (2003) Clustering of time-course gene expression data using a mixed-effects model with splines. *Bioinformatics*, **19**, 474–482.
- Rice, J. and Wu, C. (2001) Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics*, **57**, 253–259.
- Spellman, P., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
- Storch, K., Lipan, O., Leykin, I., Viswanathan, N., Davis, F.C., Wong, W.H. and Weitz, C.J. (2002) Extensive and divergent circadian gene expression in liver and heart. *Nature*, **417**, 78–83.
- Storey, J.D. and Tibshirani, R. (2001) Estimating false discovery rates under dependence, with applications to DNA microarrays. *Technical Report 2001-28*, Department of Statistics, Stanford University.
- Wang, Y. and Brown, M.M. (1996) A flexible model for human circadian rhythms. *Biometrics*, **52**, 588–596.
- Zhao, L.P., Prentice, R. and Breden, L. (2001) Statistical modeling of large microarray data sets to identify stimulus–response profiles. *Proc. Natl Acad. Sci., USA*, **98**, 5631–5636.