BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 8 2006, pages 959–966
doi:10.1093/bioinformatics/btl017

Gene expression

Polynomial model approach for resynchronization analysis of cell-cycle gene expression data

Peng Qiu[1,*], Z. Jane Wang[2] and K. J. Ray Liu[1]

[1]Department of Electrical and Computer Engineering, University of Maryland, College Park, USA and
[2]Department of Electrical and Computer Engineering, University of British Columbia, Canada

Received on October 31, 2006; revised on January 19, 2006; accepted on January 21, 2006
Advance Access publication January 24, 2006
Associate Editor: David Rocke

ABSTRACT

Motivation: Identification of genes expressed in a cell-cycle-specific periodical manner is of great interest to understand cyclic systems which play a critical role in many biological processes. However, identification of cell-cycle regulated genes by raw microarray gene expression data directly is complicated by the factor of synchronization loss, thus remains a challenging problem. Decomposing the expression measurements and extracting synchronized expression will allow to better represent the single-cell behavior and improve the accuracy in identifying periodically expressed genes.

Results: In this paper, we propose a resynchronization-based algorithm for identifying cell-cycle-related genes. We introduce a synchronization loss model by modeling the gene expression measurements as a superposition of different cell populations growing at different rates. The underlying expression profile is then reconstructed through resynchronization and is further fitted to the measurements in order to identify periodically expressed genes. Results from both simulations and real mircorarray data show that the proposed scheme is promising for identifying cyclic genes and revealing underlying gene expression profiles.

Availability: Contact the authors.

Contact: qiupeng@umd.edu

Supplementary information: Supplementary data are available at: http://dsplab.eng.umd.edu/~genomics/syn/

1 INTRODUCTION

A central goal of molecular biology is to use genetic data in order to understand the fundamental cyclic systems, such as regulatory network in yeast cell-cycle (Lee *et al*., 2002). Recent advances in high-throughput gene expression data acquisition technologies, such as microarrays, provide a rich opportunity for achieving this goal (Moor, 2001). The first critical task in understanding such cyclic systems is to identify the related genes which are periodically expressed during the cell-cycle. In the current technologies, most expression data are measured based on a population of cells which are synchronized to exhibit similar behaviors (Spellman *et al*., 1998). However, even with the most advanced synchronization method, maintaining a tight synchronization population even over a couple of cycles is a challenging research issue, since continuous synchronization loss is gradually observed owing to the diversity of individual cell growth rates (Shedden and Cooper, 2002). Therefore, in addition to the noise effect on the measurements, a significant difficulty in identifying cell-cycle regulated genes by analyzing microarray gene expression data arises from synchronization loss. Direct periodicity testing on the expression measurements could be misleading or fail due to the fact that the expression values measured are contributed by a mixed cell populations growing at different rates.

Several approaches for identifying cell-cycle regulated genes, when taking into consideration the issue of synchronization loss, have been proposed in the literature. They can be divided into two major categories, differentiated by the absence or presence of other complementary information besides gene expression data. Most studies in the literature belong to the former category, which relies solely on expression data. Fourier analysis algorithm was employed for synchronization test in Shedden and Cooper (2002), Johansson *et al*. (2003) and Whitfield *et al*. (2002). The authors presented an exact statistical test to identify periodically expressed genes by distinguishing periodicity from random processes in Wichert *et al*. (2004). In Lu *et al*. (2004), a periodic-normal mixture (PNM) model was proposed to fit transcription profiles of periodically expressed (PE) genes and a principled statistical estimation approach was developed for estimating the periodicity of gene expressions. In the second category, an algorithm combining budding index and gene expression data was proposed recently to deconvolve expression profiles in Bar-Joseph *et al*. (2004). Regardless of these developments, efforts are still needed to accurately identify cyclic genes and recover a more accurate gene profile compared with the current expression measurements.

The goal of this paper is to develop an efficient scheme for identifying periodically expressed genes and reconstructing the underlying gene expression profiles by estimating the effects of synchronization loss. The main contributions of this paper are 2-fold.

- We propose a synchronization loss model by representing the gene expression measurements as a superposition of different cell populations growing at different rates, because the model can mimic the synchronization loss observed in microarray experiments and is easy to implement. Also, we develop a model-based estimation algorithm to reconstruct the underlying single-cell gene expression profiles. In previous studies, the single-cell expression profile is often assumed to be sinusoids.

*To whom correspondence should be addressed.

However, the proposed algorithm does not require that assumption. It is able to handle a much larger variety of single-cell expression profiles.

- Using the fitting residue error as criteria, we explore a supervised learning scheme for identifying the cell-cycle regulated genes. The performance of the proposed scheme are examined via both simulations and real microarray gene expression data of *Saccharomyces cerevisiae*.

The organization of the rest paper is as follows. We start by introducing a synchronization loss model and our formulation. After that, a cyclic gene identification scheme is proposed. In Sections 4 and 5, the proposed scheme is examined and compared with two previous studies. From the results, we conclude that the proposed scheme is promising in improving quality of gene expression time series data.

2 SYSTEM MODEL AND FORMULATION

2.1 A mixture model for synchronization loss

Even with the best synchronization method currently available, cells begin to lose their synchronization in a short time. Therefore, we propose to model the observed gene expression data as a superposition from a mixed population of cells growing at slightly different rates as

$$y_i(t) = \sum_{m=0}^{N} \beta_m x_i(\rho_m t), \tag{1}$$

where $y_i(t)$ is the observed expression of gene i at the time t; $x_i(t)$ is the underlying single-cell expression profile; ρ_m represents the relative growth rates of cells with respect to standard cell-cycle; β_m represents the percentage of cells with a growth rate ρ_m, and it is assumed to be constant in one series of measurements. Although ρ_m can take continuous values in experiment, due to the limited size of microarray data, ρ_m is approximated by $N + 1$ components. Because of the different growth rates in experiment cell population, even if a gene is cell-cycle regulated, the measured expression may not exhibit clear periodicity. Therefore, it is difficult to accurately detect periodically expressed genes and distinguish them from non-periodically expressed genes based on the noisy microarray data.

Note that in Equation (1), for gene i, from the underlying expression profile $x_i(t)$ to the observation $y_i(t)$, the distortion is dictated by β_m and ρ_m, which describes the synchronization loss status of the whole cell populations. We propose to utilize some common properties of all genes to extract underlying expression profiles from the observations.

2.2 An inverse model for synchronization loss

Since the underlying expression profile $x_i(t)$ is unknown, we propose to re-write Equation (1) into the following form,

$$x_i(t) = \sum_{m=0}^{M} a_m y_i(c_m t) = [a_0, a_1, \ldots, a_M] \begin{bmatrix} y_i(c_0 t) \\ y_i(c_1 t) \\ \vdots \\ y_i(c_{Mt}) \end{bmatrix}, \tag{2}$$

where the underlying expression $x_i(t)$ is represented by the superposition of M multiple scaled versions of the observation $y_i(t)$. Parameters a_ms and c_ms describe the coefficient and scaling factor of each component. An intuitive explanation for Equation (2) is motivated by the inverse relationship between finite impulse response (FIR) filters and infinite impulse response (IIR) filters, since the structure of (1) is quite similar to that of FIR filters. Equation (1) describes an FIR-like operation which transforms $x_i(t)$ to $y_i(t)$. In order to perform the inverse transformation, an IIR-like operation is required. If the range of c_m is properly chosen, Equation (2) can be regarded as a truncated IIR-like operation, which is an approximate inverse of the FIR-like operation in Equation (1). Therefore, Equations (1) and (2) relate $x_i(t)$ and $y_i(t)$ in approximately the same way.

It is worth mentioning that the parameters a_ms and c_ms depend solely on β_ms and ρ_ms. They are common constants for all genes. Thus, we propose to utilize this common property of all genes to extract underlying expression profiles.

Equations (1) and (2) are not mathematically equivalent in general. However, if $x_i(t)$ is polynomial, Equations (1) and (2) can be equivalently represented. In this paper, we are particularly interested in the case of polynomials, since polynomial is a common tool for data fitting (Stoer and Bulirsch, 1991). As shown in the literature, polynomials are often successfully used to fit the time-series gene expression data (Bar-Joseph *et al.*, 2004).

Suppose $x_i(t)$ is a polynomial of order K such that

$$x_i(t) = \sum_{k=0}^{K} b_k t^k = [1, 1, \ldots, 1] \begin{bmatrix} b_0 t^0 \\ b_1 t^1 \\ \vdots \\ b_K t^K \end{bmatrix}, \tag{3}$$

with b_ks being the polynomial coefficients. Then, according to Equation (1), $y_i(t)$ can be expressed as

$$y_i(t) = \sum_{m=0}^{N} \beta_m x_i(t\rho_m) = [\underline{\beta}^{\mathrm{T}} \underline{\rho}^0, \underline{\beta}^{\mathrm{T}} \underline{\rho}^1, \ldots, \underline{\beta}^{\mathrm{T}} \underline{\rho}^{\mathrm{k}}] \begin{bmatrix} b_0 t^0 \\ b_1 t^1 \\ \vdots \\ b_K t^K \end{bmatrix}, \tag{4}$$

where $\underline{\beta} = [\beta_0, \beta_1, \ldots, \beta_N]^{\mathrm{T}}$, and $\underline{\rho}^k = [\rho_0^k, \rho_1^k, \ldots, \rho_N^k]^{\mathrm{T}}$. Similarly, since

$$y_i(ct) = [\underline{\beta}^{\mathrm{T}} \underline{\rho}^0 c^0, \underline{\beta}^{\mathrm{T}} \underline{\rho}^1 c^1, \ldots, \underline{\beta}^{\mathrm{T}} \underline{\rho}^K c^K] \begin{bmatrix} b_0 t^0 \\ b_1 t^1 \\ \vdots \\ b_K t^K \end{bmatrix}, \tag{5}$$

if we pick up multiple scaled version $y_i(c_m t)$ of the observation $y_i(t)$, we can write them together into the following matrix form,

$$\begin{bmatrix} y_i(c_0 t) \\ y_i(c_1 t) \\ \vdots \\ y_i(c_M t) \end{bmatrix} = \begin{bmatrix} \underline{\beta}^{\mathrm{T}} \underline{\rho}^0 c_0^0 & \underline{\beta}^{\mathrm{T}} \underline{\rho}^1 c_0^1 & \cdots & \underline{\beta}^{\mathrm{T}} \underline{\rho}^K c_0^K \\ \underline{\beta}^{\mathrm{T}} \underline{\rho}^0 c_1^0 & \underline{\beta}^{\mathrm{T}} \underline{\rho}^1 c_1^1 & & \underline{\beta}^{\mathrm{T}} \underline{\rho}^K c^K \\ \vdots & \vdots & \ddots & \vdots \\ \underline{\beta}^{\mathrm{T}} \underline{\rho}^0 c_M^0 & \underline{\beta}^{\mathrm{T}} \underline{\rho}^1 c_M^1 & \cdots & \underline{\beta}^{\mathrm{T}} \underline{\rho}^K c_M^K \end{bmatrix} \begin{bmatrix} b_0 t^0 \\ b_1 t^1 \\ \vdots \\ b_K t^K \end{bmatrix}. \tag{6}$$

Now, if we want to find a set of coefficients a_ms to represent the underlying expression profile $x_i(t)$ as in Equation (2), based on Equations (3) and (6), we will require coefficients a_ms to satisfy the following equation,

$$[a_0, a_1, \ldots, a_M] \begin{bmatrix} \underline{\beta}^{\mathrm{T}} \underline{\rho}^0 c_0^0 & \underline{\beta}^{\mathrm{T}} \underline{\rho}^1 c_0^1 & \cdots & \underline{\beta}^{\mathrm{T}} \underline{\rho}^K c_0^K \\ \underline{\beta}^{\mathrm{T}} \underline{\rho}^0 c_1^0 & \underline{\beta}^{\mathrm{T}} \underline{\rho}^1 c_1^1 & \cdots & \underline{\beta}^{\mathrm{T}} \underline{\rho}^K c_1^K \\ \vdots & \vdots & \ddots & \vdots \\ \underline{\beta}^{\mathrm{T}} \underline{\rho}^0 c_M^0 & \underline{\beta}^{\mathrm{T}} \underline{\rho}^1 c_M^1 & \cdots & \underline{\beta}^{\mathrm{T}} \underline{\rho}^K c_M^K \end{bmatrix} = [1, 1, \ldots, 1]. \tag{7}$$

Note that in the matrix in Equation (7), every element in one column shares a common factor. If we pull out the common factor, the remaining part will be a Vandermonde matrix. And the Vandermonde matrix is of full rank $\min\{M, K\}$, as long as different scaled (c_m) observations are considered, shown in Equation (8). We can show that, as long as M is greater than or equal to K, there exists at least one solution to equation (7). That is, there exists at least one set of coefficients a_ms that satisfies

Equation (7). In this case, Equations (1) and (2) are mathematically equivalent.

$$\begin{bmatrix} \underline{\beta}^{\mathrm{T}}\underline{\rho}^{0}c_{0}^{0} & \underline{\beta}^{\mathrm{T}}\underline{\rho}^{1}c_{0}^{1} & \dots & \underline{\beta}^{\mathrm{T}}\underline{\rho}^{K}c_{0}^{K} \\ \underline{\beta}^{\mathrm{T}}\underline{\rho}^{0}c_{1}^{0} & \underline{\beta}^{\mathrm{T}}\underline{\rho}^{1}c_{1}^{1} & \dots & \underline{\beta}^{\mathrm{T}}\underline{\rho}^{K}c_{1}^{K} \\ \vdots & \vdots & \ddots & \vdots \\ \underline{\beta}^{\mathrm{T}}\underline{\rho}^{0}c_{M}^{0} & \underline{\beta}^{\mathrm{T}}\underline{\rho}^{1}c_{M}^{1} & \dots & \underline{\beta}^{\mathrm{T}}\underline{\rho}^{K}c_{M}^{K} \end{bmatrix} = \begin{bmatrix} c_{0}^{0} & c_{0}^{1} & \dots & c_{0}^{K} \\ c_{1}^{0} & c_{1}^{1} & \dots & c_{1}^{K} \\ \vdots & \vdots & \ddots & \vdots \\ c_{M}^{0} & c_{M}^{1} & \dots & c_{M}^{K} \end{bmatrix} \begin{bmatrix} \underline{\beta}^{\mathrm{T}}\underline{\rho}^{0} & 0 & \dots & 0 \\ 0 & \underline{\beta}^{\mathrm{T}}\underline{\rho}^{1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \underline{\beta}^{\mathrm{T}}\underline{\rho}^{K} \end{bmatrix}. \quad (8)$$

In the above argument, the underlying expression $x_i(t)$ does not assume periodicity. However, in this study, the most interested expression signal is cell-cycle regulated, i.e. periodic,

$$x_i(t) = \sum_{k=0}^{K} b_k (t \bmod T)^k, \quad (9)$$

where mod means the modulus operator that gives the reminders after division. In microarray time series experiment, the range of relative growth rate ρ_m is not large. With c_m carefully chosen, although periodic, the above argument holds for most of the cell-cycle data. In the following section, we will demonstrate that under such a periodic condition Equation (2) is a fine approximation of the inverse of Equation (1).

2.3 Formulation for estimating a_ms

For cell-cycle regulated genes, because of periodicity, $x_i(t) = x_i(t+T)$, from Equation (2), the observations and parameters a_ms and c_ms are related as follows:

$$\sum_{m=1}^{M} a_m [y_i(c_m t) - y_i(c_m(t+T))] = 0. \quad (10)$$

Denote $\underline{y}_i(t) = [y_i(c_1 t) - y_i(c_1(t+T)), \dots, y_i(c_M t) - y_i(c_M(t+T))]^{\mathrm{T}}$ and $\underline{a} = [a_1, \dots, a_M]^{\mathrm{T}}$. Equation (10) can be re-written as

$$\underline{y}_i(t)^{\mathrm{T}}\,\underline{a} = 0. \quad (11)$$

Note that, we can evaluate Equation (11) at different time points (as long as the time-series data allows). Also, all cell-cycle regulated genes satisfy Equation (11). So, the estimation of a_m parameters can be formulated as a constrained least square problem,

$$[\underline{y}_i(t_1), \dots, \underline{y}_i(t_n), \underline{y}_j(t_1), \dots, \underline{y}_j(t_n), \dots]^{\mathrm{T}}\underline{a} = 0, \quad (12)$$

$$\text{subject to } \sum_{m=1}^{M} a_m = 1, \quad (13)$$

where genes $i, j, \dots$ are cell-cycle regulated genes; $t_1, \dots, t_n$ are the measurement time points that satisfies $(t_n + T)c_m < 2T$, for all $m = 1, \dots, M$, since in the current *S.cerevisiae* time-series gene expression data, only two cell-cycles are available. The value of n in Equation (12) is quite small, e.g. 4 or 5, depending on parameters c_m and the experiment sampling rate. Therefore it is important to use many cell-cycle regulated genes together to estimate the coefficients a_ms reliably. In this formulation, c_m are assumed known. Since in real experiment, the growth rate of different cells differ slightly. The range of the relative growth rate ρ_m is not large. In later simulations, we will show that, it is accurate enough to choose c_m to cover the range from 0.6 to 1.4. The fixed-summation constraint in Equation (13) is chosen to avoid the trivial 0-vector solution, i.e. $\underline{a} = \underline{0}$.

2.4 Fitting residue criterion

After estimating a_ms, the model in (2) is used to reconstruct the underlying periodical component $x_i(t)$ for every gene. In order to detect cell-cycle regulated genes, a criterion is needed to justify the question whether the extracted signal is the underlying periodical expression profile of a cell-cycle regulated gene, or it is the periodical component from a non-cell-cycle regulated gene. We propose a criteria based on the model in (1), using the extracted periodical signal to fit the observations. The fitting residue will serve as the criterion in detecting cell-cycle regulated genes. For a particular gene, if the fitting residue is sufficiently small, compared with a threshold, then the extracted signal could lead to the measurements owing to synchronization loss, which means the gene is highly likely to be cell-cycle regulated. On the other hand, if the fitting residue is large, then the extracted periodical signal is not closely related to experimental observation, which means the gene is more likely to be non-cell-cycle regulated. In the proposed identification scheme, the threshold of fitting residue is dynamically determined during iterations. Details are described in Sections 3 and 5.

3 THE IDENTIFICATION SCHEME

Based on the synchronization loss model and estimation approach described in Section 2, we further proceed to identify the cyclic genes. The scheme described in this section is a supervised learning scheme, since it requires an initial training set which consists of cell-cycle regulated genes previously identified by traditional biology experiments. Specifically, we propose an iterative framework to purify the training set and detect cyclic genes simultaneously. The main steps in the proposed iterative framework is described as follows:

(1) Define initial training set as cell-cycle regulated genes previously identified by traditional methods.

(2) Apply the proposed model on training set to estimate the parameters a_ms, and extract the underlying periodical signal for every gene in the training set.

(3) Based on the extracted signal $x_i(t)$, fit it to the observation model in (1). According to the fitting residue criterion, remove some non-periodically expressed genes from the training set. Then, re-estimate the parameters a_ms using the training set and use the estimated a_ms to extract periodical signal for every gene in the testing set.

(4) According to the fitting residue criterion, include some periodically expression genes into the training set. Then go back to Step 2.

Note that, under this framework, in order to purify the training set and detect the periodically expressed genes correctly, the criteria for removing and including genes in Steps 2 and 4 should be carefully designed and fine tuned for each dataset. The ultimate goal of Steps 2 and 4 is to find a set of cyclic genes as prior knowledge, such that the cyclic genes identified by the proposed scheme do not violate the prior knowledge, or maximally support the prior knowledge. It is a difficult optimization problem with numerous possible solutions. The proposed scheme, although heuristic in updating the training set, yields satisfactory results as will be demonstrated in what follows.

4 SIMULATIONS

In this section, we simulate time-series expression data with synchronization loss for both periodically expressed genes and

non-periodically expressed genes. The proposed method is used to resynchronize the simulated data and identify periodically expressed genes. To evaluate the performance of the proposed method, we compare it with the methods studied in Spellman *et al*. (1998) and Lu *et al*. (2004). We also perform sensitivity analysis to examine the robustness of the proposed method.

4.1 Simulations based on sinusoids

In this subsection, we simulate time-series expression data for 100 periodically expressed genes and 600 non-periodically expressed genes. The underlying single-cell periodical expression profile for cyclic gene i is generated by a linear combination of four sinusoids with random phases,

$$x_i(t) = \sum_{j=1}^{4} \lambda_{ij} \sin\left(\frac{2\pi j}{T} t + \phi_{ij}\right), \tag{14}$$

where the period T is set to be 60 min, same as the cell-cycle duration in the alpha experiment in Spellman *et al*. (1998). The parameter λ_{ij} is randomly chosen, different for each gene. ϕ_{ij} represents the random phase, which is uniformly distributed on $[0,2\pi)$. For the 600 non-cyclic genes, their underlying expressions are obtained through random permutations of expressions of cyclic genes.

For each gene, we simulate the synchronization loss by

$$y_i(t) = \beta_1 x_i(t * s) + \beta_2 x_i(t) + \beta_3 x_i(t * f) + v, \tag{15}$$

where $f = 1.3$ and $s = 0.7$ represent the relative growth rates. β_m is randomly generated, representing the percentage of cells growing at different rates. v represents the microarray measurement noise. It is modeled as a zero-mean Gaussian random variable. Its variance is chosen to make the signal to noise ratio (SNR) to be 5.716 dB, which is close to the SNR value estimated from the alpha dataset in Spellman *et al*. (1998). Equation (15) is applied to all genes, representing the common synchronization status of the cell populations. In the simulations, measurements are taken every 6 min from 0 to 120 min, yielding 21 time points in total.

In the simulation, 50 cyclic genes are assumed known, in order to form the initial training set. The testing set contains the left 650 genes. For a particular choice of c_m, by applying the proposed model, a_m parameters are estimated, the underlying periodical signals for all genes are extracted, and the fitting residue criterion is examined.

The parameter M is set to be $M = 7$. As mentioned in Section 2.2, we need to choose M to be larger than or equal to K. Since the exact value of K does not affect the proposed method as long as $K \leq M$, therefore, with $M = 7$, the proposed method can handle all polynomials with $K \leq 7$. And we know, that the seventh order polynomials can generate a large variety of curves, with up to six peaks and valleys. We believe the current parameters-setting can sufficiently model gene expression profiles.

As mentioned earlier, c_m should be chosen properly, in order to extract underlying expression profiles accurately. In Table 1, different choices of c_m are examined. To ensure a fair comparison, with M set to be 7, the values of c_m are chosen to be uniformly spaced in tested range. In the fitting residue criterion, ρ_m is set to be $[0.7, 0.8, \ldots, 1.3]$. From Table 1, we can see that, different choice of c_m leads to different fitting residues for both cyclic genes and non-cyclic genes. As the range of c_m increases, the fitting residues for cyclic genes tend to decrease first, and then increase. This observation can be intuitively explained by the trade-off between errors owing to the model-complexity and the data size. In one hand, from the implication of FIR and IIR filters, owing to the larger range of c_m considered, the truncation error will be smaller. However, if the range of c_m is too large, because of the limited size of time series data, the number of available time points n in Equation (11) will be small. Less training data will cause the fitting residues to increase. Therefore, based on Table 1, we choose the range of c_m to be [0.6, 1.4], since with this choice the average fitting residue for cyclic genes is small and the difference between cyclic genes and non-cyclic genes is large, resulting in a good detection performance.

Table 1. Comparison of the normalized average fitting residues for cyclic and non-cyclic genes

Range of c_m	Average fitting residue cyclic genes	Average fitting residue non-cyclic genes	Difference
[0.9, 1.1]	0.4424	0.9373	0.4949
[0.8, 1.2]	0.4164	0.9560	0.5396
[0.7, 1.3]	0.3993	0.9583	0.5590
[0.6, 1.4]	0.3917	0.9355	0.5437
[0.5, 1.5]	0.4052	0.9478	0.5426
[0.4, 1.6]	0.4354	0.9865	0.5511

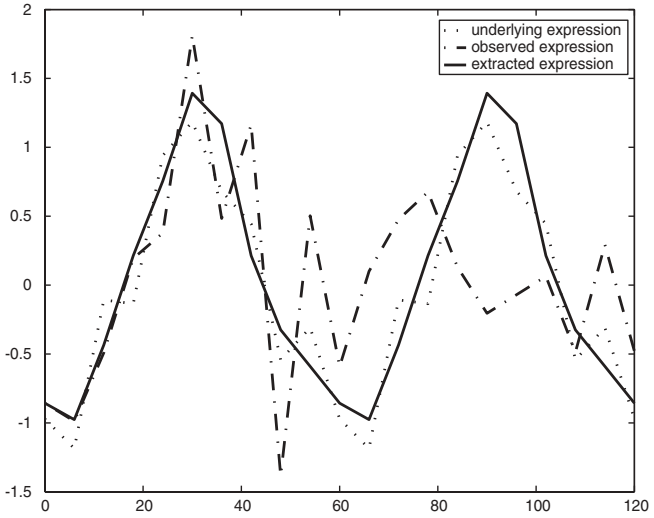

Fig. 1. The simulated sinusoid underlying periodical expression, experiment observation and extracted expression of one simulated gene.

After determining the choice of c_ms, the proposed model is applied to estimate parameters a_ms based on the training set, and extract the underlying periodical signals for genes in the training set. Figure 1 gives a typical example of genes in the training set. Although there is clear difference between the underlying periodical expression and the simulated observation, based on the proposed method, the extracted expression is quite similar to the underlying periodical expression.

Based on the a_m and c_m parameters, the proposed model is applied to extract periodical signal components for all genes, and the fitting

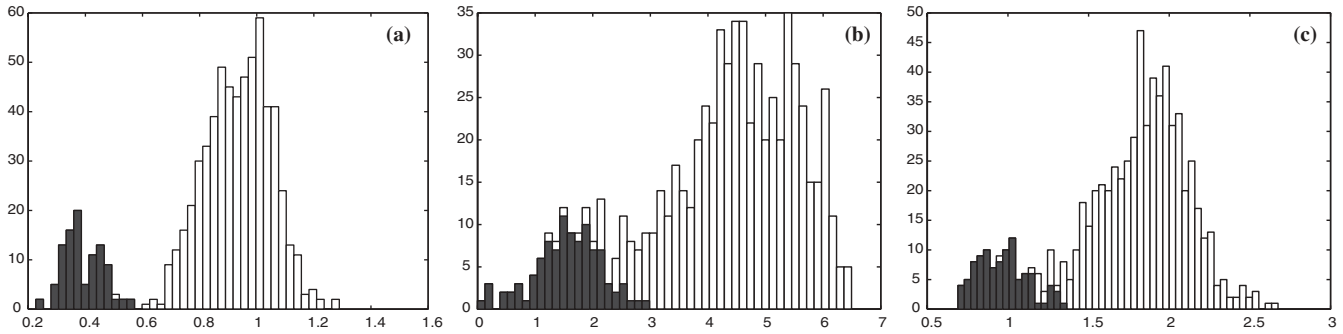

Fig. 2. The histogram of fitting residues for all genes, with the shaded area being the histogram of the 100 cyclic genes. The horizontal axis represents fitting residue, and the vertical axis represents number of genes with certain value of fitting residue. (**a**) Shows the result of the proposed method. (**b**) The result of the Fourier analysis used in Spellman *et al*. (1998). (**c**) Shows the upper bound of results from method in Lu *et al*. (2004).

residue criterion is examined. In Figure 2a, the histogram of all genes' fitting residues is shown, where the shaded part corresponds to the 100 cyclic genes. We can see that the cyclic genes have smaller fitting residues, while non-cyclic genes yield larger fitting residues statistically. Therefore, this clear separation between these two groups of genes leads to the accurate identifications of cyclic genes.

In order to examine the identification performance of the proposed method, we compare it with two previous works, Spellman *et al*. (1998) and Lu *et al*. (2004), by applying them to the same simulated time series data. In Spellman *et al*. (1998), Fourier analysis is applied to calculate the energy of the periodical components for each gene. The energy serves as a metric to identify cyclic genes. From Figure 2b, we can see that, this method can identify cyclic genes with small outage. However, its performance is worse than that of the proposed method. In Lu *et al*. (2004), a PNM model is proposed, where a probabilistic (Gaussian) distribution and Fourier analysis are combined to model the synchronization loss. Before identifying cyclic genes, the parameters of the Gaussian distribution have to be estimated. In our implementation, we skip the parameter estimation step by feeding the actual parameter values into the PNM model. Therefore, Figure 2c shows the performance upper method in Lu *et al*. (2004), which is close to that of the proposed method. However, it is worth mentioning that the PNM-based method is admittedly sensitive to the parameter estimates of the Gaussian distribution.

In Table 2, we present the results in Figure 2 in a more quantitative fashion. We employ the Neyman–Pearson framework in detection theory (Poor, 1994). During comparison, we fix the probability of correctly detecting cyclic genes and examine the probability of false positive of different methods. That is, under the condition that certain amount of cyclic genes are correctly detected, how many non-cyclic genes will be falsely detected as cyclic. From Table 2, we can see that, when fixing the probability of detection, the proposed method has much less false positives, compared with the two previous studies.

In this subsection, the time series are simulated with the underlying signal $x_i(t)$ being sinusoids. Together with the fact that Fourier analysis is employed, both previous works have nice performance in identifying cyclic genes. However, if the underlying signal is based on polynomials, the result could be different.

Table 2. Comparison of the proposed method and two previous studies

Probability of detection	False positive of proposed method	False positive of Spellman *et al*. (1998)	False positive of Lu *et al*. (2004)
0.75	0	0.0741	0.0132
0.80	0	0.0805	0.0123
0.85	0	0.1053	0.0116
0.90	0	0.1262	0.0217
0.95	0	0.1518	0.1121
1.00	0.01	0.2857	0.1597

When the probability of correctly detecting cyclic genes is fixed, we compare the probability of false positive, which means the probability of detecting a non-cyclic gene as cyclic.

4.2 Simulation based on polynomials

In this subsection, we simulate time-series expression data based on polynomial models. Again, 100 cyclic genes and 600 non-cyclic genes are simulated. The underlying periodical expression profile for cyclic gene i is generated by polynomials of order $K = 6$,

$$x_i(t) = \sum_{k=0}^{K=6} a_k (t \bmod T)^k, \tag{16}$$

where the period T is set to be 60 min, same as the cell-cycle duration in the alpha experiments. The parameter a_k is randomly chosen in $[-1, 1]$, different for each gene. For the 600 non-cyclic genes, the underlying expressions are obtained through random permutations as in previous subsection.

For each gene, we simulate the synchronization loss by Equation (15). All parameters are set to be the same as previous subsection. A total of 50 cyclic genes are assumed known, forming the training set. For a particular choice of c_m, by applying the proposed model, a_m parameters are estimated based on the training set, the underlying periodical signals for all genes are extracted, and the fitting residue criterion is examined. Again, M is set to be 7, and different choices of c_m are examined. From Table 3, similar result is observed. We choose the range of c_m to be [0.6, 1.4], because the average fitting residue for cyclic genes is small, and the difference between cyclic genes and non-cyclic genes is large.

Table 3. Comparison of the normalized average fitting residues for cyclic and non-cyclic genes

Range of c_m	Average fitting residue cyclic genes	Average fitting residue non-cyclic genes	Difference
[0.9, 1.1]	0.2240	0.9439	0.7199
[0.8, 1.2]	0.2164	1.0267	0.8104
[0.7, 1.3]	0.2157	1.0256	0.8099
[0.6, 1.4]	0.2284	1.0567	0.8283
[0.5, 1.5]	0.2455	1.1299	0.8845
[0.4, 1.6]	0.3411	1.0531	0.7120

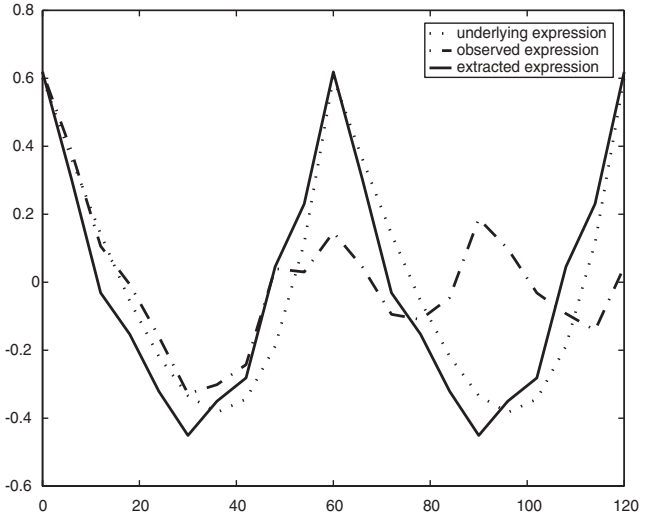

Fig. 3. The simulated polynomial underlying periodical expression, experiment observation and extracted expression of one simulated gene.

Figure 3 is a typical example of genes in the training set. We can see that the simulated observations is quite different from the underlying periodical expression profile. Owing to synchronization loss, the observed time-series does not exhibit a clear periodicity, especially in the second cycle. From poorly synchronized observations, the proposed method can successfully recover the underlying periodical expression profile.

Based on the estimates of a_ms and c_ms, the proposed model is applied to extract periodical signal components for all genes, and the fitting residue criterion is examined. In Figure 4 a the histogram of residues shows that the cyclic genes and non-cyclic genes are well separated, meaning that the proposed method can successfully identify the cyclic genes.

The methods in Spellman *et al*. (1998) and Lu *et al*. (2004) are also examined in this subsection, with results shown in Figure 4b and 4c. From these figures, we note that both previous methods failed to separate cyclic and non-cyclic genes in the case that underlying expression profiles being polynomials. Similar to the previous subsection, the result is shown in a more quantitative way, in Table 4. As it is easy to see, the proposed method outperforms previous studies in the simulation based on polynomials. It is encouraging to see that the proposed method works well for a much larger variety of the underlying single-cell expressions.

4.3 Sensitivity analysis

In our discussions so far, the standard cell-cycle duration T is assumed to be known as a prior knowledge. However, the cell-cycle duration may vary because of various environmental and experimental factors. In this subsection, we examine the performance of the proposed method when inexact prior knowledge of the cell-cycle duration T is considered.

The sensitivity analysis is conducted based on the simulated data by sinusoids. In the simulated data, the true cell-cycle length is $T = 60$. However, when applying the proposed method, we do not know the cell-cycle length exactly as prior knowledge. In Figure 5, we can see that, when the prior knowledge is inexact, the separation of fitting residues between cyclic and non-cyclic genes is not affected much. In Table 5, we quantitatively examine the sensitivity of the proposed method in terms of probability of detection and false positive. In Table 5, each row corresponds to a certain requirement of probability of detection; each column corresponds to a case where certain value of T is taken as prior knowledge; and each element is the probability of false positive. From this table, as long as we do not require probability of detection to be extremely high (i.e. 100%), only when the prior knowledge is significantly different from the truth (i.e. the prior $T \leq 40$ or $T \geq 70$), will the performance degrade severely. This simulation result demonstrates the robustness of the proposed method with respect to the cell-cycle duration.

5 REAL DATASETS

In this study, three real datasets are investigated, alpha, cdc15 in (Spellman *et al*., 1998) and cdc28 in (Cho *et al*., 1998). From Spellman *et al*. (1998), 93 cell-cycle regulated genes previously identified by traditional methods are selected as initial training set. Since there is no guarantee that all those 93 genes will behave periodically in a particular experiment, we employ the iterative framework to purify the training set and identify cyclic genes simultaneously. During each iteration, we adopt simple removing and including criterion in Steps 2 and 4. In Step 2, the size of training set is reduced to half in order to purify the training set. In Step 4, 200 genes with smallest fitting residues are included into the training set. In this way, we hope to purify the training set. As mentioned before, the ultimate goal of Steps 2 and 4 is to find a set of cyclic genes as prior knowledge, such that the identified cyclic genes identified by the proposed method do not violate the prior knowledge, or maximally support the prior knowledge. It is a quite difficult optimization problem, with numerous possible solutions. However, with the simple criterion we adopted, this goal can be achieved within several iterations (5–10). As results, the histograms of fitting residues for the alfa, cdc15 and cdc28 datasets are shown in Figure 6, where the identified cyclic genes in the training set have small fitting residues.

To make a fair comparison with Spellman *et al*. (1998) and Lu *et al*. (2004), 800 genes with smallest fitting residues are identified as cyclic genes. In Figure 7, a Venn diagram showing the overlap of genes identified by different studies. The proposed method and Spellman *et al*. (1998) is 403; the intersection between proposed method and Lu *et al*. (2004) is 433; the intersection between Spellman *et al*. (1998) and Lu *et al*. (2004) is 541; the intersection among all three studies is 355. It is encouraging to see the large overlaps illustrated in Figure 7, an indication of consistency of the

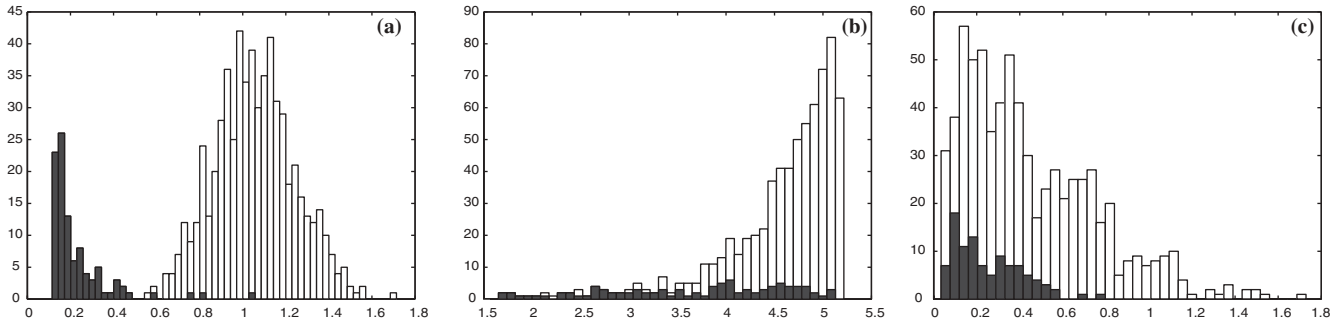

Fig. 4. Simulation based on polynomials. The histogram of fitting residues for all genes, with the shaded area being the histogram of the 100 cyclic genes. The horizontal axis represents fitting residue, and the vertical axis represents number of genes with certain value of fitting residue. (**a**) Shows the result of the proposed method. (**b**) The result of the Fourier analysis used in Spellman *et al*. (1998). (**c**) Shows the upper bound of results from method in (Lu *et al*., 2004).

Table 4. Comparison of the proposed method and two previous studies

Probability of detection	False positive of proposed method	False positive of (Spellman *et al*., 1998)	False positive of (Lu *et al*., 2004)
0.75	0	0.6622	0.7768
0.80	0	0.6887	0.7838
0.85	0	0.7028	0.7870
0.90	0	0.7443	0.7897
0.95	0	0.7765	0.7894
1.00	0.7375	0.8415	0.8353

When the probability of correctly detecting cyclic genes is fixed, we compare the probability of false positive, which means the probability of detecting a non-cyclic gene as cyclic.

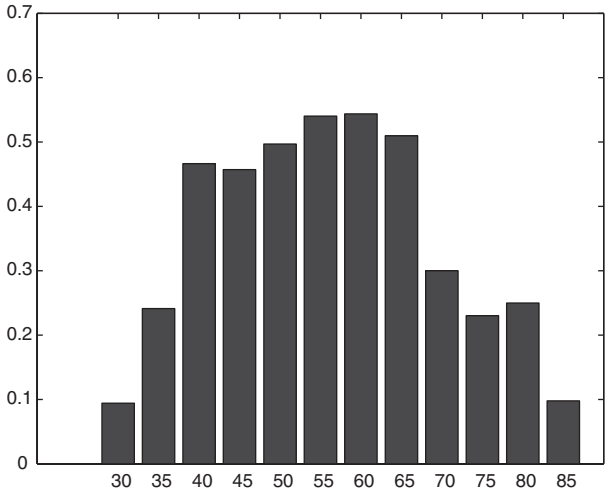

Fig. 5. The horizontal axis is the prior knowledge of cell-cycle length, though it may not be the true cell-cycle length $T = 60$. The vertical axis is the difference of fitting residues between cyclic and non-cyclic genes.

proposed method to the previous researches. In Supplementary Material, we show some examples of genes identified by both the proposed method, (Spellman *et al*., 1998), and traditional experimental methods. Both the observed expression and extracted expression are shown. We can see that, for the cyclic genes that already exhibit periodical expression, the extracted expression is closed to experiment observed expression. And for the cyclic genes that do not exhibit periodical expression, the proposed method can recover the periodicity.

Although the genes identified by the proposed method have large overlap with those of the previous studies, it is interesting to examine the non-overlapping genes identified by the proposed method, but not identified in the previous studies, neither Spellman *et al*. (1998) nor Lu *et al*. (2004). In the Supplementary Material, some examples are shown. Since both previous studies relied on Fourier analysis, genes without clear periodicity may not be identified. However, the proposed method may be able to identify them, because synchronization loss is estimated and recovered. We need to further investigate the genes identified by the proposed method only, and to validate the identified genes through biology experiments or previous biology knowledge. One possible validation method is to validate the biological relevance of such identified cell-cycle genes by semantic analysis based on the Gene Ontology (GO) terms. To achieve this purpose, an online tool is applied, the SGD GO Term Finder (http://db.yeastgenome.org/cgi-bin/GO/goTermFinder). We analyzed the set of non-overlapping genes which are identified by one method, but not by the other two methods. The top GO terms associated with each method's results can be found in the Supplementary Material. For the proposed method, in the top 25 GO terms, there are several cell-cycle related terms, such as 'M phase', 'cell-cycle', 'mitotic cell cycle' and 'M phase of mitotic cell cycle', including 84 genes. It suggests that some genes identified by the proposed method but not by the other two methods are cell-cycle related. For the sets of non-overlapping genes identified by the two reference methods, it is noted that none of the above four cell-cycle related GO terms appears in the top 25 GO terms. Details of top GO terms associated with results of each method can be found in the Supplementary Materials. Theseencouraging observations demonstrate that the proposed method is promising for identifying cyclic genes.

6 CONCLUSION

Synchronization loss is a major concern in identifying cyclic genes to understand the fundamental cyclic systems. We developed a model-based framework for identifying cell-cycle regulated

Table 5. The performance sensitivity to inexact prior knowledge of cell-cycle length

Probability of detection	$T = 35$	$T = 40$	$T = 45$	$T = 50$	$T = 55$	$T = 60$	$T = 65$	$T = 70$	$T = 75$	$T = 80$
0.75	0.0964	0	0.0132	0	0	0	0	0.0506	0.1176	0.1573
0.80	0.1011	0	0.0123	0	0	0	0	0.0476	0.1667	0.1489
0.85	0.1237	0	0.0116	0	0	0	0	0.0761	0.1827	0.1827
0.90	0.1818	0	0.0110	0	0	0	0	0.1089	0.2373	0.2373
0.95	0.3791	0.0104	0.0104	0.0206	0.0104	0	0.0104	0.1364	0.2857	0.2460
1.00	0.6032	0.0909	0.1228	0.0476	0.0099	0.0099	0.0099	0.2188	0.3939	0.5215

When the probability of correctly detecting cyclic genes is fixed, we compare the probability of false positive, under different prior knowledge of cell-cycle T.

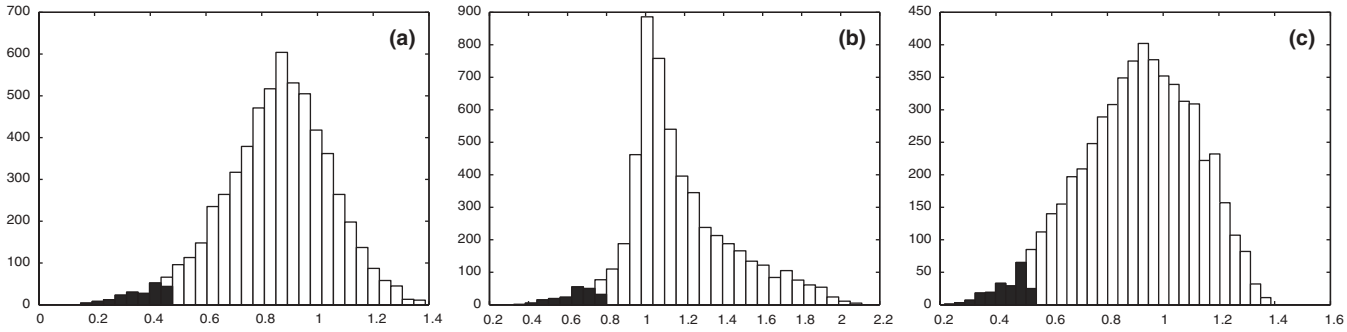

Fig. 6. Histogram of fitting residues for the cdc28 dataset. (**a**) alpha, (**b**) cdc15 and (**c**) cdc28. Solid curve represents the histogram of fitting residues for training gene set.

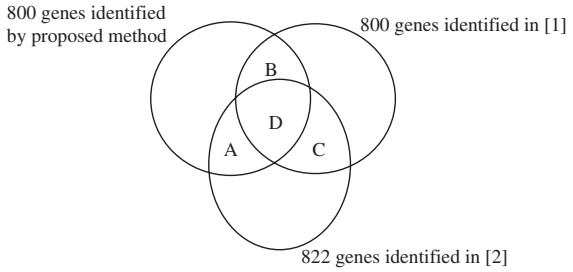

Fig. 7. Venn diagram of genes identified by proposed method and Spellman *et al*. (1998), Lu *et al*. (2004). The intersection between proposed method and Spellman *et al*. (1998) is 403 (B + D); the intersection between proposed method and Lu *et al*. (2004) is 433 (A + D); the intersection between Spellman *et al*., 1998 and Lu *et al*., 2004 is 541 (C + D); the intersection among all three studies is 355 (D).

genes through resynchronization and reconstructing the underlying gene expression profiles, which representing a single-cell behavior more accurately. We consider a simple synchronization loss model where the gene expression measurements are regarded as superposition of mixed cell populations with different growth rates. The proposed scheme is shown feasible, promising and robust via simulations. Results from real mircoarray data analysis reveal that the reconstructed profiles represent a more accurate expression profiles and improve our ability to identify cyclic genes. We will further investigate the proposed method by combining complementary information such as budding index.

Conflict of Interest: none declared.

REFERENCES

Bar-Joseph,Z. *et al*. (2004) Deconvolving cell cycle expression data with complementary information. *Bioinformatics*, **20** (Suppl. 1), i23–i30.

Cho,R.J. *et al*. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.

Johansson,D. *et al*. (2003) A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription. *Bioinformatics*, **19**, 467–473.

Lee,T.I. *et al*. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.

Lu,X. *et al*. (2004) Statistical resynchronization and Bayesian detection of peroidically expressed genes. *Nucleic Acids Res*., **32**, 447–455.

Moore,S. (2001) Making chips to probe genes: biotechnology''. *IEEE Spectrum*., **38**, 54–60.

Shedden,K. and Cooper,S. (2002a) Analysis of cell-cycle gene expression in *Saccharmoyces cerevisiae* using microarray and multiple synchronization methods. *Nucleic Acids Res*., **30**, 2920–2929.

Shedden,K. and Cooper,S. (2002b) Analysis of cell-cycle-specific gene expression in human cells as determined by microarrays and double-thymidine block synchronization. *Proc. Natl Acad. Sci. USA*, **99**, 4379–4384.

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

Stoer,J. and Bulirsch,R. (1991) *Introduction to Numerical Analysis*. Springer.

Poor,H.V. (1994) *An Introduction to Signal Detection and Estimation*. Springer.

Whitfield,M.L. *et al*. (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, **13**, 1977–2000.

Wichert,S. *et al*. (2004) Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, **20**, 5–20.