Proceedings of the 2005 IEEE
Engineering in Medicine and Biology 27th Annual Conference
Shanghai, China, September 1-4, 2005

Tracking the Herd: Resynchronization Analysis of Cell-Cycle Gene Expression Data in *Saccharomyces Cerevisiae*

Peng Qiu, Z. Jane Wang[2], and K. J. Ray Liu
Department of Electrical and Computer Engineering, University of Maryland, College Park, USA
[2] Department of Electrical and Computer Engineering, University of British Columbia, Canada

Abstract— **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 microarray gene expression data is complicated by the factor of synchronization loss, thus remains a challenging problem. Decomposing the expression measurements will allow to better represent the single-cell behavior and improve the accuracy in identifying periodically expressed genes. In this paper, we propose a resynchronization-based algorithm for identifying cell-cycle-related genes, where we present a simple 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 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 cycling genes and revealing underlying gene expression profiles.**

I. 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 [1]. Recent advances in highthroughput gene expression data acquisition technologies, such as microarrays, provide a rich opportunity for achieving this goal. The first critical task in understanding such cyclic systems is to identify the genes which are periodically expressed during the cell-cycle. In the current technologies, many expression data are measured based on a population of cells which are synchronized to exhibit similar behavior. 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 due to the diversity of individual cell growth rates. Therefore, except the noise effect in the measurements, a significant difficulty in identifying cell-cycle regulated genes by analyzing microarray gene expression data arises from synchronization loss. Because direct periodicity testing on the expression measurements could fail or be misleading due to the fact that the expression values are from a mixed cell populations growing at different rates.

Several approaches for identifying cell-cycle regulated genes by 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. The former category relying only on expression data was mainly studied in the literature. A Fourier analysis algorithm was employed for synchronization test of different arrest methods in [5]. The authors presented an exact statistical test to identify periodically expressed genes by distinguishing periodic from random processes [6]. In [2], 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. Along the second line, an algorithm combining budding index and gene expression data to deconvolve expression profiles was proposed recently in [4]. Regardless these developments, efforts are still needed to accurately identify cyclic genes and recover a more accurate gene profile compared with the current expression measurements.

Our goal in 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 is two fold.

- We propose a synchronization loss model by representing the gene expression measurements as a superposition of different cell populations growing at different rates, and we develop a model-based estimation algorithm to reconstruct the underlying single-cell gene expression profiles.
- Using the fitting error as criteria, we explore a supervised learning scheme for identifying the cell-cycle regulated genes. The performance of the proposed scheme are examined by both simulations and real microarray gene expression data of *Saccharomyces Cerevisiae*.

II. SYSTEM MODEL AND FORMULATION

A. Mixture Model for Synchronization Loss

Even with the best synchronization method currently available, cells begin to lose their synchronization shortly. We propose to model the observed gene expression data as a superposition from a mixed population of cells growing at slightly different rate, as

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

0-7803-8740-6/05/$20.00 ©2005 IEEE.

where $y_i(t)$ is the observed expression of gene i; $x_i(t)$ is the underlying single-cell expression profile; ρ_m represent the relative growth rates of cells with respect to cell-cycle; β_m represent the percentage of cells with growth rate ρ_m, and it is assumed to be constant in one series of measurements. Because of different growth rates in experiment cell populations, even if a gene is cell-cycle regulated, the observed expression may not exhibit clear periodicity. Therefore, it is difficult to detect periodically expressed genes, and distinguish them from non-periodically expressed genes.

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 describe the synchronization loss status of the cell populations. Here we propose to utilize some common properties of all genes to extract underlying expression profiles from the observations.

B. Formulation

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

$$x_i(t) = \sum_{m=1}^{M} a_m y_i(t * c_m), \tag{2}$$

where parameters a_m and c_m are not the same with β_m and ρ_m. They are related in a complicated nature. Mathematically, equation (1) and (2) are not equivalent. However, the intuitive idea is that: if c_m can cover a larger range than ρ_m, model (2) could be accurate enough. This is motivated by the idea of approximating an Infinite Impulse Response (IIR) filter using a Finite Impulse Response (FIR) filter, since the relationship between (1) and (2) is similar to that of Finite FIR filters and IIR 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 large enough, 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, equation (1) and equation (2) relates $x_i(t)$ and $y_i(t)$ in approximately the same way.

Again, note that, the parameters a_m and c_m depend solely on β_m and ρ_m. They are the common constants for all genes. We proposed to utilize this common property of all genes to extract underlying expression profiles.

For cell-cycle regulated genes, because of periodicity, $x_i(t) = x_i(t+T)$, through equation (2), the observations and a_m, c_m parameters are related as follows:

$$\sum_{m=1}^{M} a_m [y_i(t * c_m) - y_i((t+T) * c_m)] = 0. \tag{3}$$

Denote $\underline{y_i}(t) = [y_i(t * c_1) - y_i((t+T) * c_1), ..., y_i(t * c_M) - y_i((t+T) * c_M)]^T$, and $\underline{a} = [a_1, ..., a_M]^T$. Equation (3) can be written as,

$$\underline{y_i}(t)^T \underline{a} = 0, \tag{4}$$

Note that, we can evaluate equation (4) at different time points (as long as data allows). Also, the cell-cycle regulated genes all satisfy equation (4). So, the estimation of a_m parameters can be formulated as a constrained least square problem,

$$[\underline{y_i}(t_1), ..., \underline{y_i}(t_n), \underline{y_j}(t_1), ..., \underline{y_j}(t_n), ...]^T \underline{a} = 0, \tag{5}$$

$$\text{subject to } \sum_{m=1}^{M} a_m = 1 \tag{6}$$

where genes $i, j, ...$ are cell-cycle regulated genes. In the current *Saccharomyces Cerevisiae* time-series gene expression data, usually only two cell-cycles are covered, and the sampling interval is large. The value of n in equation (5) is quite small, e.g. 18. Therefore it is important to use many cell-cycle regulated genes together to estimate coefficients a_m's reliably. In this formulation, c_m's 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 reasonably small. In later simulations, we will show that, it is accurate enough to choose c_m to cover the range from 0.5 to 1.5. The fixed-summation constraint in equation (6) is chosen to avoid the trivial 0-vector solution, i.e. $\underline{a} = \underline{0}$.

C. Fitting Residue Criterion

After estimating a_m's, the model in (2) is used to extract 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 of noises 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 small, then taking into the effects of synchronization loss into consideration, the extracted signal could lead to the measurements with high probability, 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 periodic signal is not closely related to experiment observation, which means the gene is more likely not cell-cycle regulated.

III. PROPOSED IDENTIFICATION SCHEME

Based on the synchronization loss model and estimation approach described in Section II, 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_m's, and extract the underlying periodical signal for every gene in the training set.

3) Based on the extracted signal x_i, 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_m' using the training set and use the estimated a_m's 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. And go to Step 2.

Note that,under this framework, in order to purify training set and detect periodically expressed genes correctly, the criteria for removing and including genes in step 2 and step 4 should be carefully designed and fine tuned for each data set. It is admitted that we apply a heuristical way to update the training set in the current stage of our study.

IV. SIMULATION

In this section, we simulate time-series expression data for 100 periodically expressed genes and 600 non-periodically expressed genes. For simplicity, the underlying periodic expression profile for cell-cycle regulated gene i is generated by a linear combination of sinusoids with random phases,

$$x_i(t) = \sum_{j=1}^{4} \lambda_{ij} sin(\frac{2\pi j}{T} t + \phi_{ij}), \qquad (7)$$

where the period T is set to be 60 minutes, same as the cell-cycle duration in the alpha experiment in [1]. The parameter λ_{ij} is randomly chosen, different for each gene. ϕ_{ij} represents the random phase, which is uniformly distributed within $[0, 2\pi)$. For the 600 non-periodical genes, their underlying expressions are obtained through random permutations of expressions of periodical genes.

For each gene, we simulate the synchronization loss by

$$y_i(t) = \beta_1 x_i(t * s) + \beta_2 x_i(t) + \beta_3 x_i(t * f), \qquad (8)$$

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. Equation (8) is applied on all genes, which represents the common synchronization status of the cell populations. Similar to the alpha experiment in [1], samples are taken every 7 minutes from 0 to 119 minutes, giving 18 time points in total.

In the simulation, 50 periodically expressed genes are assumed known, in order to form the training set. The testing set contains the other 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.

As mentioned earlier, c_m should cover a large enough range, in order to extract underlying expression profiles accurately. In table I, different range of c_m is examined. To ensure a fair comparison, M is set to be 7, and values of c_m are chosen to be uniformly spaced in tested range. In the fitting residue criterion, ρ_m is set to be $[0.7, 0.8, ..., 1.3]$. From table I, we can see that, different choice of c_m leads to different fitting residues for both periodical genes and non-periodical genes. Regarding this particular simulation setting, we can see that, range of c_m being $0.5 \sim 1.5$ is an appropriate choice. Because the average fitting residue for periodical genes is small, and the difference between periodical genes and non-periodical genes is large.

range of c_m	avg fitting residue periodical genes	avg fitting residue non-periodical genes	diff
[0.9, 1.1]	0.4352	0.9619	0.5268
[0.8, 1.2]	0.4141	0.9613	0.5472
[0.7, 1.3]	0.2980	0.9636	0.6656
[0.6, 1.4]	0.2687	0.9278	0.6591
[0.5, 1.5]	0.2665	0.9425	0.6760
[0.4, 1.4]	0.2963	0.9351	0.6389

TABLE I

COMPARISON OF THE NORMALIZED AVERAGE FITTING RESIDUES FOR PERIODICAL AND NON-PERIODICAL GENES.

After determining c_m's, the proposed model is applied to estimate a_m's based on the training set, and extract the underlying periodical signals for genes in the training set. Fig 1 is a typical example of genes in the training set, where we can see that the simulated observations is quite different from the underlying periodical expression profile. Due to synchronization loss, the observed time-series does not exhibit a clear periodicity, especially in the second cycle. It is encouraging to see that, from poorly synchronized observations, the proposed scheme can successfully recover the underlying periodic expression profile.

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 residue criterion is examined. In Figure 2, the histogram of residues shows that, the periodically expressed genes and non-periodically expressed genes are well separated, meaning that the proposed scheme can successfully identify the cyclic genes and reconstruct the underlying periodic profiles.

V. REAL DATASETS

In this study, we investigated three real datasets, alpha, cdc15 in [1] and cdc28 in [3]. From [1], 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 periodically expressed genes simultaneously. During each iteration, we adopt simple removing and including criterion in step 2 and step 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 training set. As an example, the histogram of fitting residues for the cdc28 dataset is shown in Fig 3.

To make a fair comparison with [1] and [2], 800 genes with smallest fitting residues are identified as cyclic genes. In Fig 4, a venn diagram showing the overlap of genes identified by different studies. It is encouraging to see the large overlaps

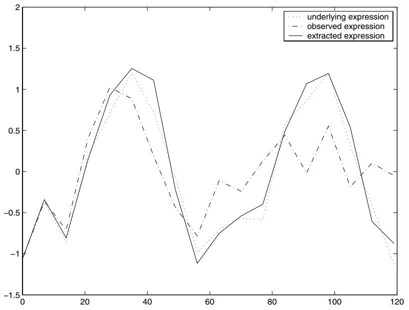

Fig. 1. The underlying periodical expression, experiment observation and extracted expression of one simulated gene.

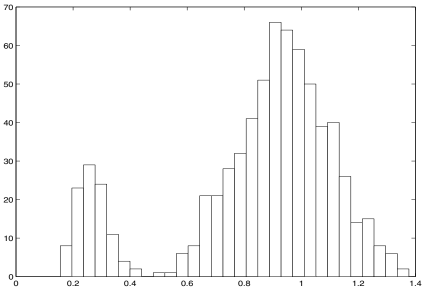

Fig. 2. The histogram of fitting residues for all genes. The horizontal axis represents fitting residue, and the vertical axis represents number of genes with certain value of fitting residue.

illustrated in Fig 4, an indication of consistency of the proposed scheme to the previous researches.

Though genes identified by the proposed method have large overlap with previous studies, it is interesting to examine genes identified by the proposed method, but not identified in either [1] or [2]. Fig 5 is a typical example. Since [1] [2] are based on Fourier analysis, genes without clear periodicity may not be identified. The proposed method may be able to identify them, because synchronization loss is estimated and recovered. We need further investigate whether genes identified by the proposed method only are cell-cycle regulated. Due to the space limitation, more details regarding the results in yeast cell-cycle study can be found in [7].

VI. 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 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 is regarded as superposition of mixed cell populations with different growth rates. The proposed scheme is shown feasible and promising 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 cycling genes. We will further investigate the proposed scheme by combining complementary information such as budding index.

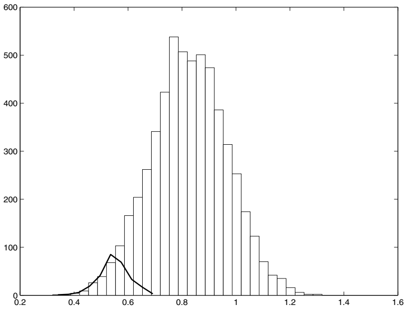

Fig. 3. Histogram of fitting residues for the cdc28 dataset. Solid curve represents the histogram of fitting residues for training gene set.

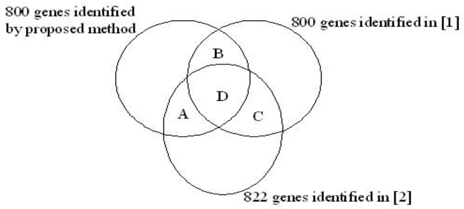

Fig. 4. Venn diagram of genes identified by proposed method and [1], [2]. The intersection between proposed method and [1] is 428 (B+D); the intersection between proposed method and [2] is 426 (A+D); the intersection between [1] and [2] is 540 (C+D); the intersection among all three studies is 368 (D).

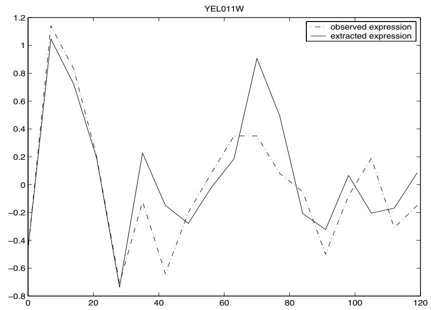

Fig. 5. A typical example identified by the proposed method, while missed by [1] [2]. The observed expression does not exhibit clear periodicity. However, the proposed method can recover the underlying periodical expression.

REFERENCES

[1] P.T.,Spellman, G.,Sherlock, M.,Zhang, V.,Iyer, K.,Anders, M.,Eisen, P.,Brown, ,D.,Botstein and B.,Futcher, “Comprehensive identi- fication of cell cycle-regulated genes of the yeast Saccharomyces cerevisia by microarray hybridization”, *Mol. Biol. Cell*, 9, 32733297, 1998

[2] X.,Lu, W.,Zhang, Z.,Qin, K.,Kwast, and J.,Liu, “Statistical resynchronization and Bayesian detection of peroidically expressed genes”, *Nucleic Acids Research*, 32, 447455, 2004

[3] R.J.,Cho, M.J.,Campbell, E.A.,Winzeler, L.,Steinmetz, A.,Conway, L.,Wodicka, T.G.,Wolfsberg, A.E.,Gabrielian, D.,Landsman, D.J.,Lockhart, and R.W.,Davis, “A genome-wide transcriptional analysis of the mitotic cell cycle”, *Mol. Biol. Cell*, 2, 6573, 1998

[4] Ziv Bar-Joseph et. al., “Deconvolving cell cycle expression data with complementary information”, *Bioinformatics*, vol. 20 (Suppl. 1), pp.ı23-i30, 2004.

[5] K. Shedden and S. Cooper, “Analysis of cell-cycle gene expression in Saccharmoyces cerevisiae using microarray and multiple synchronization methods”, *Nucleic Acids Research*, vol. 30, no. 13, pp. 2920-2929, 2002.

[6] S. Wichert, K. Fokianos, and K. Strimmer, “Identifying periodically expressed transcripts in microarray time series data”, *Bioinformatics*, vol. 20, no. 1, pp. 5-20, 2004.

[7] P. Qiu, Z. Jane Wang, and K. J. Ray Liu,“Indentifying Cyclic Genes Via Resynchronization Using Microarray Expression Data”, in submitting to *Bioinformatics*, 2005.