Detecting Periodically Expressed Genes based on Time-frequency Analysis and L-curve Method

Detecting Periodically Expressed Genes based on Time-frequency Analysis and L-curve Method

Xiangchao Gan[1], Alan Wee-Chung Liew[2] and Hong Yan[1,3]
[1]*Department of Electronic Engineering*
City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong
[2]*Department of Computer Science and Engineering*
The Chinese University of Hong Kong, Shatin, Hong Kong
[3]*School of Electrical and Information Engineering*
University of Sydney, NSW 2006, Australia

Abstract

In microarray experiments, gene expression profiles are often affected by biological properties, such as synchronization loss, and show some non-stationarity. Worse still, the microarray data usually suffers from missing values. The conventional spectrum-based methods, when used to identify a subset of genes that are periodically expressed, are degraded by these factors. In this paper, we use the Wigner-Ville distribution analysis and L-curve method for detection of periodically expressed genes. We provide a graphical exploratory device for assessment of the presence of periodically expressed genes. Then, we identify the subset of genes actually involved in the cell cycle using the L-curve method. The experiments on several widely used datasets show that our algorithm can effectively reduce the effect of non-stationarity and missing values problems.

1. Introduction

We often need to analyze cell cycle gene expression data in microarray experiments to find the evidence of the periodicity and identify a subset of genes involved during the cell cycle. The gene expression data are quite different from the classical time series data. They often have small number of measurements per gene and large number of genes needed to be identified. Furthermore, the data are often degraded by high levels of non-normal random noise and plenty of missing values. Difficulties in identifying the periodically expressed genes have also brought controversies about the statistical significance of some published results. Shedden and Cooper [1] questioned the presence of cyclic signals, which are analyzed by Spellman *et al.*[2]. Thus, identifying periodically expressed genes is a significant problem and there is a clear need to develop reliable methods for solving the problem.

In this paper, we propose a spectrum-based method to detect periodically expressed genes, which consists of two complementary procedures. First, we provide a graphical exploratory device for assessment of the presence of periodically expressed genes. Second, we identify the subset of genes actually involved in the cell cycle using a so-called *L*-curve method, which is developed based on the time-frequency energy spectrogram.

Spectrum-based methods have many applications to gene expression data analysis. Wichert *et al.* [3] use average periodogram to identify periodically expressed transcripts. In this method, it is assumed that the gene expression data is stationary. However, a typical gene expression data set has a small number of measurements per gene and the time series data for each gene is usually non-stationary. Furthermore, the gene expression data is often affected by many biological phenomena and they pose many different tendencies for gene time series data values. For example, Bar-Joseph *et al.* [4] find that synchronization loss will cause the peak expression values to be lower in the second cycle and the lowest expression values to be higher for most cell-cycle regulated genes. This kind of non-stationarity of the gene data cannot be properly processed by conventional spectral methods, but can be dealt with using time-frequency based techniques. In practice, there are often some missing values in gene expression data. This is caused by diverse reasons, including spotting problems, slide scratches, hybridization error and image corruption. In most cases, missing values are often replaced by imputing values or simply by

The 18th International Conference on Pattern Recognition (ICPR'06)
0-7695-2521-0/06 $20.00 © 2006 IEEE

average values of a gene. Obviously, these values are not reliable. For conventional spectrum-based method, these unreliable values will affect the whole signal spectrum. In time-frequency energy spectrogram, the contribution of a single value is limited to a certain range, thus the degradation caused by an unreliable value is effectively reduced.

Another method used by our algorithm is the *L*-curve, which provides a convenient graphical tool for analysis of ill-posed problems [7]. It has many applications in regularization-based image processing methods. It is the first time that this method is used for detecting periodically expressed genes and our experiment results show that it is very effective.

2. Methods

Outline

To analyze the cell cycle data, we are interested in two questions: whether there exist cell-cycle–specific signals and which genes are relevant for the cell cycle. In our method, we provide two complementary procedures. First, we provide a graphical exploratory device for assessment of the presence of periodically expressed genes. Second, we identify a subset of genes actually involved in the cell cycle using the *L*-curve method based on the time-frequency energy spectrogram. Unlike the common threshold method, the *L*-curve method is easy to interpret and takes account of multiple tests.

Average time-frequency energy spectrogram

A time-frequency distribution which is particularly useful for gene expression data analysis is the *Wigner-Ville* distribution (WVD). A *WVD* of signal x is defined as,

$$W(t,v) = \int_{+\infty}^{-\infty} x(t+\tau/2)x^*(t+\tau/2)e^{-j2\pi v\tau}d\tau \qquad (1)$$

The above definition has a windowed version as

$$W(t,v) = \int_{+\infty}^{-\infty} h(\tau)x(t+\tau/2)x^*(t+\tau/2)e^{-j2\pi v\tau}d\tau \qquad (2)$$

where $h(\tau)$ is a regular window function. The function in Equation (2) is also called as pseudo-WVD.

In a typical microarray experiment, genes may or may not be cell-cycle regulated. To identify a subset of genes which are cell-cycle regulated among many random time series, we use a graphical device called average time-frequency energy spectrogram (TFES). Let a time-frequency energy spectrogram of a gene time series data be $T(t, v)$. We binarize $T(t, v)$ to form an image $I(t, v)$ as follows: if a value is larger or equal to the 95% percentile of T(t, v) (i.e., it is among the 5% largest values), replace it with one, and otherwise replace it with zero. A typical binary time-frequency energy spectrogram of a cell-cycle-regulated signal is shown in Figure 1(a). If there are N genes in a microarray experiment, then the average time-frequency energy spectrogram is represented as

$$P(t,v) = \frac{1}{N}\sum_{i=1}^{N} I_i(t,v) \qquad (3)$$

For a group of random time series data, the average time-frequency energy spectrogram will be a uniform background in a 2-D image. For cell-cycle-regulated gene expression data, there exists a line in the corresponding TF energy spectrogram. When there are genes which are cell cyclic, the average TF energy spectrogram will appear as an image with a line embedded in a noise background (Figures 1 (b) and (c)).

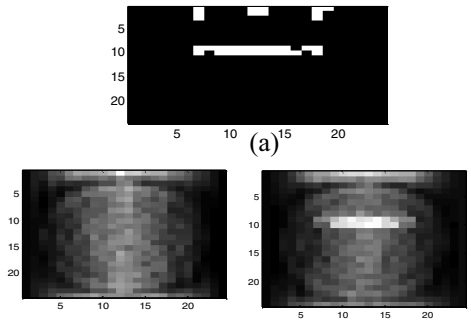


Fig. 1. Average time-frequency energy spectrogram (TFES) for simulated time series data of length 24. (a) The binary time-frequency energy spectrogram of a time series with 2 cycles and a random phase. (b) The average TFES for 1000 random time series. (c) The average TFES for 900 random time series and 100 time series with 2 cycles and random phases. The embedded line is very obvious and easy to detect.

The *L*-curve method for periodicity detection

In Figure 1, we see that when a time series is relevant for the cell cycle, there is a line in its average TFES. The more obvious is the line, the more significant the periodic property of the data is. To judge whether a gene is cell-cycle regulated, we need to provide an objective evaluation. We make use of the weighted line length (WLL) (Here term "weighted" is used since width of the line changes with the frequency resolution of our TFES. If the frequency sampling number is twice of the length of the signal, the width of line in our TFES is 2 and we use a weighted length). Using the WLL as the measurement, we then provide the L-curve method to test whether a gene is periodically expressed.

The *L*-curve is a convenient graphical tool for analysis of the ill-posed problem [7]. It has many applications in regularization-based image processing method. However, it is the first time that this method is used for detecting periodically expressed genes. The *L*

curve clearly displays the compromise between two quantities, the number of genes and the WLL in our present gene expression data analysis problem problem. The generic form of the *L*-curve is shown in Figure 2(a) (Hansen, 1994). When plotted the two quantities always have a characteristic L-shaped appearance (hence its name) with a distinct corner separating the vertical and the horizontal parts of the curve.

In microarray time series data, usually there is only a small fraction of the genes under investigation exhibits some evidence of periodically varying expression during cell cycle so that the overall signal in the data is dominated by non-periodic components (Wichert *et al.*, 2004.). When we calculate the WLL for each gene, sort the values in descend and then plot them in a diagram, we get an L-curve. The distinct corner means that WLL values of many genes are less than or close to the corner value. Since we assume that most of the genes are not relevant to the cell cycle, the value at the corner is the best threshold for separating periodically expressed genes and non-periodic genes.

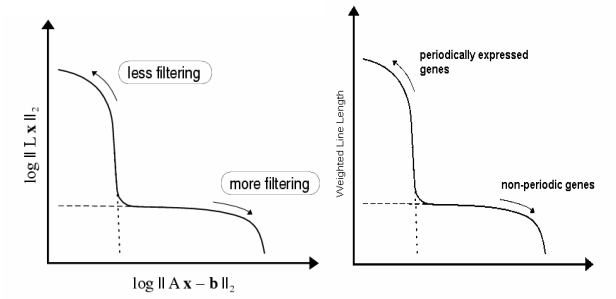


Fig. 2. Illustration of the *L*-curve method. (a) A typical example of the L-curve method. (b) Using *L*-curve method to detect periodically expressed genes.

Procedure for the proposed algorithm

To summarize, when detecting periodically expressed genes in a microarray experiment, we propose following steps:

(1) Using the average time-frequency energy spectrogram to explore whether or not there are periodically expressed genes.
(2) Calculate the weighted line length (WLL) for each gene.
(3) Based on sorted WLL, using the *L*-curve to identify which genes is periodically expressed.

In Step (3), it is easy to determine the *L*-curve corner visually. If we want to identify it automatically, a computationally convenient definition of the L-curve's corner is the point with maximum curvature. Some tools can help us locate the corner in the *L*-curve (Hansen, 1994).

3. Experiments

We now test our algorithm on yeast *Saccharomyces cereisiae* microarray dataset. This experiment is performed by Spellman *et al.* [2] (http://cellcycle-www.stanford.edu) and has been widely used as benchmark data set in previous studies. It contains expression profiles for 6178 genes under different experimental conditions, i.e., cdc15, and cdc28, alpha factor and elutriation experiments. For all datasets, the missing values are imputed using a projection onto convex sets (POCS) method [5].

The average time-frequency energy spectrograms computed for the yeast data are shown in Figure 3. These plots display a clear signal of periodicity.

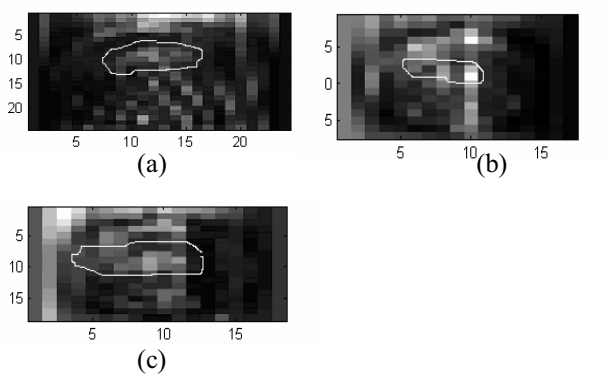
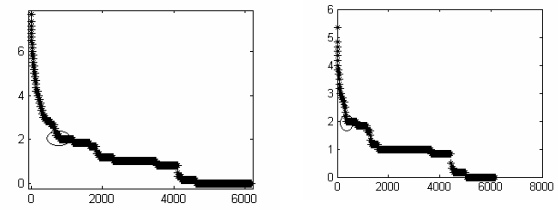
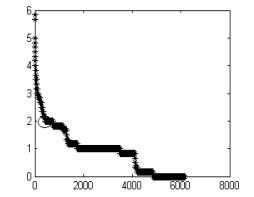


Fig. 3. Average time-frequency energy spectrogram for the yeast data sets (a) cdc15, (b) cdc28 and (c) alpha.

To identify which gene is cell cycle regulated, we calculate the weighted line length of each gene and then use *L*-curve to do the multiple test. The genes with WLL higher than the corner value of the *L*-curve are identified as cell-cycle-regulated. The *L*-curve for each experiment is provided in Figure 4. The result of our algorithm for each experiment is given in Table 1. For cdc15, 22 genes which are confirmed to by cell-cycle regulated by corresponding annotations are detected. For cdc28, the number of genes we detected is almost 2-fold of the result of conventional spectrum-method. de Lichtenberg et al. (2005) have found that normalization to this dataset can help to detect more cell-cycle regulated genes. This demonstrates that our algorithm has more resistibility to data degradation.



(a) (b)



(c)

Fig. 4. The *L*-curve of weighted line length for the yeast data sets (a) cdc15, (b) cdc28 and (c) alpha. The corner values are represented by ellipses.

Cell type	Experiments	N	G	C	C/G
Yeast	cdc15	24	5673	805	14.2
Yeast	cdc28	17	6125	362	5.9
Yeast	alpha	18	6075	412	6.8

Table 1. The results of periodic gene detection using the proposed algorithm for the yeast dataset. Here *N* is the number of time points, *G* the total number of genes, and *C* the number of periodic genes

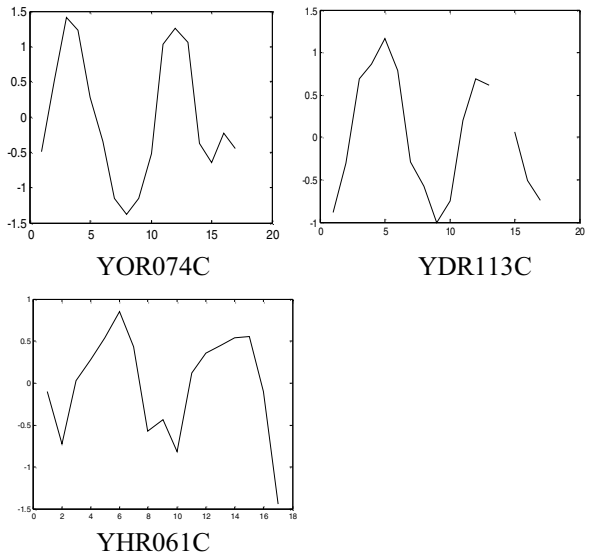


Fig. 5. The three most significant periodic genes in cdc28 experiment, which are found using the proposed method. There is a missing value in the data of YDR113C.

In Figures 5, we plot the profiles of the three most significant periodic genes in cdc28 dataset as time functions. The property of the cell cycle is very obvious. Although some gene expression data profiles contain missing values, our algorithm still can correctly identify them. Compared to the result of Wichert et al.'s algorithm, it is obvious our algorithm has more resistibility for the missing values problem.

4. Conclusions

In this paper, we have proposed a time-frequency analysis based method for detecting cell-cycle regulated genes. In our method, the Wigner-Ville distribution is used to compute the time-frequency energy spectrogram of gene expression data. The average of the thresholded spectrograms from all gene expression data contain a distinctive line if there are a subset of genes that are periodically expressed. An *L*-curve based method can then be used to identify the periodic genes. Compared with existing methods based on spectral analysis, our method can reduce the degradation caused by synchronization loss of gene time series data and missing values. Experiments on several widely used reference datasets show that our algorithm is very effective.

10. References

[1] Sheden,K. and Cooper,S. (2002) Anaysis of cell-cycle gene expression in *Saccharomyces cerevisiae* using microarrays and multiple synchronization methods. *Nucleic Acids Res.*, **30**, 2920-2929

[2] Spellman,P.T. (1998) Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization. *Mol. Biol. Cell.*, **9**, 3273-3297.

[3] Wichert,S., Fokianos, K. and Strimmer,K. (2004) Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, **20**, 5-20

[4] Bar-Joseph,Z., Farkash,S., Gifford,D.K., Simon,I. and Rosenfeld,R. (2004) Deconvolving cell cycle expression data with complementary information. *Bioinformatics*, **20**, i23-i30

[5] Gan.X, Liew,A.W.C. and Yan,H. (2004) Missing Microarray Data Estimation Based on Projection onto Convex Sets Method. *Proceeding of ICPR 2004*, **3**, 782-785

[6] de Lichtenberg, U. et al., (2005) Comparison of computational methods for the identification of cell cycle-regulated genes, *Bioinformatics*, 21, 1164-1171

[7] Hansen, P.C. (1994) Regularization Tools: A Matlab package for analysis and solution of discrete ill-posed problems, *Numerical Algorithms*, **6,** 1-35.

[8] Lu,X., Zhang,W., Qin,Z.S., Kwast,K.E. and Liu,J.S. (2004) Statistical resynchronization and Bayesian detection of periodically expressed genes. *Nucleic Acids Research*, **32**, 447-455