Mathematical Biology and Bioinformatics
2017. V. 12. № 1. P. 198–203. doi: 10.17537/2017.12.198

========== **BIOINFORMATICS** ==========

UDC: 57.087.1:577.218

False discovery rate of classification as a function of periodicity strength of time-course gene expression

©2017 Farzad Najafi Amiri*[1], Mahnaz Khalafi†[1], Masoud Golalipour‡[2], Majid Azimmohseni§[1]

[1]*Department of Statistics, Faculty of Science, Golestan University, Gorgan, Iran*
[2]*Medical cellular and molecular Research center, Golestan University of Medical Sciences, Gorgan, Iran*

Abstract. Classification of genes provides valuable information about similar types of gene expressions. The periodic structure of time-course gene expression is a reliable characterization to classify two genes with the same periodic pattern in the same class. The strength of periodicity may differ from one gene to another. In this article, using Lomb-Scargle and JTK methods, three types of cyclic time-course patterns of genes are introduced according to periodicity strength. We proposed that the periodicity is an important factor for gene discrimination according to time-course expression profile. Based on the *Saccharomyces cerevisiae* data set, genes with different periodicity were discriminated, according to both the amounts of phase shift and time-course expression. Then, false discovery rates were computed under all circumstances. As a result, the false discovery rate increased when the strength of periodicity decreased. The false discovery rate of genes with strong periodic structure was 20 % whereas it was 45 % for weak periodic ones. The data set comprised 79 % of genes with a weak periodicity, that deviated the result of discrimination.

Key words: *discrimination, JTK, Lomb-Scargle method, phase shift, Saccharomyces cerevisiae data.*

1. INTRODUCTION

Clustering and discrimination of genes according to their time-course expression data is important while investigating genes with the similar expression patterns. Several methods have been proposed to explore time-course expression pattern [1, 2].

A common approach to cluster and discriminate genes is frequency-based similarity measures. In this method at first, Fourier transform for each gene is calculated and then the distance measure between Fourier transforms of genes is computed. Finally, genes are grouped according to the values of distance measures. Spellman et al. (1998) [3] used frequency-based approach to cluster and discriminate genes based on the time-course expression. Eisen et al. (1998) [4] used this method for clustering microarray expression data. Glynn et al. (2006) [5] used Lomb-Scargle periodogram as an efficient method to estimate frequency density and frequency-based clustering and discrimination. Lomb-Scargle method is useful for unevenly spaced gene expression data or gene expression data with missing values. Zhao et al.

*f.najafi@stu.gu.ac.ir
†Corresponding Author E-mail: m.khalafi@gu.ac.ir
‡gapmasood@goums.ac.ir
§m.azim@gu.ac.ir

(2009) [6] conclude that the frequency-based methods provide trustworthy approaches for gene classification. The most crucial assumption in the frequency-based method is the existence of a cyclic pattern in time-course gene expression profile. There are several reasons for the periodic pattern of gene regulation. This regulation shows the proper functioning of mechanisms that maintain order during cell division and gene expression control. Cell environment and growth factors play a crucial role in regulation and fluctuation of gene expression [1]. The choice of algorithm and method to identify periodicity is a challenging problem, Wichert et al. (2004) [8]. A common method to assess the hypothesis of periodicity is periodogram-based Fisher′s test. Zhao et al. (2001) [9] used a single pulse model to find the periodic pattern in the microarray profiles. Chen et al. (2005) [5] propose two statistical hypothesis-testing procedures to detect significant periodically expressed genes. Ahdesmaki et al. (2005) [10] propose the use of a robust Fisher′s test that has better performance when the data deviate from normal assumption. Glynn et al. (2006) [5] propose Lomb-Scargle test statistic for detecting a cyclic pattern in gene expression data. Hughs et al. (2010) [11] used the Jonckheere-Terpstra-Kendall (JTK) method as an efficient nonparametric algorithm for detecting the periodic pattern in genome-scale data sets. Several issues arise while applying different algorithms to detect the cyclic pattern. The first problem is the lack of apparent periodicity in genes expression data; yet if they had nominated to be fluctuating genes. In other words, the strength of periodicity in genes expression is not high enough to be detected by the common tools. The second issue is that the algorithms for detecting whether there is a cyclic pattern in the time-course expression of a specific gene differ from one another. Since the knowledge concerning the periodic nature of gene expression is necessary for gene ontology studies, the mentioned problems may lead to unreliable results.

In the present study, the effect of the cyclic pattern of gene expression on clustering and discrimination of genes was investigated. Moreover, the false discovery rate was calculated as a function of the periodicity strength to show the decrease in precision of classification for weak periodicity strength.

The rest of this paper is organized as follows. According to Lomb-Scargle and JTK methods, in section 2 three types of cyclic patterns in gene expression data are introduced. In section 3, to illustrate the method, the yeast expression data [3] are used to discriminate cell-cycle dependent genes in the *S. cerevisiae*. A set of 800 genes is identified with the cyclic regulation pattern in Spellman study. Besides, they identified the phase of these genes using Fourier transform of the data. Based on this information, discrimination of genes is performed according to their type of cyclic pattern and the numerical results are provided. The main conclusions of this study are listed in section 4.

2. THE PERIODICITY TYPE IN GENE EXPRESSION DATA

According to the Lomb-Scargle and JTK methods, we introduce three types of periodicity in the time-course genes expression data.

The Lomb-Scargle periodogram method

The algorithm of Lomb-Scargle method assumes that time-course genes expression observations are similar to a deterministic periodic function. More precisely, for each gene g and expression level observation at time t_i, denote the time-course data by $Y_g(t_i)$ with the following model:

$$Y_g(t_i) = f_g(t_i) + \epsilon_g(t_i) \quad i = 1, \cdots, N, \;\; g = 1, \cdots, G, \tag{1}$$

where $f_g(t_i)$ is a periodic function with period T_g and ϵ_g is the error around the function f_g. Moreover, G and N stand for numbers of genes and times respectively. The Lomb-Scargle periodogram at a frequency ω is defined as follows:

$$P_g(\omega) = \frac{1}{2\hat{\sigma}^2} + \{\frac{(\sum_{i=1}^{N}(y_g(t_i) - \bar{y_g})\cos[\omega(t_i - \tau)])^2}{\sum_{i=1}^{N}\sin^2[\omega(t_i - \tau)]} + \frac{(\sum_{i=1}^{N}(y_g(t_i) - \bar{y_g})\sin[\omega(t_i - \tau)])^2}{\sum_{i=1}^{N}\cos^2[\omega(t_i - \tau)]}\}, \quad (2)$$

where $\hat{\sigma^2} = \frac{1}{N-1}\sum_{i=1}^{N}(y_g(t_i) - \bar{y_g})^2$, $\bar{y_g} = \frac{1}{N}\sum_{i=1}^{N} y_g(t_i)$ and τ is defined by $\tan(2\omega\tau) = \frac{\sum_{i=1}^{N}\sin(2\omega t_i)}{\sum_{i=1}^{N}\cos(2\omega t_i)}$. Scargle (1982) [12] showed that under the assumption of non-periodicity, the distribution of Lomb-Scargle periodogram at a given frequency ω is exponentially distributed. Therefore, the p-value according to the test of the existence of a cyclic regulation at the frequency ω_k among M frequencies $\omega_1, \cdots, \omega_M$ in which $\max_{1 \le j \le M} P_g(\omega_j) = P_g(\omega_k)$, is given by p-value $= 1 - (1 - exp(-x_g))^M$, where $x_g = P_g(\omega_k)$. The smaller values of p-value (p-value < 0.05) show the existence of at least a peak at some ω_j in Lomb-Scargle periodogram that shows the periodicity related to those frequencies.

Jonckheere-Terpstra-Kendall (JTK) method

For k independent samples from k groups, Jonckheere-Terpstra (JT) test is the nonparametric one for detecting monotonic orderings of data over k groups. More precisely, if $F_1, \cdots, F_k$ are distribution functions corresponding to k groups, the JT method tests the hypothesis of: $H_0 : F_1(x) = \cdots = F_k(x)$ versus $H_A : F_1(x) \le \cdots \le F_k(x)$ or $H_A : F_1(x) \ge \cdots \ge F_k(x)$. This test can be applied for testing the hypothesis or periodicity in a set of time-course gene expression data. Let us assume that there is a periodic pattern in every time-course gene expression data. It is also assumed that all periods take place in an interval with a predetermined period length. If a periodicity with period T exists in time-course gene expression data, it is expected that a significant correlation exists between expression values and cosine curve with period T. Then, an order is assigned to each data according to cosine curve (for example higher orders to higher values of cosine function etc.). Suppose that the orders of data range from 1 to k. Therefore, the data are categorized into k separated groups according to their orders. Now, JTK test can be used for measuring periodicity pattern in time-course expression with period T. By changing the amount of T in the interval of fixed periods and calculating the p-value for each test, one can find the optimum value of period for each gene. To sum up, the test with minimum p-value identifies the optimum value of a period.

Three types of cyclic regulation

According to our study of many time-course gene expression data sets, the Lomb-Scargle method is more conservative than JTK to determine the cyclic regulation in time-course gene expression data. In other words, JTK method reports a gene to be periodic but the Lomb-Scargle would not confirm the periodicity. Therefore, we define three types of periodic time-course patterns.

Periodic time-course pattern type 1: The periodicity that is confirmed by Lomb-scargle periodogram method.

Periodic time-course pattern type 2: The periodicity that is recognized by JTK but not Lomb-Scargle method.

Periodic time-course pattern type 3: The periodicity that is recognized neither by JTK nor by Lomb-Scargle method.

By categorizing genes in aforementioned groups, we investigate the role of periodicity strength in discrimination analysis of genes based on their time-course expression.

3. NUMERICAL COMPUTATION

A numerical study was conducted of the set of *S. cerevisiae* gene expression data to show the effect of the type of periodicity on the accuracy of discrimination.

S. cerevisiae data set

Microarray experiments measure RNA levels of several thousands of genes during a time interval. Spellman et al, (1998) [3] studied cell-cycle dependent genes expression in the budding yeast *S. cerevisiae* under different synchronizations. Based on the combination of periodicity and correlation algorithms, 800 genes were identified that related to cell-cycle regulation. In this paper, the time-course genes expression data under $cdc15$ synchronization were analyzed using Fourier transform. An expression profile for each gene for the five stages ($G1$, S, $S/G2$, $G2/M$ and $M/G1$) of the cell cycle were obtained. Then, genes were assigned to one of the five stages according to the time of peak expression across the experiment. Table 1 presents the distribution of the 800 genes to each cell cycle stage.

Table 1. Numbers of gene expression profiles

Phase	G1	G2/M	M/G1	S	S/G2
Number of genes	299	192	112	70	121

Discrimination of genes

According to the gene expression profiles, discrimination analysis was performed to explore the influence of the periodicity strength on gene classification. A fraction of genes with different types of periodicity were used as test set and the rest of genes were used as train set. Since all genes are presumed to have the cyclic expression, the amplitude and phase shift were computed according to time-course expression. The value of phase shift approximately represents the profile of gene. Therefore, genes were discriminated according to both the amount of phase shift and the time-course expression. Using the Lomb-scargle and JTK methods, the type of periodicity was determined. Table 2 illustrates the distribution of genes corresponding to periodicity type and profile names.

Table 2. Numbers of genes according to their periodicity type and profile names

	G1	G2/M	M/G1	S	S/G2	Total number of genes
Type 1	41	43	31	16	22	153
Type 2	5	4	1	1	1	12
Type 3	253	145	80	53	98	629

Figure 1 demonstrates the difference between profiles after discrimination of genes based on time-course expression.

.

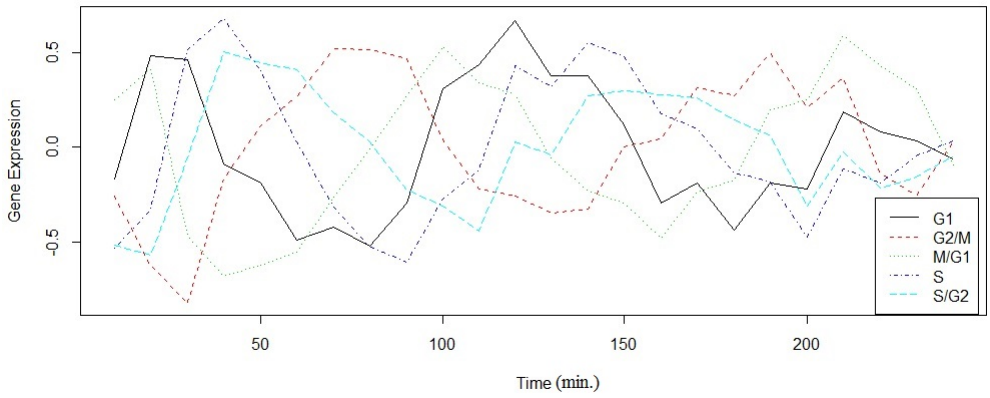

Fig. 1. Time-course centers of genes expression profiles.

Moreover, table 3 shows the true discovery rates for genes in each type of periodicity. The true discovery rate for genes with type 3 periodicity is significantly less than types 1 and 2. Also, the genes with periodicity types 1 and 2 are assigned to true profiles with the same high accuracy. The genes with type 3 periodicity have a low cyclic pattern and as a result, the variation around the cosine curve is high. This is more apparent when a number of phase shifts of genes considered instead of time-course expression data. Based on the phase shifts, the false discovery rates rise for type 1 and 2 as well as for type 3. Note that 79% of genes have type 3 periodicity and can deviate the centers of profiles during the discrimination. Using phase shifts, the deviation of centers from true values is more considerable.

Table 3. True discovery rates of genes according to their periodicity

	Type 1	Type 2	Type 3
Time-course Expression	80	80	55
Phase Shift	58	60	40

4. CONCLUSION

In this article, we introduced three types of periodicity in time-course genes expression. Then, the effect of periodicity strength on the accuracy of discrimination was investigated. We concluded that the false discovery rates for the genes with the weak cyclic pattern are noticeably high when the data set consists of a considerable number of this type of genes. It is worth mentioning that discrimination analysis is a supervised method to discover groups. For cluster analysis that is an unsupervised method, this problem can be more influential; in which the cluster analysis may produce unreliable groups. The present work emphasizes that for similar studies, the periodicity strength should be considered in analyses.

REFRENCES

1. Fokianosa K., Promponas V.J. Biological applications of time series frequency domain clustering. *Journal of Time series Analysis*. 2011. P. 744. P. 21.
2. Gonze D., Pinloche S., Gascuel O., Van Helden J. Discrimination of yeast genes involved in methionine and phosphate metabolism on the basis of upstream motifs. *Bioinformatics*. 2005.
3. Spellman P.T., Sherlock G., Zhang M.Q., Iyer V.R., Anders K., Eisen M.B., Brown P.O., Botstein D., Futcher B. Comprehensive identification of cell cycle-regulated genes of the yeast *saccharomyces cerevisiae* by microarray hybridization. *Molecular Biology of the Cell*. 1998. P. 3273.
4. Eisen M.B., Spellman P.T., Brown P.O., Botstein D. Cluster analysis and display of genomewide expression patterns. *Proceedings of the National Academy of Sciences*. 1998. V. 95. P. 14863–14868.
5. Glynn E.F., Chen J., Mushegian A.R. Detecting periodic patterns in unevenly spaced gene expression time series using Lomb-Scargle periodograms. *Bioinformatics*. 2006. V. 22. P. 310.
6. Zhao W., Serpedin E., Dougherty E.R. Spectral preprocessing for clustering time-series gene expressions. *EURASIP Journal on Bioinformatics & Systems Biology*. 2009.
7. Nugent R., Meila M. An overview of clustering applied to molecular biology. *Methods in Molecular Biology*. 2010. P. 369.
8. Wichert S., Fokianos K., Strimmer K. Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*. 2004. P. 5.
9. Zhao L.P., Prentice R., Breeden L. Statistical modeling of large microarray data sets to identify stimulus response profiles. *Proc. Natl Acad. Sci. USA*. 2001. V. 98. No. 10. P. 5631–5636.
10. Ahdesmaki M., Lahdesmaki H., Pearson R., Huttunen H., Yli-Harja O. Robust detection of periodic time series measured from biological systems. *BMC Bioinformatics*. 2005. V. 6. P. 117.
11. Hughes M.E., Hogenesch J.B., Kornacker K. JTK-CYCLE: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. *Journal of biological rhythms*. 2010. P. 372.
12. Scargle J.D. Statistical aspects of spectral analysis of unevenly spaced data. *Astrophys. J.* 1982. P. 835.

Accepted 16.02.2017.
Revised 10.04.2017.
Published 18.05.2017.