Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 52 (2008) 1387–1398
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

# Testing the significance of cell-cycle patterns in time-course microarray data using nonparametric quadratic inference functions

Guei-Feng Tsai[a], Annie Qu[b,*]

[a] *Center for Drug Evaluation, Taipei, Taiwan*
[b] *Department of Statistics, Oregon State University, Corvallis, OR 97331, USA*

Received 25 June 2006; received in revised form 20 March 2007; accepted 20 March 2007
Available online 24 March 2007

**Abstract**

We develop an approach to analyze time-course microarray data which are obtained from a single sample at multiple time points and to identify which genes are cell-cycle regulated. Since some genes have similar gene expression patterns, to reduce the amount of hypothesis testing, we first perform a clustering analysis to group genes into classes with similar cell-cycle patterns, including a class with no cell-cycle phenomena at all. Then we build a statistical model and an inference function assuming that genes within a cluster share the same mean model. A varying coefficient nonparametric approach is employed to be more flexible to fit the time-course data. In order to incorporate the correlation of longitudinal measurements, the quadratic inference function method is applied to obtain more efficient estimators and more powerful tests. Furthermore, this method allows us to perform chi-squared tests to determine whether certain genes are cell-cycle regulated. A data example on cell-cycle microarray data as well as simulations are illustrated.


*Keywords:* Cell-cycle microarray data; Gene grouping; Varying coefficient model; Quadratic inference function; Chi-squared test

## 1. Introduction

Microarray has become a powerful tool in molecular biology because it can measure gene expression levels for thousands of genes simultaneously. In this paper we are interested in time-course microarray data. There are two types of time-course microarray data, longitudinal data with no replication which are obtained from a single sample at multiple time points, and longitudinal data with replication obtained from multiple samples at multiple time points (Qin and Self, 2006). Our main focus is on longitudinal data with no replication, and in particular on yeast cell-cycle microarray data. Cell-cycle microarray data has measurements of gene expressions over multiple time points during the progression of the cell cycle. Regulation of cell cycles plays an important role in the normal development of multicellular organisms, since diseases such as cancer are a consequence of uncontrolled cell growth.

The primary goal in analyzing cell-cycle microarray data is to identify cell-cycle regulated genes, where expression levels change periodically during the cell cycle (Cho et al., 1998; Spellman et al., 1998). A number of statistical methods have been proposed for this purpose, including Fourier transformation (Spellman et al., 1998), the single

* Corresponding author. Tel.: +1 541 7372239; fax: +1 541 7373489.
*E-mail address:* qu@science.oregonstate.edu (A. Qu).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.03.018

pulse model (Zhao et al., 2001), a nested model using the idea of principal component analysis (Li et al., 2002), and a nonparametric model approach using B-spline functions (Luan and Li, 2004). Despite the different forms of these models, these approaches all test each gene individually whether they are cell-cycle regulated or not. Under their settings, thousands of hypotheses are formulated and are assumed to be independent from each other, and the false discovery rate (Benjamini and Hochberg, 1995) might be either inflated or underestimated based on the theoretical null where genes are assumed to be independent (Efron, 2005). In addition, most of these methods also do not take within-gene correlation into account, which could lead to inefficient estimation and inference.

Clustering techniques are commonly used for reducing the complexity of large data sets and are also effective for grouping genes with similar expression patterns. A number of clustering algorithms have been proposed for cell-cycle microarray data, including hierarchical clustering (Eisen et al., 1998), k-means (Tavazoie et al., 1999), self-organizing maps (Tamayo et al., 1999), model-based clustering (Yeung et al., 2001), and the clustering of regression models (Qin and Self, 2006). Although these methods provide useful clustering techniques and visual detection of expression patterns, a statistical method to test whether genes are cell-cycle regulated is still lacking. Additionally, these clustering methods assign all genes into clusters. However, in microarray experiments, many genes may show some variations unrelated to any clusters. These gene, called sporadic genes, should not be assigned to specific clusters (Tseng and Wong, 2003). If a cluster contains sporadic genes, the pattern may be misrepresented.

To reduce the complexity of high-dimensional data and make inferences for time-course microarray data, we propose an approach which combines a clustering method, time-varying coefficients modeling, and the quadratic inference function (QIF). First, we group genes with similar cell-cycle patterns into the same class, or into a class with no cell-cycle phenomena at all. Then a marginal varying coefficient model is fitted for genes within the same cluster by assuming that genes within the same cluster share the same mean model. In order to incorporate within-gene correlations, the QIF is applied to estimate the regression coefficients. The QIF also provides an inference function for testing whether genes within the same cluster are cell-cycle regulated or not.

The proposed approach has several advantages. First, it groups genes with similar cell-cycle patterns together and therefore reduces the number of hypotheses and minimizes inaccurate estimation of false positive rates. Scientifically, it is also more meaningful to examine genes with similar cell-cycle patterns together instead of a single gene, since there might be an association between genes within the same cluster. Secondly, it employs nonparametric modeling which allows more flexibility for time-course data than the parametric approach. Specifically, we use a time-varying coefficient model with a polynomial truncated spline basis function with knots and also with periodic basis functions. Applying periodic basis functions in a varying coefficient model is equivalent to testing whether or not genes are cell-cycle regulated, since an appropriate periodic basis function is able to capture cell-cycle phenomena. In addition, it does not require the specification of the likelihood function, but also enables us to account for within-gene correlations. Lastly, it provides an asymptotic distribution of test statistics to test whether genes have cell-cycle patterns.

The outline of this article is as follows. In Section 2.1, we describe a clustering technique to classify genes into classes with similar expression patterns. The nonparametric varying coefficient model is proposed in Section 2.2. Section 2.3 provides the QIF approach for cell-cycle data. The simulation results are illustrated in Section 3. Section 4 demonstrates a cell-cycle microarray data example using our approach. A brief discussion is provided in Section 5.

## 2. Methods

### *2.1. Gene clustering analysis*

The primary goal of cell-cycle microarray experiments is to investigate whether gene expressions change over time. This is a challenge for statistical modeling because there are large numbers of parameters associated with genes, yet sometime there are no replications. Furthermore, gene expressions may have very different patterns with large heterogeneous variation between different groups of genes. For that reason, we first classify genes with similar expression patterns into the same group so that each group has a more homogeneous pattern.

Most clustering algorithms assign all genes into clusters which include assigning unrelated sporadic genes to any specific cluster (Tseng and Wong, 2003). If a cluster contains sporadic genes, the pattern may be misrepresented. The objective of our clustering is to avoid unrelated genes being assigned into the same cluster and to ensure that only clusters of genes with similar patterns will be formed.