Available online at www.sciencedirect.com
ScienceDirect
Computational Biology and Chemistry
Computational Biology and Chemistry 31 (2007) 265–274
www.elsevier.com/locate/compbiolchem

Clustering of time-course gene expression data using functional data analysis

Joon Jin Song [a,*], Ho-Jin Lee [b], Jeffrey S. Morris [c], Sanghoon Kang [d]

[a] *Department of Mathematical Sciences, University of Arkansas, Fayetteville, AR 72701, USA*
[b] *Schering-Plough Research Institute, 2015 Galloping Hill Road, K-15-2-2125, Kenilworth, NJ 07033-1300, USA*
[c] *Department of Biostatistics and Applied Mathematics, The University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, USA*
[d] *Institute for Environmental Genomics, The University of Oklahoma, Norman, OK 73019, USA*

Received 18 April 2007; received in revised form 29 May 2007; accepted 30 May 2007

Abstract

Clustering of gene expression data collected across time is receiving growing attention in the biological literature since time-course experiments allow one to understand dynamic biological processes and identify genes governed by the same processes. It is believed that genes demonstrating similar expression profiles over time might give an informative insight into how underlying biological mechanisms work. In this paper, we propose a method based on *functional data analysis* (FNDA) to cluster time-dependent gene expression profiles. Consideration of clustering problems using the FNDA setting provides ways to take time dependency into account by using basis function expansion to describe the partially observed curves. We also discuss how to choose the number of bases in the basis function expansion in FNDA. A synthetic cycle data and a real data are used to demonstrate the proposed method and some comparisons between the proposed and existing approaches using the adjusted Rand indices are made.

Keywords: Time-course gene expression; Functional data analysis; Clustering; Principal component analysis

1. Introduction

Microarray technologies in molecular biology make it possible to simultaneously measure the expression levels of thousands of genes for a certain organism. They allow us to gain biological insight at the genome scale and to study the behaviour of thousands of genes simultaneously, under various conditions. Gene expression can be examined from two points of view, static and dynamic. The gene expression in static microarray experiments is a snapshot at a single time, whereas, in time-course experiments the expression profiles of genes are repeatedly measured over a time period. In particular, time-course microarray experiments are effective not only in studying gene expression profile levels over a period of time but also in exploring functions of genes and the interactions with their products. Since biological processes are dynamic and complex systems, such characteristics are essential factors in understanding how the underlying mechanisms regulate cellular processes and gene functions. Time-course microarray experiments are the tools for understanding temporal patterns of gene expression and detecting periodically expressed genes.

A number of statistical methods have been recently proposed to analyze time-course gene expression data. Peddada et al. (2003) proposed the order-restricted inference method to cluster and select genes in accordance with temporal or dose profiles arisen from microarray experiments. However, the approach resulted in that the gene profiles with a monotonic pattern but distinct accelerations in the profiles are identified as the same cluster. Johansson et al. (2003) treated genes as variables and employed the method of partial least squares to identify genes with periodic fluctuations in expression levels, coupled with the cell cycle in the budding yeast. The measure used for gene selection was the magnitude of frequencies of sinusoidal functions that fit the cyclically expressed data. Schliep et al. (2003) used Hidden Markov Models (HMM) that take time dependency of time-course data into account, where a set of clusters was obtained by the method of maximum likelihood. Luan and Li (2003) introduced the mixed-effects model using B-splines to analyze time-course gene expression data and carried out gene clustering in the framework of a mixture model. The clustering

* Corresponding author. Tel.: +1 479 575 6319; fax: +1 479 575 8630.
E-mail address: jjsong@uark.edu (J.J. Song).

1476-9271/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2007.05.006

problem is viewed as a mixture model problem by introducing the cluster indicator to be estimated and to be treated as missing data in the estimation of the parameter associated with a mixture model using the EM algorithm. They also compared the proposed method with the model-based clustering method proposed by Fraley and Raftery (2002).

In this paper, we propose a unified approach for gene clustering and dimension reduction based on *functional data analysis* (FNDA) to group observed curves with respect to their shapes or patterns by using the sample information in time-course microarray experiments.[1] The fundamental idea behind FNDA is that the atom, or unit of observation, is considered to be the entire curve rather than just a set of observations (Ramsay and Silverman, 1997, 2002). Our clustering is built upon a basis-space approach, which reduces the dimensionality of the data and allows the time points to be non-equally spaced and to vary between subjects.

We apply this method to a time course microarray data set on the yeast cell cycle, and demonstrate that our method is able to identify tight clusters of genes with expression profiles focused on particular phases of the cell cycle.

2. Methods

2.1. Functional data analysis

Functional data refer to data in which each observation is a partially observed function or curve on some interval where these functions are often assumed to be smooth. What distinguishes FNDA from other conventional statistics is the datum or data unit. Many statistical methods treat numbers or vectors as the units of data. In FNDA, however, data units are functions or curves defined on some interval, rather than focusing on the observed values at particular points in the interval. The nature of functional data makes it necessary to consider function spaces such as Hilbert spaces, and each functional observation is viewed as a realization generated by a random mechanism in these spaces. The books by Ramsay and Silverman (1997, 2002) give useful accounts of the basic considerations of FNDA.

FNDA has a wide range of flexibility in the sense that the observation times are not required to be equally spaced for the subjects, and furthermore, these times can vary from one subject to another. Functional data do not necessarily assume that the values observed at different times for a single subject are independent although it often assumes that data from different subjects are independent.

Consider the situation where we observe sample curves which are partially observed on the subset of the interval. Let $\{X(t), t \in T\}$ be a second order stochastic process defined on T, e.g., $X \in L^2[a, b]$. The stochastic process is a collection $\{X(t), t \in T\}$ defined on a common probability space (Ω, F, P), where (Ω, F) is a measurable space and P is a measure on F with $P(\Omega) = 1$. In order to clarify the use of the index sets in stochastic processes, one needs to write $X(t)$ as a function $X(\omega, t)$ of two variables, where t is the time and $\omega \in \Omega$ is the random element. For fixed $t \in T$, the function $X(\cdot, t)$ is a measurable map from Ω into $\Re$. For fixed $\omega \in \Omega$, the function $X(\omega, \cdot)$ becomes a sample path of the stochastic process. Denoted by $\mu(t)$,

$$\mu(t) := EX(\omega, t) = \int X(\omega, t)\, \mathrm{d}F_X,$$

for fixed t, where F_X is the distribution function of a probability P on (Ω, F).

For fixed ω, a sample path $X(\omega, t)$ is an equivalent class of functions in the function space L^2. Since functions in the space can be expressed in terms of basis functions generating the space, a separable Hilbert space, each function in the space can be written as a countable linear combination of the basis functions. Let $\{\phi_k\}$ be a set of basis functions of L^2, then we see that for each $X(\omega, t)$ with fixed ω, there is a unique $\mathbf{c}^T = (c_1, c_2, \ldots) \in l^2$ such that

$$X(t) = \sum_{k=1}^{\infty} c_k \phi_k(t),$$

where l^2 is a discrete analogue of L^2 space. It should be emphasized that the stochastic process is decomposed into two parts c_k and $\phi_k(t)$, and the random mechanism only involves in the coefficients $c_k = c_k(\omega)$ unless setting ω to be fixed.

Once the representation by basis functions is adopted, three types of computational issues need to be addressed: (a) choosing an appropriate type of basis function, (b) determine the number of basis functions, and (c) computing the best linear combination.

The choice of the number of basis functions clearly has implications in determining the assumed underlying smoothness of the process and the degree of dimension reduction provided by the basis representation. Ramsay and Silverman (1997) suggest that 20–30 basis functions are in general enough to extract prominent features. In this paper, we propose a way to select the number of basis functions analogous to determining the number of clusters using the Bayesian information criterion (BIC) in model-based clustering illustrated below. In this context, the number of basis functions with the maximum BIC score is selected for representing functional data as basis functions.

Choosing a basis is a more controversial issue since no basis will be universally optimal for all data sets. However, there are advisable guidelines depending on specific occasions. For example, if the paths are uniformly smooth with limited features and especially if the curves appear to be periodic, then the Fourier basis seems to be a good choice. On the other hand, a spline basis or a wavelet basis may be a better choice if there are a number of local features which may be relevant for the statistical analysis. Note that for some basis functions, more computationally efficient alternatives are available (e.g., FFT for Fourier and DWT for wavelet). We may write

$$X(t) \approx \sum_{k=1}^{K} c_k \phi_k(t), \tag{1}$$

[1] FNDA is an acronym for functional data analysis instead of FDA because FDA traditionally stands for US Food and Drug Administration.