Gene expression

Clustering of change patterns using Fourier coefficients

Jaehee Kim¹,* and Haseong Kim²
¹Department of Statistics, Duksung Women’s University and ²Bioinformatics and Biostatistics Laboratory, Seoul National University, Seoul, S. Korea

Received on March 21, 2007; revised on November 7, 2007; accepted on November 8, 2007

Advance Access publication November 19, 2007
Associate Editor: Martin Bishop

ABSTRACT

Motivation: To understand the behavior of genes, it is important to explore how the patterns of gene expression change over a time period because biologically related gene groups can share the same change patterns. Many clustering algorithms have been proposed to group observation data. However, because of the complexity of the underlying functions there have not been many studies on grouping data based on change patterns. In this study, the problem of finding similar change patterns is induced to clustering with the derivative Fourier coefficients. The sample Fourier coefficients not only provide information about the underlying functions, but also reduce the dimension. In addition, as their limiting distribution is a multivariate normal, a model-based clustering method incorporating statistical properties would be appropriate.

Results: This work is aimed at discovering gene groups with similar change patterns that share similar biological properties. We developed a statistical model using derivative Fourier coefficients to identify similar change patterns of gene expression. We used a model-based method to cluster the Fourier series estimation of derivatives. The model-based method is advantageous over other methods in our proposed model because the sample Fourier coefficients asymptotically follow the multivariate normal distribution. Change patterns are automatically estimated with the Fourier representation in our model. Our model was tested in simulations and on real gene data sets. The simulation results showed that the model-based clustering method with the sample Fourier coefficients has a lower clustering error rate than K-means clustering. Even when the number of repeated time points was small, the same results were obtained. We also applied our model to cluster change patterns of yeast cell cycle microarray expression data with alpha-factor synchronization. It showed that, as the method clusters with the probability-neighboring data, the model-based clustering with our proposed model yielded biologically interpretable results. We expect that our proposed Fourier analysis with suitably chosen smoothing parameters could serve as a useful tool in classifying genes and interpreting possible biological change patterns.

Availability: The R program is available upon the request.

Contact: jahee@duksung.ac.kr

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

One of the most common approaches in genome science is the analysis of gene expression patterns or change patterns. If the expressions of genes are measured at various time points during the course of an experimental study, each gene can be characterized by its pattern of change. Grouping genes that share similar expression profiles into clusters is usually the first step in understanding the huge amount of DNA microarray data associated with complicated biological networks. However, most research on gene clustering has been performed with the observed expression data, while ignoring the change patterns. The motivation of this research is to derive an efficient and robust statistical method for an area where little research has been done yet the needs from a biological standpoint are numerous.

The researcher is frequently interested in studying gene expression changes along time and evaluating trend differences between the various experimental groups. For example, rather than comparing and grouping genes with the same pattern from corn oil or fish oil diet on colon cancer, and only considering observation values, change values and change patterns need to be investigated. The researcher is then interested in detecting biologically meaningful gene expression trends and in spotting differences between the various experimental groups.

Due to the differences in the initial levels of background noise in the experiment, difference values or derivatives need to be used as a measure of change. Also a basic premise is that the genes sharing similar change profiles may be functionally related or co-regulated. As such, microarray derivative data provide further insight into gene–gene interactions, gene functions and pathways. Derivative functions also provide statistical convenience in that: (1) functions with a constant amount of difference have the same derivatives (2) difference values give information about their changes as well as about their original functions. Nevertheless, few of the previous methods took derivative functions into account.

We propose to use Fourier coefficients in clustering expression patterns and change patterns. Fourier coefficients have several advantages over other methods. Some of these advantages are: (1) the dimension of a data set can be reduced to several Fourier coefficients (2) the estimated Fourier coefficients give information about the underlying function and enable automatic estimation of the change or pattern function (3) the Fourier coefficient estimation does not depend
strongly on the covariance structure (4) as the sample Fourier coefficients asymptotically follow the multivariate normal distributions. A Gaussian mixture model that incorporates underlying probability distributions can be effective.

There has been considerable research about discovering patterns using clustering and testing. Serban and Wasserman (2005) proposed clustering after transformation and smoothing as a technique for non-parametrically estimating and clustering a large number of curves. To discover change patterns in gene expressions, Ernst et al. (2005) clustered short time series gene expression data by selecting a set of potential expression profiles. Li and Wong (2002) proposed an effective discretization and gene selection method using the concept of emerging patterns. Park et al. (2003) proposed a method to test the statistical significance of time-dependent gene expression data and to identify genes with significant change based on an ANOVA model. Lai et al. (2004) proposed a method for selecting genes that have differential gene–gene co-expression patterns with the idea of correlation difference.


Smoothing away noise-induced wiggles with Fourier series has been studied by some researchers. Zhang et al. (2003) used the first harmonic of discrete Fourier transform to translate the multi-dimensional time series microarray expression data into a two-dimensional scatter plot. Murthy and Hua (2004) proposed improved Fourier method considering irregular or monotonic component of cell-cycle expression. Kim et al. (2006) suggested a two-step procedure for clustering periodic patterns of gene expression profiles. They used the least squares non-linear curve fitting based on a Fourier series approximation with frequency and amplitude of order one. Though they considered the periodicity and mixture model-based likelihood for the estimated parameters, change patterns of the gene expression were not taken into account.

There has been much research on clustering microarray data, mostly on grouping common expression patterns. However, there are many cases in gene study in which grouping change patterns is of interest. In this research, we propose a new method for clustering change patterns with derivative Fourier coefficients. The proposed method consists of four main steps. The first and second steps consist of representing a gene profile with sample Fourier coefficients, and then the calculation of derivatives from the Fourier coefficients. The third step is to cluster the derivative Fourier coefficients using model-based clustering. In the final step, genes with the same change pattern are clustered and the underlying change pattern is automatically estimated using the Fourier representation.

We demonstrated the usefulness of the Fourier analysis and model-based clustering by applying the method to simulated data. We also extended the application of our model to real gene expression data resulting in interpretable genes.

2 THE MODEL

Consider the data $Y_{iu}$, $u$th observation on the $i$th curve, of the form

$$Y_{iu} = f(t_{iu}) + e_{iu}, \quad i = 1, 2, \ldots, n, u = 1, 2, \ldots, m$$

where $E(e_{iu}) = 0$ and $\text{Var}(e_{iu}) = \sigma^2$. In the microarray experiment $Y_{iu}$ is the log gene expression of gene $i$ at time $t_{iu}$.

We assume that the curve $f_i$ belongs to a class of smooth functions $F$ as defined below:

$$f_i(t) = \int \phi(t) b_i(t) dt, \quad (2)$$

where $\{b_i\}$ is an orthonormal basis system and $\phi(t) = \int f_i(t) b_i(t) dt$.

We can estimate $f_i$ using Fourier coefficients by

$$\hat{f}_i(t) = \sum_{j=1}^J \hat{\phi}_j b_j(t), \quad (4)$$

which is the projection onto the first $J$ basis functions where $J, 1 \leq J \leq m$, is a smoothing parameter to be chosen based on the data.

The sample Fourier estimate can be estimated as

$$\hat{\phi}_j = \frac{1}{m} \sum_{i=1}^m Y_{iu} b_j(t_i), \quad (5)$$

with $t_i = r/m$ and $t \in [0,1]$.

With regard to changes, the difference data

$$D_{iu} = Y_{iu} - Y_{iu-1} \quad (6)$$

can be approximated by $f_i(t_{iu})$, the derivative of $f_i$ at $t_{iu}$ and $t_{iu} - t_{iu-1}$, assuming that the first order derivative exists. Therefore, the following model can be considered:

$$D_{iu} = \frac{1}{m} f_i(t_{iu}) + \eta_{iu}, \quad i = 1, 2, \ldots, n, u = 2, \ldots, m \quad (7)$$

where

$$\eta_{iu} = e_{iu} - e_{iu-1}$$

This setup can be extended to the cases where the design or time points are not the same for all curves. We want to classify the same patterns with differences or derivatives that give information about the underlying change pattern.

3 TRIGONOMETRIC FOURIER SERIES

ESTIMATION OF DERIVATIVES

The function represented with a Fourier series with the cosine bases is given as

$$f_i(t) = \phi_0 + \sum_{j=1}^{\infty} \phi_j \sqrt{2} \cos \pi j t. \quad (8)$$

We can estimate $f_i$ with $J$ terms of Fourier coefficients as

$$\hat{f}_i(t) = \hat{\phi}_0 + \sum_{j=1}^{J} \hat{\phi}_j \sqrt{2} \cos \pi j t. \quad (9)$$
where Fourier coefficients are estimated as
\[ \hat{\phi}_j = \frac{1}{m} \sum_{i=1}^{m} Y_i \sqrt{2} \cos j\pi t_i, \quad j = 0, 1, 2, \ldots, m. \] (10)

We also estimate the derivative of \( f_j \) as
\[ \hat{f}_j(t) = \frac{\hat{\phi}_j(t)}{d t} = \pi \sum_{j=1}^{j} \hat{\phi}_j \sqrt{2} \sin j\pi t. \] (11)

Note that the Fourier coefficients of the derivatives are calculated by weighting the coefficients from the original functions. The coefficients of the derivative have more weight \( j \) on the latter terms of the Fourier coefficients. This means that the higher frequency terms have more information about the derivative pattern of ups and downs.

The model in (7) can be expressed as
\[ D_{\text{in}} = \pi \sum_{j=1}^{j} \psi_j \sqrt{2} \sin j\pi t + \eta_{\text{in}} \] (12)

where \( \psi_j = (j/m) \hat{\phi}_j \). Therefore the Fourier coefficients of change can be estimated by \( \hat{\psi}_j = (j/m) \hat{\phi}_j \). Since \( \psi_j \) is a Fourier coefficient of the derivative function, we call \( \hat{\psi}_j \) the derivative Fourier coefficient and \( \hat{\psi}_j \) the estimated derivative Fourier coefficient from the sample.

With the independent \( \varepsilon_j \)'s, \( \text{var}(\eta_{\text{in}}) = 2\sigma^2 \), and
\[ \text{Cov}(\eta_{\text{in}}, \eta_{\text{in}}) = \begin{cases} -\sigma^2, & l = j - 1 \\ 0, & \text{otherwise}. \end{cases} \]

To estimate Fourier coefficients, the covariance structure is not considered in our approach since the covariance matrix of a finite set of estimated Fourier coefficients is asymptotically proportional to the identity matrix.

The parameter \( J \) controls the amount of smoothing and should be determined based on the data. Even though the optimal choice for \( J \) varies from function to function, we choose to use a single smoothing parameter that operates reasonably well for all of the curves. There has been some research on optimal choices for \( J \). For example, to find global smoothing parameter, Serban and Wasserman (2005) calculated \( J \) as the minimizer of the total regret. Eubank and Hart (1992) also suggested choosing the smoothing parameter \( J \) minimizing the risk or mean-squared error.

With a large number of gene curves and various functional shapes, a universal rule for an optimal choice for \( J \) does not exist. Therefore, instead, we capitalize on the convergence property of Fourier transforms. Since the Fourier estimator converges to the true function, usually the first few Fourier coefficients contribute to the estimation of the whole function. In practice, we can select a smaller \( J \) for linear or smooth functions and a larger \( J \) for wigglier functions.

### 4 CLUSTERING CURVES OF THE SAME CHANGE

The similarity of cluster derivatives \( \psi_i = (\psi_{i1}, \psi_{i2}, \ldots, \psi_{iJ}) \) and \( \psi_j = (\psi_{j1}, \psi_{j2}, \ldots, \psi_{jJ}) \) can be measured with Euclidean distance or other distance measures. It may be of interest to check the equivalence of the similarity of the estimated Fourier coefficients with the similarity of the estimated functions. As Beran and Dumbgen (1998) demonstrated, a reasonable coordinate system via a Fourier transform of data has as much correct asymptotic coverage probability as the untransformed data. As such, the sample Fourier coefficients can be used instead of the underlying functions.

After clustering with the estimated Fourier coefficients \( \hat{\phi}_j \)'s for the original function and with \( \hat{\psi}_j \)'s for the derivative function, we can estimate the function of each gene with these estimated Fourier coefficients using (4). The change pattern can also be estimated with derivative Fourier coefficients using (11). This automatic estimation is another capability of Fourier representation. These estimated periodic functions show the functional shape and periodicity.

#### 4.1 Mixture model of derivative Fourier coefficients

Clustering using a mixture model assumes that each group of the data is generated by an underlying probability distribution. Suppose that data \( X_1, \ldots, X_n \) are multivariate observations.

In a Gaussian mixture model, each group \( k \) is modeled by the multivariate normal distribution with parameters \( \mu_k \) (mean vector) and \( \Sigma_k \) (covariance matrix):
\[ f_k(X_i | \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{n/2} |\Sigma_k|^{1/2}} \exp \left\{ -\frac{1}{2} (X_i - \mu_k)^T \Sigma_k^{-1} (X_i - \mu_k) \right\}. \] (13)

Geometric features (shape, volume, orientation) of each group \( k \) are determined by the covariance matrix \( \Sigma_k \). Banfield and Raftery (1993) proposed a general framework for exploiting the representation of the covariance matrix in terms of its eigenvalue decomposition. Each elliptical model is implemented in Mclust (Fraley and Raftery, 1999).

We consider model-based clustering with the estimated Fourier coefficients of change \( \psi_i = (\psi_{i1}, \psi_{i2}, \ldots, \psi_{iJ}) \). The sample Fourier coefficient \( \hat{\phi}_j \) in (5) is a form of weighted average of random variables with variance \( O(n^{-1}) \). Freedman and Lane (1980) showed that the empirical distribution of Fourier coefficients is normal. By Central Limit Theorem for independent and identically distributed samples, the sample Fourier coefficient \( \hat{\phi}_j \) is asymptotically normally distributed as \( n \to \infty \). As \( m \to \infty \) and fixed \( J \ll m \), \( \hat{\phi}_j = (\phi_{j1}, \phi_{j2}, \ldots, \phi_{jJ}) \) has an asymptotically, \( J \)-dimensional multivariate normal distribution. Therefore \( \psi_j = (\psi_{j1}, \psi_{j2}, \ldots, \psi_{jJ}) \) has an asymptotically multivariate normal distribution as a linear function of \( \phi_j = (\phi_{j1}, \phi_{j2}, \ldots, \phi_{jJ}) \). With this asymptotic property, we can use the Gaussian mixture model for clustering.

Model-based hierarchical agglomerative clustering is an approach to compute an approximate maximum of the classification likelihood,
\[ L_d(\theta_1, \ldots, \theta_J; g_1, \ldots, g_n | \psi) = \prod_{j=1}^{n} L_d(\psi_{i1}, \ldots, \psi_{iJ} | \theta_{j1}) \]

where the \( g_i \) are labels indicating a unique classification for each observation and \( L_d \) is the probability function of the estimated Fourier coefficients of the \( i \)-th gene. In the above Gaussian mixture likelihood, each component is weighted by the probability that a sample Fourier coefficient belongs to that component. Our clustering strategy is model-based.
Clustering of change patterns

agglomerative hierarchical clustering and selection of the model and the number of clusters using approximate Bayes factors with the BIC approximation.

4.2 Cluster validity

A major challenge in cluster analysis is the estimation of the optimal number of clusters. To identify the partition of clusters for which a measure of quality is optimal, as a cluster validity technique silhouette method was proposed by Kaufman and Rousseeuw (1987).

The silhouette width for the \( i \)th sample in the \( J \)th cluster is defined as:

\[
s(i) = \frac{b(i) - a(i)}{\max[a(i), b(i)]}
\]

where \( a(i) \) is the average distance between the \( i \)th sample and all other samples included in the \( J \)th cluster, \( b(i) \) is the minimum average distance between the \( i \)th sample and all of the samples clustered in \( k \)th cluster for \( k \neq J \). A point is regarded as well clustered if \( s(i) \) is large. The overall average silhouette value can be used as an effective validity index for any partition. Kaufman and Rousseeuw (1990) proposed choosing the optimal number of clusters as the value maximizing the average silhouette width.

The simulated data consist of 2000 curves originating from nine different functions, 1200 \( f_i \)'s and 100 curves of each \( f_2, \ldots, f_9 \), to reflect typical gene expression data. There are only five different change patterns \( f_1, f_2, f_4, f_5, f_6 \). We assume that the noise is normally distributed: for low noise \( \sigma \sim \text{Unif}(0,0.4) \) and for high noise \( \sigma \sim \text{Unif}(1.0,1.2) \). \( m = 5, 10, 20, 30, 50 \) repeated design points are considered.

Pollard (1982) showed that under weak conditions, as the sample size tends to infinity, the set of cluster centers as the minimizer of the distance of samples in each cluster converges almost surely to the population cluster centers and converges in distribution to the multivariate normal distribution. Since the means with K-means clustering can satisfy this property, we consider K-means clustering in the comparison study.

Let \( T \) be a clustering map defined as

\[
T(f, g) = \begin{cases} 
1, & \text{if } f \text{ and } g \text{ are in the same cluster} \\
0, & \text{otherwise.}
\end{cases}
\]

Regarding the estimation error, the clustering estimation error rate \( \eta(K) \) is defined as

\[
\eta(K) = \frac{1}{N^2} \sum_{i<j} I(T_K(f_i, f_j) \neq \hat{T}_K(f_i, f_j))
\]

where \( \hat{C} = \{f_1, \ldots, f_N\} \) denote the true curves and \( \hat{C} = \{\hat{f}_1, \ldots, \hat{f}_N\} \) denote the estimated curves. Let \( T \) and \( \hat{T} \) represent the corresponding cluster maps and \( K \) denotes the number of clusters. \( \eta(K) \) then is the fraction of all pairs that are incorrectly put in separate clusters depending on \( K \) clusters, as described in Serban and Wasserman (2005).

Table 1 shows the clustering estimation error rates for the model-based method and K-means methods with Fourier coefficients and also with difference data for the number of Fourier terms and the number of repeated design points. The clustering estimation error is smaller in the model-based method with the Fourier coefficients than in the K-means with the Fourier coefficients. Also the clustering estimation error is smaller in the model-based with Fourier coefficients than in the model-based with the difference data. With Fourier coefficients, the clustering estimation error becomes smaller as \( m \) becomes larger. The number of clusters is determined to be 5 in accordance with Bayesian Information Criteron (BIC). Once \( J \) exceeds 5, the clustering estimation error rate does not change appreciably. Therefore, we suggest using \( J \) around 5 for dimension reduction and to perform the biological interpretation. Optimal \( J \) values are highlighted in Table 1. Supplementary Table S1 shows the similar result as Table 1 with high noise data. Figure 1 shows the functions grouped in each cluster with \( J = 2 \) with low noise simulated data. It shows the true functions of the same derivatives.

5 RESULTS

5.1 Simulated data set

Since real expression data sets are generally noisy and their clusters may not be fully reflective of the class information, we first evaluate the performance of our method with simulated data, where the classes are known.

We simulate data according to the regression model

\[
Y_{iu} = f_i(t_{iu}) + \sigma e_{iu}
\]

with \( i = 1, 2, \ldots, 9 \), \( u = 1, 2, \ldots, m \), \( t_{iu} = u/m \). The regression functions for \( f_i \) are:

\[
\begin{align*}
  f_1(t) &= 0 \\
  f_2(t) &= \frac{2 - 5t}{2} \wedge \left( \left( \frac{5t - 2}{3} \right)^2 + \sin \frac{5\pi t}{2} \right) \\
  f_3(t) &= f_2(t) + 5 \\
  f_4(t) &= -f_2(t) \\
  f_5(t) &= f_2(t) - f_4(t) = f_4(t) - 5 \\
  f_6(t) &= 2 \cos(2\pi t) \\
  f_7(t) &= 5 + f_6(t) \\
  f_8(t) &= 10 t \\
  f_9(t) &= 5 + f_6(t).
\end{align*}
\]

The simulated data consist of 2000 curves originating from nine different functions, 1200 \( f_i \)'s and 100 curves of each \( f_2, \ldots, f_9 \), to reflect typical gene expression data. There are only five different change patterns \( f_1, f_2, f_4, f_5, f_6 \). We assume that the noise is normally distributed: for low noise \( \sigma \sim \text{Unif}(0,0.4) \) and for high noise \( \sigma \sim \text{Unif}(1.0,1.2) \). \( m = 5, 10, 20, 30, 50 \) repeated design points are considered.

Pollard (1982) showed that under weak conditions, as the sample size tends to infinity, the set of cluster centers as the minimizer of the distance of samples in each cluster converges almost surely to the population cluster centers and converges in distribution to the multivariate normal distribution. Since the means with K-means clustering can satisfy this property, we consider K-means clustering in the comparison study.

Let \( T \) be a clustering map defined as

\[
T(f, g) = \begin{cases} 
1, & \text{if } f \text{ and } g \text{ are in the same cluster} \\
0, & \text{otherwise.}
\end{cases}
\]

5.2 Yeast cell cycle microarray expression data example

5.2.1 Yeast cell cycle data

We also applied our method to yeast cell cycle data. Cell cycle is important in understanding cell replication, malignancy and reproductive disease that are associated with genomic instability and abnormal cell division. Biologists have been studying the cell cycle with budding yeast.
Table 1. Comparison of clustering estimation error rate (%) of model-based versus K-means cluster method with derivative Fourier coefficients (FC) and difference data (dif) in 5 clusters of the simulated data with low noise level in 100 repetitions and \( m \) repeated design points

<table>
<thead>
<tr>
<th>( J )</th>
<th>Method</th>
<th>( m = 5 )</th>
<th>( m = 10 )</th>
<th>( m = 20 )</th>
<th>( m = 30 )</th>
<th>( m = 50 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Mc(FC)</td>
<td>63.69</td>
<td>60.14</td>
<td>58.55</td>
<td>58.09</td>
<td>58.03</td>
</tr>
<tr>
<td></td>
<td>Kc(FC)</td>
<td>74.23</td>
<td>70.68</td>
<td>69.44</td>
<td>66.59</td>
<td>65.55</td>
</tr>
<tr>
<td>Low Noise</td>
<td>Mc(FC)</td>
<td>42.51</td>
<td>31.07</td>
<td>18.03</td>
<td>24.07</td>
<td>15.03</td>
</tr>
<tr>
<td>2</td>
<td>Kc(FC)</td>
<td>76.23</td>
<td>58.60</td>
<td>54.02</td>
<td>52.18</td>
<td>54.37</td>
</tr>
<tr>
<td></td>
<td>Mc(FC)</td>
<td>47.00</td>
<td>40.30</td>
<td>25.85</td>
<td>15.59</td>
<td>9.31</td>
</tr>
<tr>
<td>3</td>
<td>Kc(FC)</td>
<td>76.13</td>
<td>72.06</td>
<td>52.08</td>
<td>52.23</td>
<td>54.30</td>
</tr>
<tr>
<td></td>
<td>Mc(FC)</td>
<td>46.67</td>
<td>41.43</td>
<td>25.12</td>
<td>14.48</td>
<td>6.81</td>
</tr>
<tr>
<td>4</td>
<td>Kc(FC)</td>
<td>77.07</td>
<td>75.06</td>
<td>66.95</td>
<td>55.66</td>
<td>53.44</td>
</tr>
<tr>
<td>5</td>
<td>Mc(FC)</td>
<td>41.60</td>
<td>25.09</td>
<td>13.80</td>
<td>5.78</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kc(FC)</td>
<td>76.78</td>
<td>73.94</td>
<td>60.06</td>
<td>53.32</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Mc(FC)</td>
<td>42.27</td>
<td>32.22</td>
<td>22.84</td>
<td>6.12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kc(FC)</td>
<td>77.53</td>
<td>76.27</td>
<td>71.27</td>
<td>54.39</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Mc(FC)</td>
<td>39.62</td>
<td>37.71</td>
<td>2.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kc(FC)</td>
<td>77.67</td>
<td>77.49</td>
<td>75.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Mc(FC)</td>
<td>42.62</td>
<td>37.87</td>
<td>30.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kc(FC)</td>
<td>81.29</td>
<td>79.14</td>
<td>77.81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Mc(FC)</td>
<td>40.62</td>
<td>40.54</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kc(FC)</td>
<td>82.10</td>
<td>80.34</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Mc(FC)</td>
<td>38.62</td>
<td>82.38</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kc(FC)</td>
<td>85.62</td>
<td>85.52</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mc(FC): model-based clustering with FC.
Mc(dif): model-based clustering with dif.
Kc(FC): K-means clustering with FC.
Kc(dif): K-means clustering with dif.

Saccharomyces cerevisiae that is a free living, eukaryotic and single cell but highly complex organism.

Spellman et al. (1998) created a comprehensive catalog of yeast genes whose transcript levels vary periodically within the cell cycle. They used DNA microarrays and samples from yeast cultures synchronized by three independent methods: a-factor arrest, elutriation and the arrest of a cdc15 temperature-sensitive mutant.

We applied our method of clustering to yeast cell cycle data downloaded from http://genome-www.stanford.edu/cellcycle/. We used yeast alpha data collected at 18 time points for 120 min during two full cell cycles. After removing genes with the missing values, there were 4489 genes remaining.

5.2.2 Choice of Fourier coefficients and clusters To determine the \( J \) value and the number of clusters, we considered several \( J \) values and Bayesian Information Criterion (BIC) with the assumption that each cluster covariance has the same elliptical volume and shape. Since we found that the optimal \( J \) value varied for each function, we surmised that a true optimal \( J \) value may not exist. As such, we experimented with the model-based clustering using various numbers of clusters and \( J \) values.

Table 2 shows the median and average silhouette values with Euclidean distance between samples by model-based and K-means clusterings for various \( J \) values in 5 clusters. Although \( J = 1 \) yields higher overall silhouette widths using both K-means and model-based clustering, we think a larger number than 1 is appropriate to extract enough information about the underlying change patterns. Judging from the highest overall silhouette value, the model-based with 4 Fourier coefficients and 5 clusters was considered most appropriate. With K-means, silhouette value with \( J = 1 \) is the largest. Silhouette value of K-means with \( J = 3 \) is larger than that of the model-based clustering with \( J = 4 \). Therefore, it should be noted that silhouette values of Euclidean distance between two clustering models may not be the only criterion for model comparison. Rather as in the following gene ontology analysis biological interpretation should be done to validate clustering. However, the model-based method including density connects the probability-neighboring data, while K-means method measures intra-cluster homogeneity as cluster compactness.
Using the model-based and \( J = 4 \), each partition of 5 clusters has the following number of genes 3032, 401, 164, 400 and 492. Figure 2 shows means, 5% and 95% of Fourier estimated gene scores in 5 clusters with sample derivative Fourier coefficients. The graph in the bottom right-hand corner of Figure 2 shows the estimated change patterns of the 5 clusters altogether. Supplementary Figure S1 shows the means of 4 derivative Fourier coefficients as a cluster profile and gives the variation between clusters. Supplementary Figure S2 shows chisquare plots of each cluster for multivariate normality with a dimension of 4. If the four derivative Fourier coefficients follow a multivariate normal distribution, this assumption can also be checked with chisquare plots. Except for cluster 4, they appear to have a slightly heavier tail than a normal distribution.

Supplementary Figure S3 and S4 show plots of \((\psi_1, \psi_2)\) and \((\psi_3, \psi_4)\) hinting at the shape of clusters using both the model-based and K-means method. The clusters are elliptical with the different size and shape combined with the probability distributions in Figure S3. The clusters are grouped closer within cluster in Figure S4. Gaussian mixture model clustering allows clusters to have different orientation or sizes while preserving some common features, such as an ellipsoidal shape. Cluster 5 in particular has a wide elliptical shape incorporated with the probability distribution.

Owing to noise and the high dimensionality of data, careful consideration of statistical and biological validity is needed when analyzing the real microarray data.

5.2.3 Gene ontology analysis In order to evaluate the result of the clustering analysis, we obtained Gene Ontology (GO) information for the clustered genes’ biological processes, molecular functions and cellular components. The GO database provides a useful tool to annotate and analyze the functions of a large number of genes. We searched statistically overrepresented GO annotations using GOstat for evaluating statistical significance of overrepresented functional and molecular mechanisms (Beissbarth and Speed, 2004). GOstat allows us to identify which annotations are typical for the group of genes. GOstat simply derives the statistical significance between expected and observed functional categories based on the Fisher’s exact test.

In order to compare our method with other clustering methods, we also applied K-means clustering (MacQueen, 1967) to yeast cell cycle data. Table 3 shows some results of the overrepresented biological processes from the proposed method and the K-means clustering method for various values of \( k \) from 5–15.

In Table 3, the first column shows the cluster number of the proposed method. The second column summarizes the list of the selected overrepresented biological processes that had their children GO terms in the same cluster. For example, we first selected total 81 GO terms in cluster 1 by using GOSTat and then selected 6 GO terms that had as many children nodes as possible in cluster 1. In the same way, K-means clustering results were obtained. We compared the list of the overrepresented GO terms from the proposed method (second column) with that from the K-means clustering method. The black dots in Table 3 represented the GO terms that were selected by both methods. In summary, there are some GO terms that can only be detected by the proposed method such as GO:0000209, GO:0000079, GO:0009086 and GO:0005978. In particular, all GO terms in cluster 5 of our proposed method are closely related to biosynthesis. The three GO terms in cluster 5, GO:0009086, GO:0006537 and GO:0005978, are rarely overrepresented by the K-means clustering method. Our proposed method not only found the GO terms that were not identified by the K-means method but also grouped them in the same cluster.

Furthermore, the genes in cluster 5 are closely related to the glucose metabolic pathway. For example, GLC3 (GO:0005978) encodes 1,4-glucan-6-(1,4-glucano)-transferase, involved in glycogen accumulation. Glycogen in turn serves as a major storage carbohydrate (glucose) (Rowen et al., 1992). Free glucose is oxidized to pyruvate. The other genes from GO:0006537, GO:0006526 and GO:0009086 are related to the synthesis of amino acids in the citric acid cycle, 15ATPs and 3CO2 are produced from one pyruvate molecule. IDP1 (GO:0006526) catalyzes the oxidation of isocitrate to alpha-ketoglutarate (Haselbeck and McAlist-Henn, 1993). GLT1 (GO:0006537) synthesizes glutamate from glutamine and alpha-ketoglutarate (Valenzuela et al., 1998). ARG1, ARG3 and ARG4 in GO:0006526 are involved in the synthesis of arginine from the glutamate (Crabeel et al., 1988; Juaiaux et al., 1978). Oxaloacetate an intermediary in the citric acid cycle, is the entry point for the metabolism of the underlying carbon structure of the amino acids aspartate and asparagine. MET2 (GO:0009086) is involved in the synthesis of methionine from the aspartate (Masselot and De Robichon-Szulmajster, 1975). It catalyzes the conversion of homoserine to O-acetyl...
homoserine using one molecule of acetyl coenzyme A (acetyl-CoA) (Thomas and Surdin-Kerjan, 1997). These findings illustrate that our proposed methodology can identify genes that are biologically interpretable.

6 CONCLUDING REMARKS

The method proposed in this study provides an efficient tool for clustering curves of the same change pattern by Fourier estimation.

Because Fourier coefficients can give information on both the original underlying functions and their derivatives, we used the sample Fourier coefficients of derivatives to summarize the change patterns. We demonstrated the effectiveness of our approach using model-based clustering of change patterns. Although we assumed that the residuals within each curve over time were independent and had constant variance, due to the large number of repetitions, we found that it is not necessary to assume independence between curves.

There are several areas that deserve further research. Determining the number of Fourier coefficients and selecting the number of clusters are topics that many researchers are actively pursuing. Also, there needs a study to handle the instability of Fourier estimation affected by outliers when only a small number of repeated time points is available. Another topic of future research is to develop validity measures incorporating the probability framework for clusters. When further information about correlations is available, a time series analysis approach would also be an area worthy of consideration.

Table 3. Result of comparison between overrepresented biological processes by using proposed method and K-means method

<table>
<thead>
<tr>
<th>Numbers</th>
<th>GO terms in biological process of the proposed method</th>
<th>S C</th>
<th>Number of centers (k) for the K-means clustering method</th>
<th>5 6 7 8 9 10 11 12 13 14 15</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GO:0006511,ubiquitin-dependent protein catabolism</td>
<td>26 27</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0006445,regulation of translation</td>
<td>20 27</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0006418,9RNA aminoacylation for protein translation</td>
<td>19 29</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0030150,protein import into mitochondrial matrix</td>
<td>17 19</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0006413,translational initiation</td>
<td>16 17</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0000209,protein polyubiquitination</td>
<td>16 17</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>GO:0006365,35S primary transcript processing</td>
<td>19 20</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0030490,processing of 20S pre-tRNA</td>
<td>19 20</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0000027,ribosomal large subunit assembly and maintenance</td>
<td>14 16</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0000055,ribosomal large subunit export from nucleus</td>
<td>14 26</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0030488,tRNA methylation</td>
<td>14 18</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0000154,tRNA modification</td>
<td>13 15</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0006325,establishment and/or maintenance of chromatin arch.</td>
<td>13 17</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>GO:0000750,pheromone-dependent signal transduction during conj.</td>
<td>21 28</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0015892,siderophore-iron transport</td>
<td>11 15</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0006827,high affinity iron ion transport</td>
<td>10 14</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0000079,regulation of cyclin dependent protein kinase activity</td>
<td>9 16</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>GO:0000710,meiotic mismatch repair</td>
<td>14 30</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0007064,mitotic sister chromatid cohesion</td>
<td>10 14</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0000731,DNA synthesis during DNA repair</td>
<td>8 18</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0006301,postreplication repair</td>
<td>7 17</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0006284,base-excision repair</td>
<td>7 17</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0006289,nucleotide-excision repair</td>
<td>7 17</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0006269,DNA replication, synthesis of RNA primer</td>
<td>6 16</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>GO:0009086,methionine biosynthesis</td>
<td>15 26</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0005626,arginine biosynthesis</td>
<td>14 23</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0006537,glutamate biosynthesis</td>
<td>13 22</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GO:0005978,glycogen biosynthesis</td>
<td>9 27</td>
<td>* * * * * * * * * * * * * * * * * *</td>
<td></td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

We are grateful to Dr Carroll, Dr Hart and Dr Vannucci for their motivation and advice. We also thank the referees for their constructive comments. This work took place during J.K.’s visit to the Bioinformatics Training Program at the Department of Statistics at Texas A&M University and her research was supported by the Korea Research Foundation (R04-2004-000-10138-0). The work of H.K. was supported by the National Research Laboratory Program of Korea Science and Engineering Foundation (M10500000126).

Conflict of Interest: none declared.

REFERENCES