International Scholarly Research Network
ISRN Bioinformatics
Volume 2012, Article ID 537217, 12 pages
doi:10.5402/2012/537217

Research Article

Dynamic Clustering of Gene Expression

Lingling An[1] and R. W. Doerge[2]

[1] *Department of Agricultural and Biosystems Engineering, University of Arizona, Tucson, AZ 85721, USA*
[2] *Department of Statistics, Purdue University, West Lafayette, IN 47907, USA*

Correspondence should be addressed to R. W. Doerge, doerge@purdue.edu

Received 11 July 2012; Accepted 5 August 2012

Academic Editors: T. Can, Z. Gáspári, and A. Pulvirenti

It is well accepted that genes are simultaneously involved in multiple biological processes and that genes are coordinated over the duration of such events. Unfortunately, clustering methodologies that group genes for the purpose of novel gene discovery fail to acknowledge the dynamic nature of biological processes and provide static clusters, even when the expression of genes is assessed across time or developmental stages. By taking advantage of techniques and theories from time frequency analysis, periodic gene expression profiles are dynamically clustered based on the assumption that different spectral frequencies characterize different biological processes. A two-step cluster validation approach is proposed to statistically estimate both the optimal number of clusters and to distinguish significant clusters from noise. The resulting clusters reveal coordinated coexpressed genes. This novel dynamic clustering approach has broad applicability to a vast range of sequential data scenarios where the order of the series is of interest.

1. Introduction

Microarray and next-generation sequencing (RNA-seq) technologies enable researchers to study any genomewide transcriptome at coordinated and varying stages. Since biological processes are time varying [1], they may be best described by time series gene expression rather than by a static gene expression analysis. Acknowledging the nature of genes that are involved in dynamic biological processes (e.g., developmental processes, mechanisms of cell cycle regulation, etc.) has potential to provide insight into the complex associations between genes that are involved.

Functional discovery is a common goal of clustering gene expression data. In fact, the functionality of genes can be inferred if their expression patterns, or profiles, are similar to genes of known function. There are published clustering methods that include into the analysis the duration of the experimental stages, or the staged dependence structure of gene expression. The results from these approaches are certainly more informative and realistic than groupings that are gained from static clustering methods (i.e., clustering at a single-staged experimental point), but their results are limited in interpretation. The seminal work from Luan and Li [2] is a good example of a clustering application that takes the time dependent nature of genes into account. More realistic, though, is the fact that some biological processes typically start and end at identifiable stages, or time points, and that the genes in a process may be dynamically regulated at different stages of the biological process [3]. In other words, genes can be coregulated over a finite series of points (i.e., only a portion of points represent the total when the transcriptome is being sampled).

A variety of subspace clustering methodologies have attempted to address the time-dependent nature of transcriptome experiments through biclustering [4], or plaid models [5]. Although these bicluster (i.e., clusters obtained by any subspace clustering method are referred to as biclusters from this point forward) approaches are popular, they have limitations. Namely, they restrict subspace clusters to consecutive time points [6–9]. For example, Madeira and Oliveira [8] discretized real-valued gene expression data as upregulated, downregulated, and unchanged according to the slope of expression change from one time point to the next. They then rely on string processing techniques to develop an algorithm that identifies contiguous column coherent biclusters. Alternatively, Zhang et al. [9] alter original expression data by deleting and inserting border time points, and then use an algorithm based on a mean squared residue score to cluster the modified expression data.

We are motivated by the fact that the genes involved in the biclusters that are obtained by [8, 9] have the same starting and ending time point(s). Even though it is well known that time lags exist for many genes that are involved in the same biological process and that genes with the same function may give rise to unique expression patterns/profiles, to our knowledge this information has not been incorporated into any statistical approach for clustering. Ji and Tan [6] focus on extracting time-lagged gene clusters known as q-clusters, where q is the time length of a bicluster (i.e., the number of consecutive time points in the bicluster), that can have different time lengths, but genes in the same cluster must have the same durations over time, even though time lags exist among the genes. Song et al. [7] proposed to use a wavelet-based cluster method to detect time shift/delay situation. To our knowledge, none of the current or existing subspace clustering methodologies is able to provide biclusters that are varying in their duration of time length.

We know that standard exploratory clustering methods are useful for grouping items that behave in a similar fashion. However, when these standard approaches are applied to experiments that evaluate the transcriptome over coordinated experimental stages, they fail to acknowledge the dynamic nature of such processes. As such, this work focuses on the dynamic and nonconsistent nature of gene activity (Figure 1). Although presented here in the context of coordinated transcriptome data, this novel dynamic clustering approach is applicable to a vast range of sequential data scenarios where the order of the series is of interest [10].

2. Methods

One important aspect of periodic (gene expression) profiles is spectral frequency. Specifically, there may be discontinuities (e.g., a gene may enter or exit a biological process at any time point), or more generally, a time-varying spectral frequency in nonstationary or piecewise stationary time series (i.e., signal) that can be studied using techniques and theories from signal decomposition [11]. “Time series” and “signal” are used exchangeable from this point forward. Once decomposed, the components can be quantified based on their constrained coherency (CoCo, details in the Section 2.2) and gathered via an agglomerative hierarchical clustering that involves determining the number of clusters and separating signal from noise.

2.1. Time Frequency Analysis. For a nonstationary time series, time-frequency analysis is usually employed to map one-dimensional time series onto a two-dimensional time-frequency domain so that both the frequency and time information can be considered [12, 13]. Addison [14] justifies the continuous wavelet transformation (CWT) based time-frequency analysis and states that it has many benefits over other time-frequency representations of multiple-component signals (i.e., signals that contain multiple frequencies) or signals with discontinuities in frequency [13]. We assume that even though a gene may be expressed multiple times in a time series, it may also be involved in multiple biological processes, and therefore multiple spectral components need to be considered [15]. Toward this end, we employ CWT to decompose gene expression signals, $s(t)$:

$$W(b,a) = \frac{1}{a}\int s(t)\varphi\left(\frac{t-b}{a}\right)dt, \tag{1}$$

where $\varphi(t)$ is the mother wavelet, a is the scale parameter, b is the shift parameter, and the Morlet wavelet [16] is employed. The significant frequency values and their associated starting and ending time points for an expression signal, $s(t) = (s_1, s_2, \ldots, s_N)$, are determined as follows.

(1) Perform a continuous wavelet transformation (CWT) on the time series $s(t)$.

(2) Generate a white noise time series $w(t) = (w_1,w_2,\ldots, w_N)$, where w_i's are distributed as $N(0,\sigma^2)$, where σ^2 is the variance of the time series $s(t)$.

(3) Perform a CWT on the white noise time series from Step 2 and obtain the time-frequency representation.

(4) Repeat Steps 2-3 M times (usually, $M \geq 1000$). For each frequency f at each time point t, the 95th percentile of M modulus serves as the threshold. The threshold surface is constructed by connecting all thresholds across both frequency and time.

Let $W_0(t, f)$ represent the threshold surface calculated from white noise information, and let $W(t, f)$ represent the CWT outcome of the original gene expression signal $s(t)$. Determining the significant components whose CWT values are above the threshold surface is equivalent to identifying the components from W':

$$W'(t, f) = \max\{0, W(t, f) - W_0(t, f)\}, \tag{2}$$

where t represents time and f denotes frequency. It is worth noting that the meaningful component on the sliced CWT has some width in the frequency domain, and that a signal may contain several significant components. A “Crazy Climber” ridge extraction algorithm is employed [12] to extract significant frequencies and their related starting and ending time points.

2.2. Similarity Measurement. The actual clustering (of genes) requires a similarity metric that measures the pairwise relationship between genes via their decomposed components. Since frequency is of interest, a coherency [17] is used. For two signals x and y, coherency is

$$C_{xy}(f) = \frac{P^2_{xy}(f)}{P_{xx}(f)P_{yy}(f)}, \tag{3}$$

where $P_{xx}(f)$ and $P_{yy}(f)$ are the respective power spectral densities of the two signals and $P_{xy}(f)$ is their cross-spectral

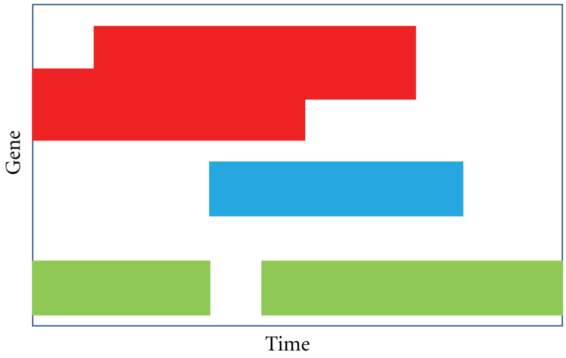

Figure 1: Three (red, blue, green) hypothetical gene clusters across time. The specific genes in the red cluster is varying, and of different duration, while the genes in the green cluster reenter the cluster after a time of not being in the cluster, and all of the genes in the blue cluster enter and leave the process at the same time.

density. Typically when calculating the coherency function between two signals, the lengths of the signals are the same. However, since the component signals may differ in length, the median length for all frequencies is used to represent a uniform time length. Further, since coherency is a function of frequency, its range is from zero to half of the sampling frequency of a discrete signal (i.e., a continuous signal measured at discrete time points; [18]). In general, the sampling frequency, or sampling rate, is the number of time points per second as measured in Hertz. Given two signals with frequencies f_x and f_y, we use the average of the coherency function in the interval $[f_x, f_y]$ to represent the coherency similarity and refer to it as "coherency" for simplicity.

Although it is expected that coherency will decrease as the difference between two frequencies (from two signals) increases, there are situations (Figure 2) when a nonmonotonic coherency pattern identifiable from a coherency plot exists. To ensure monotonicity, we provide a modified coherency that acknowledges that the valleys of the coherency curves are nonincreasing along the frequency differences when we construct a representative curve that is monotonic in frequency difference. Although many other similarity measures can be employed (e.g., Brownian Distance measurement [19]), our modification (CoCo) is data-driven and provides a constrained coherency that serves as a similarity measurement for clustering (Figure 2).

2.3. Clustering. We rely on agglomerative hierarchical clustering coupled with Ward linkage [21] for clustering. Although the Ward linkage finds compact and homogenous clusters by minimizing the variance of objects, there are two issues that we need to address: determining the number of clusters and differentiating meaningful clusters from noise clusters (i.e., clusters containing only noise).

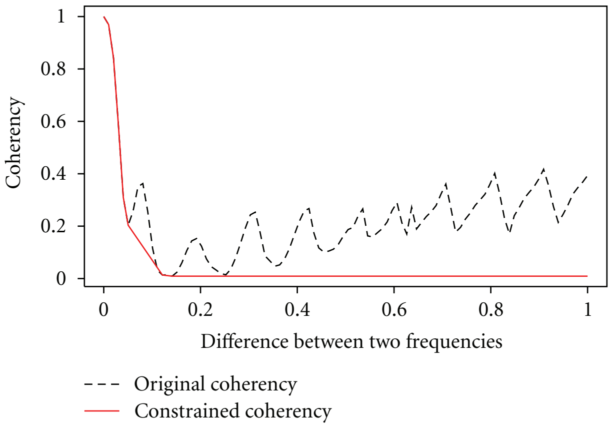

Figure 2: The nonmonotonic pattern of frequency differences between two signals requires a modification to the coherency. A constrained coherency (CoCo) identifies the valleys of the original curve as nonincreasing along the frequency differences and provides a monotonic representation of coherency. Illustration taken from An and Doerge [20].

2.4. Cluster Validation

2.4.1. Number of Clusters. A variety of approaches have been suggested for determining the number of clusters [22–24]. One well-known approach involves finding the "elbow" (or, change point) of an error curve [25]. Unfortunately, most work that relies on the elbow lacks statistical justification [26]. Although others have attempted to compare the error curve of the original data to the error curve of the data generated from a null reference distribution (i.e., uniform distribution) by employing the Gap statistic [27], this approach is not applicable to overlapping clusters, nor is it appropriate for noisy data [24, 25].

We approach the issue of determining the number of clusters by globally evaluating the merge distance plot, which is represented by the height of joint nodes in a cluster tree. A null reference distribution is able to provide a merge-distance threshold that can be compared to the original merge distance. Specifically, a minimal convex set, or a convex hull, is formed based on a set of convex combinations of all points of interest [28]. The merge distance is obtained by performing clustering on data generated from a uniform distribution from the convex hull. Munneke et al. [29] used a convex hull to generate randomly distributed gene expression values so that distinctions between clusters arise with statistical confidence. Here, a convex hull is employed to assess the statistical significance of the merge distance of hierarchical clusters. By evaluating the merge distance globally, the number of clusters can be determined.

For a set X containing n objects, $\{x_1, \ldots, x_n\}$, the convex hull of X is

$$H(X) = \left\{ \sum_{i=1}^{n} \alpha_i x_i \mid x_i \in X, \alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i = 1 \right\}, \qquad (4)$$

where x_i can be one or more dimensions. For time series gene expression data, n is the total number of decomposed components and the x_i's are the component frequencies. The following steps are used to determine the number of clusters.

(1) Perform hierarchical clustering with Ward linkage and CoCo similarity on the original data (x_i) containing n points and obtain the merge distance set $M_0 = (d_2, \ldots, d_n)$.

(2) Randomly choose n objects from the uniform distribution $[\min(x_i), \max(x_i)]$.

(3) Perform hierarchical clustering on the random data set from Step (2), and obtain a new merge distance set.

(4) Repeat Steps (2)-(3) M times (usually $M \geq 1000$). For each possible number of clusters ($k, k = 2, \ldots, n$) the 95th percentile of M merge distances, d_k^*, serves as a threshold and the 95% threshold curve is constructed as $M^* = (d_2^*, \ldots, d_n^*)$.

(5) Compare M_0 with M^*; the largest k which satisfies $d_k > d_k^*$ is the optimal number of clusters k_0.

Since some gene expression data can be quite noisy, an additional step in cluster validation necessitates differentiating the noise cluster from meaningful (i.e., significant) clusters.

2.4.2. Significant Clusters. A noise cluster differs from statistically significant clusters in terms of compactness and separation. The objects in a noise cluster are scattered, while the objects in a statistically significant cluster, which are similar to the tight clusters in [30], are dense. Here, the silhouette metric [31], a measure of tightness and separation of clusters, is used to assess the level of statistical significance of clusters. For each object i, its silhouette width is defined as

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{5}$$

where $a(i)$ is the average dissimilarity of object i to other objects in the same cluster and $b(i)$ is the average dissimilarity of object i to objects in its nearest neighbor cluster. The range of the silhouette width is $[-1, 1]$. The average silhouette width of objects in a cluster represents the quality of the cluster in terms of compactness and separation. The average silhouette of a noise cluster (if it exists) should be low, while for statistically significant clusters it should be high.

When evaluating the silhouettes for k_0 clusters, a uniform reference distribution is employed to generate independently located objects upon which the hierarchical clustering operates. The silhouettes are obtained for the reference data for the same cluster number k_0 as follows [20].

(1) For k_0 clusters from the original data, compute their silhouettes.

(2) Randomly choose n objects from the uniform distribution on the convex hull of the frequencies of the original data.

(3) Perform hierarchical clustering on data from Step (2), choose k_0 clusters and obtain their silhouettes.

(4) Repeat Steps (2)-(3) M times (usually $M \geq 1000$) and obtain M sets of k_0 silhouettes. For each of k_0 silhouettes of the original data, calculate its P value from $M^* k_0$ values. Cluster significance is represented by its P-value.

(5) The significance level is α. A cluster is significant if its P value $< \alpha$; otherwise, it is noise.

2.5. Dynamic Cluster. Up to this point, we have described a two-step cluster validation that provides the number of clusters and differentiates significant clusters from a noise cluster. Since a gene's expression over time can be described in terms of frequencies and then decomposed into components that each may have unique start and stop points, the time-dependent structure of the data is retained, and the concept of a dynamic cluster evolves naturally.

In anticipation of assessing the performance of the proposed dynamic clustering algorithm via simulation, we realize two further issues. First, to our knowledge, there are no clustering approaches, which provide gene sets at different (time) points. This makes comparing our approach with existing approaches fruitless. Second, because we work from simulated data, we need a metric to compare the clusters that result from dynamic clustering with the cluster-scenario from which they were simulated. To address these issues, we develop a discovery index.

2.6. Discovery Index. The proposed cluster validation algorithm objectively evaluates the quality of clustering results using information from the data (i.e., silhouette width). When prior information (e.g., pathway, etc.) is available, obviously it contributes even more information upon which to base cluster validation. Fortunately, this is exactly the situation for time series gene expression with time-varying frequencies; information on both true cluster membership and true time duration for genes in clusters is available. Since traditional criteria [32, 33] only involve the true cluster membership, we provide a new criterion that takes into account both time information and cluster characteristics.

Assume an estimated cluster E obtained via a clustering method represents the true class C such that some genes in E differ from genes in class C in terms of the gene identification number or the gene time information. For each gene in either true class C, or estimated cluster E, consider the true time interval and estimated time interval. The discovery index for gene g in class j or cluster j is defined as

$$D_{gj} = \frac{P_{gj} \cap Q_{gj}}{P_{gj} \cup Q_{gj}}, \tag{6}$$

where P_{gj} denotes the true time interval and Q_{gj} denotes the estimated time interval. By this definition, $0 \leq D_{gj} \leq 1$. Specifically, when gene g appears in the class but is not detected in the corresponding cluster we have $D_{gj} = 0$ since $Q_{gj} = 0$; similarly, when it is detected in the cluster but is not in the corresponding class we get $D_{gj} = 0$ due to $P_{gj} = 0$. For a class it is possible to define its cluster by using all pairwise comparisons and a specified (small) cut-off. We can find the closest cluster (i.e., estimate) for each class. Specifically, if the distance between a class and its closest cluster (i.e., the difference between two frequency values) is less than the predefined cutoff, then the cluster will be called its estimate. At times not all of the classes will be detected, or some cluster may be superfluous since it may not have a matching class. In these situations, the discovery indices for the genes in either that class or that cluster are all zero. Finally, the overall discovery index for the genes across the corresponding clusters and classes is

$$D = \frac{\sum_{j=1}^{J} \sum_{g=1}^{G} D_{gj}}{\sum_{j=1}^{J} O_j}, \qquad (7)$$

where G the number of genes, J is the maximum of the number of clusters and the number of classes, O_j is the number of genes in the jth cluster or class, and D_{gj} is the discovery index for gene g in the jth cluster or class. Clearly, $\sum_{g=1}^{G} D_{gj} \le O_j$, so $0 \le D \le 1$. The discovery index evaluates the value of combining clustering and signal decomposition.

3. Simulation Studies

Because the proposed approach includes two subprocedures, namely, time-frequency decomposition and clustering, and because the result of the first subprocedure impacts the results of the second subprocedure, a power study will be conducted on these two subprocedures separately. The power of the signal decomposition is investigated relative to the noise level, frequency level, difference between frequencies of components in a gene signal, and time lengths of components. Interestingly, gene expression contains various features, for example, some genes may have one component with a certain frequency; some genes may have two components with other frequencies; others may have more than two components with different frequencies. We acknowledge that investigating the performance of the signal decomposition, and the proposed clustering method, for these types of data is challenging simply because it is difficult to summarize the distribution of the frequencies for multiple components from multiple genes. Furthermore, the time durations of the components affect statistical power.

3.1. Time-Frequency Analysis. Signal decomposition is influenced by many parameters/factors, including frequency, frequency difference (for multiple component signals), time length of signal, ratio of the amplitude of the components, and noise level. Because of the limitation of displaying multiple factor effects simultaneously, a power study of signal decomposition on two-component signals is performed for investigating the effect of frequency, frequency difference, amplitude ratio, and noise. Additional simulation studies can be found in An's paper [34]. In general, the two-component signals are simulated from

$$\begin{aligned} s(t) = {} & \cos(2\pi f t + \varphi_1) + A\cos(2\pi(f + \Delta f)t + \varphi_2) \\ & + \text{noise}_{\text{level}} * N(0,1), \end{aligned} \qquad (8)$$

where t is time from 0 to 10 seconds, f represents spectral frequency that varies from 0.1 Hz to 1.0 Hz, Δf denotes the frequency difference between two components (from 0.1 Hz to 1.0 Hz), A represents the amplitude ratio between two components (0.5, 0.2, 1.0, 2.0, 5.0; the amplitude of the first component is 1, as a baseline), and φ_1 and φ_2 are the phase shifts that are randomly chosen from $[0, 2\pi]$. The noise level varies from 0 to 1.0, in increments of 0.10. In spectral analysis, particularly in fast Fourier transformations, the length of a signal is usually a power of two. We consider 64 time points that equally partition the sample space.

The power of decomposing each gene expression time series is defined as

$$\frac{\sum_{k=1}^{c} w_k}{\sum_{i=1}^{q} u_i + \sum_{j=1}^{m} v_j - \sum_{k=1}^{c} w_k}, \qquad (9)$$

where $\sum_{i=1}^{q} u_i$ is the total time duration of the q true components ($q = 2$ for the two-component signal decomposition), $\sum_{j=1}^{m} v_j$ is the total time duration of the m estimated components, and $\sum_{k=1}^{c} w_k$ is the total time duration of the $c$overlap components between the true and estimated components. For each parameter combination, the overall power of the signal decomposition is the average power for decomposing 1000 signals.

Figure 3 illustrates the power study of decomposing two-component signals with amplitude ratio 1.0 without noise. A larger difference between two frequencies results in greater power, and a lower frequency is more likely to be detected than a higher frequency. Similar conclusions are obtained for decomposing signals with different amplitude ratios and different noise levels (plots not shown). The effect of time duration is also investigated. Components with longer time duration are more likely to be detected than those having short time duration (plots not shown).

3.2. Dynamic Clustering. The performance of the dynamic clustering approach coupled with the proposed validation method is investigated using the discovery index, which reflects the effect of both the signal decomposition and the clustering. It is a considerable challenge to display the discovery index for a set of data across all possible combinations of parameters. Fortunately, it is possible to assess the effect due to noise in the discovery index while holding the other parameter settings fixed. The performance of the dynamic clustering is assessed via fixed parameters over increasing noise. Further, the effect of time, when other parameter settings are fixed, is also assessed. Relying on the previous simulation, we use 140 simulated genes to illustrate the dynamic nature of the simulated time series.

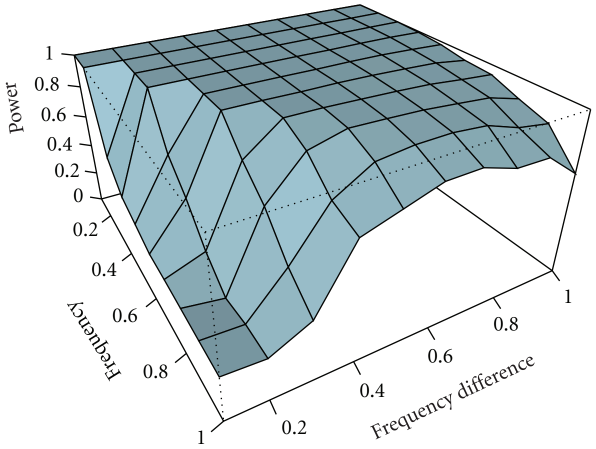

Figure 3: The decomposition power of CWT for two-component signals without noise. The two components, both with the same amplitude (i.e., ratio 1), have different frequencies.

3.2.1. Power Study: Noise Effect. Time series expression data for 140 genes in three groups are simulated as follows:

$$\begin{aligned} \text{group } 1 &= \cos(2\pi * 0.1t + \varphi_1) + \cos(2\pi * 0.8t + \varphi_2) \\ &\quad + \text{noise}_{\text{level}} * N(0,1), \\ \text{group } 2 &= \cos(2\pi * 0.4t + \varphi_3) + \cos(2\pi * 0.8t + \varphi_4) \\ &\quad + \text{noise}_{\text{level}} * N(0,1), \\ \text{group } 3 &= \cos(2\pi * 0.1t + \varphi_5) + \cos(2\pi * 0.4t + \varphi_6) \\ &\quad + \text{noise}_{\text{level}} * N(0,1). \end{aligned} \tag{10}$$

The time t and the phase shifts φ_j $(j = 1, \dots, 6)$ are simulated as in the previous power study. There are 20, 40, and 80 genes in groups 1, 2, and 3, respectively. Each gene expression profile contains two components whose frequencies are 0.1 Hz, 0.4 Hz, and 0.8 Hz. Each pair of genes from different groups shares one common component. Sixty-four time points equally partition the sample space $[0, 10]$. Noise level varies at 0.5, 1, and 1.5.

The discovery index is calculated for each scenario, and the average discovery index-calculated for each of the 1000 simulated data sets (Table 1). Figure 4 illustrates single gene expression profiles with two components, 0.1 and 0.4 Hz, at three different noise levels (0.5, 1, and 1.5). Interestingly, the discovery index is very high even though the data are moderately noisy (i.e., noise level 1.0 and the energy ratio from signal and noise is 1 : 1). Dynamic clustering using the proposed similarity metric and validation methods is able to capture meaningful information from relatively noisy data.

3.2.2. Dynamic Clustering for Various Time Lengths. The components in the previous simulation are simulated (3) under varying times (0 to 10), across the entire time interval. Since some genes may belong to a cluster in only a portion of a time interval (simply because of noise), dynamic clusters are demonstrated using simulated data where time intervals for some components are only a portion of entire time interval. One hundred and forty genes are simulated as follows:

Table 1: Dynamic indices for simulated data (1000 datasets) at three different noise levels.

	Noise level		
	0.5	1.0	1.5
Average of dynamic index	0.999	0.910	0.653
St. dev. of dynamic index	$8e-4$	0.012	0.024

$$\begin{aligned} \text{group } 1 &= \cos(2\pi * 0.1t + \varphi_1) + N(0, 0.5), \quad 0 \le t < 5 \\ &= \cos(2\pi * 0.1t + \varphi_2) + \cos(2\pi * 0.8t + \varphi_3) \\ &\quad + N(0, 0.5), \quad 5 \le t \le 10, \\ \text{group } 2 &= \cos(2\pi * 0.4t + \varphi_4) + \cos(2\pi * 0.8t + \varphi_5) \\ &\quad + N(0, 0.5), \quad 0 \le t < 5 \\ &= \cos(2\pi * 0.4t + \varphi_6) + N(0, 0.5), \quad 5 \le t \le 10, \\ \text{group } 3 &= \cos(2\pi * 0.1t + \varphi_7) + N(0, 0.5), \quad 0 \le t < 5 \\ &= \cos(2\pi * 0.1t + \varphi_8) \\ &\quad + \cos(2\pi * 0.4t + \varphi_9) + N(0, 0.5), \quad 5 \le t \le 10. \end{aligned} \tag{11}$$

The true clusters (i.e., the clusters from which the 140 genes are simulated) are illustrated in Figure 5 where different colors represent different clusters. In each panel, the segmentation represents the starting and ending time points of the corresponding genes involved in that cluster. The cluster characteristics are represented by the component frequencies, 0.1, 0.4, and 0.8 Hz, and are listed on the top of each panel.

Three significant clusters are detected by cluster validation. The dynamic property of the clusters is displayed in Figure 6 and is visibly comparable to Figure 5 (simulation setting). In the (red) cluster with frequency 0.1 Hz, all genes from group 1 and group 3 are involved during the whole time interval. For the (green) cluster with frequency 0.4 Hz, all genes from group 2 appear in the entire interval. In this cluster, some genes from group 3 remain in the second half of the time interval and some genes are in the entire time interval. For the third (blue) cluster with frequency of 0.83 Hz, most genes in group 1 are active in the second half interval and most genes in group 2 are active in the first half interval. Figure 6 reveals that components with a lower frequency, and of long duration, are unlikely to be affected by noise during the decomposition. Interestingly, the edge effect of time series is involved in the results (Figure 6). Specifically, the starting or ending time points for the genes in the third (blue) cluster may not be accurately estimated, yet they can be estimated precisely for the genes of a long duration (e.g., the genes in the first (red) cluster and genes from the second group in the second (green) cluster).

Although we relied on the 140 genes from our earlier simulation to demonstrate the performance of the proposed

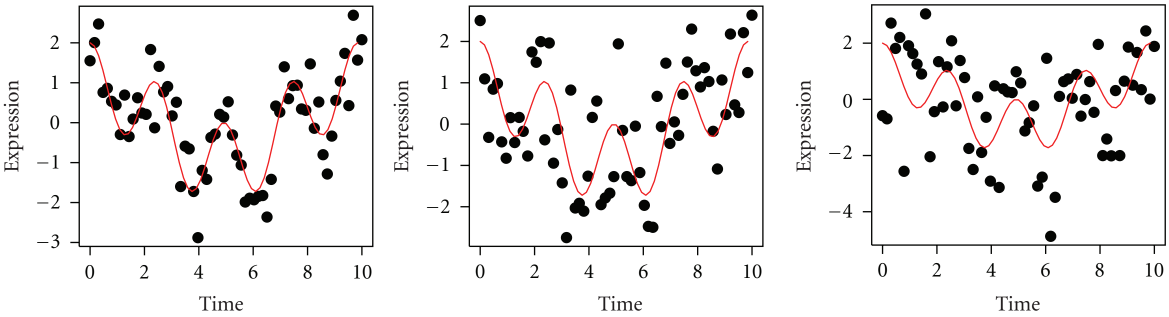

Figure 4: A single simulated gene expression profile with components 0.1 Hz and 0.4 Hz at three different noise levels: 0.5, 1, and 1.5 (from left to right). The red line represents each profile without noise.

method, it is worth noting that performance improves as the number of genes increases. Specifically, increasing gene number allows our algorithm to more accurately identify the noise cluster, thus separating the gene cluster(s) of interest more precisely. While our simulation studies demonstrate that our approach is able to both capture meaningful signals from very noisy data and group them very well, we cannot compare our method with existing methods [36] simply because no information about time is contained in the clusters obtained by other methods.

4. Real Data Application

Many microarray experiments have been conducted for the purpose of understanding complex dynamic biological processes and gene function of cell cycle (e.g., yeast, human fibroblasts, human cancer cell lines, and *Plasmodium falciparum*) [37–40]. Applicable here is the fact that it is essential cell-cycle genes exhibit periodic expression over time. Dynamic clustering is applied to cell-cycle *Plasmodium falciparum* (known to cause malaria in humans) expression data from [35].

4.1. Plasmodium falciparum. Between the mosquito vector and human host, *Plasmodium falciparum* has a complex life cycle. Its genome is sequenced and has over 5,000 cell-cycle genes; 530 of them are annotated into 14 functional groups [35]. A more complete understanding of the life cycle and gene regulation will provide the foundation for drug and vaccine development, for example, shortening its life cycle may control transmission. Most of studies on the *P. falciparum* data focus on either detecting periodic genes [41, 42] or static clustering [43, 44]. Dynamic clustering is applied to the 530 gene representations measured at 46 time points spanning 48 hours during the Intraerythrocytic Developmental Cycle (IDC) with 1 hour time resolution for the HB3 strain. The data are downloaded from http://dx.doi.org/10.1371/journal.pbio.0000005.st002. The missing data (for time point 23 and 29) are imputed by the k-nearest neighbor algorithm [45] with $k = 12$.

4.2. Results. Using the continuous wavelet transformation and ridge extraction, 530 time series gene expression profiles are decomposed into a set of 1,019 component signals whose frequencies are centralized for the purpose of calculation. Hierarchical clustering using CoCo similarity is employed. Two significant clusters and one noise cluster are detected (Figure 7). The number of genes in each (and between) significant cluster is summarized in Figure 8.

In the signal decomposition, the phase or phase shift can be obtained for each component. Figure 9 summarizes genes that are ordered by the phase (shift) of the corresponding component. Red represents high values and green denotes low values. The cluster characteristic is represented by period (reciprocal of frequency). 444 genes are in the cluster with period of 31.9 hours and 528 genes are in the cluster with period of 63.8 hours. The periods of 31.9 hours and 63.8 hours are equivalent to1.5 cycles and 0.75 cycles, respectively. These are consistent with the findings in the original research [35]. Specifically, that the majority of the periodic gene profiles exhibit an overall expression period of 0.75∼1.5 cycles in 48 hour interval.

Since the components contain time information (i.e., starting and ending time points), the number of genes in a cluster may vary across time. The dynamic property of clusters from the *P. falciparum* data can be summarized in terms of number of genes at each time point (Figure 10). A nonparametric bootstrap [46] is employed to calculate the 95% confidence interval for the statistic. Hence the 95% confidence band for the curve of the number of genes is constructed by calculating the confidence intervals across all time points.

As illustrated in Figure 10, the number of genes in Cluster 2 varies with time while the number of genes in Cluster 1 (i.e., the ones with longer period) remains constant. This confirms the findings from the simulation study that genes with lower spectral frequency are more likely to be detected than those with higher frequency. Apparently, multiple periods in *P. falciparum* data have not been studied (i.e., no published results); therefore, the expression pattern in Figure 10 may warrant further investigation.

4.3. Gene Ontology Analysis of the Clustered Genes. We employed the web-accessible programs DAVID (Database for Annotation, Visualization and Integrated Discovery [47, 48]) and PlasmoDB (Plasmodium Genomics Resource [49])

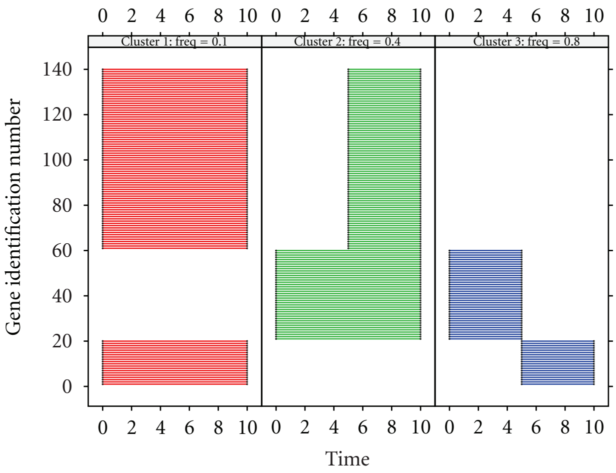

FIGURE 5: Panel plot for the true three clusters of 140 genes simulated using (8). Each segment represents a component (gene id on the y-axis) with the starting and ending time points detected via signal decomposition. The frequency value in each cluster is given at the top of each panel.

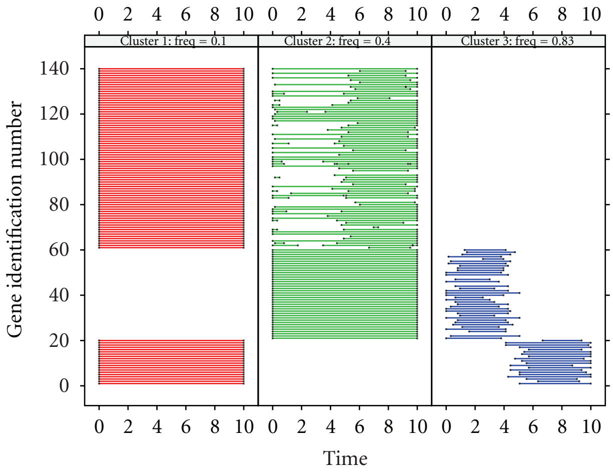

FIGURE 6: Panel plot for the estimated three clusters from 140 genes simulated using (8). Each segment represents a component (gene id on the y-axis) with the starting and ending time points detected via signal decomposition. The estimated frequency value for each cluster is shown at the top of each panel.

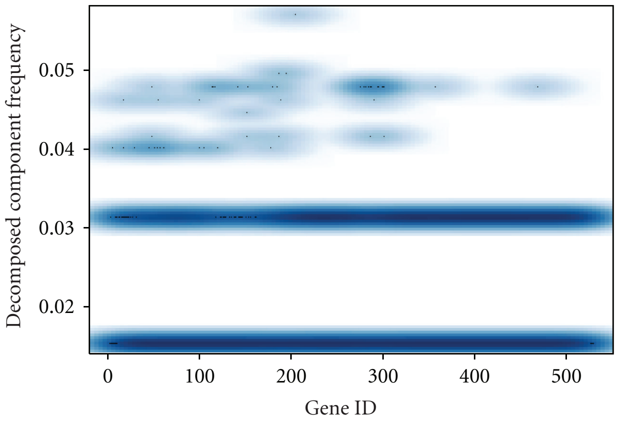

FIGURE 7: Smoothed scatter plot of decomposed frequency (centralized) of 1,019 components for 530 *P. falciparum* genes [35]. The dark blue area represents the high density of the points.

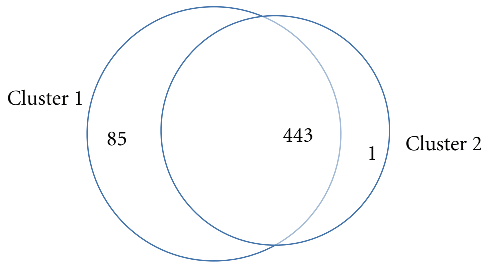

FIGURE 8: Distribution of the gene counts in the two significant clusters that result from an application of dynamic clustering to *P. falciparum* data from Bozdech et al. [35].

for the gene ontology (GO) analysis. In DAVID, the GO-BP FAT term is used to report enrichment results, as it "attempts to filter the broadest terms so that they do not overshadow the more specific terms" [50]. The biological process "translation" is the only GO term that is detected as enriched (with FDR = $8.9e - 19$) for the genes appearing in Cluster 1. The genes involved in both clusters with significant GO terms are listed in Table 2.

There are a quite few genes involved in multiple (independent) processes. For example, we find that the processes "amino acid activation" and "tRNA metabolic process" share 19 genes in common, while these two processes have no ancestor—child relationship. There is only one gene in Cluster 2 that is not in Cluster 1. This gene is involved in the "glycolysis." As a point of future research, we noticed that the phase information is often used to cluster cell cycle gene expression profiles [35, 44, 51]. As such, a nature evolution of our approach is to employ phase information in clustering (see Section 5).

5. Discussion and Conclusion

When the application is gene expression, methods from signal processing have proven successful in decomposing the nature of complicated time series that contain multiple component signals. A two-step cluster validation is proposed to statistically determine the optimal number of clusters and to select the statistically significant clusters. To our knowledge, there are no clustering approaches that provide unique gene sets at different time points (i.e., genes in the same cluster may have different starting and ending points, and even have different time durations in the cluster). A simulation study demonstrates the benefits of our approach by showing that it is able to capture meaningful signals and separate them, even for very noisy data. Finally, we understand and acknowledge that it would be useful and encouraging to compare our method with other existing methods. Unfortunately, this is not possible simply because

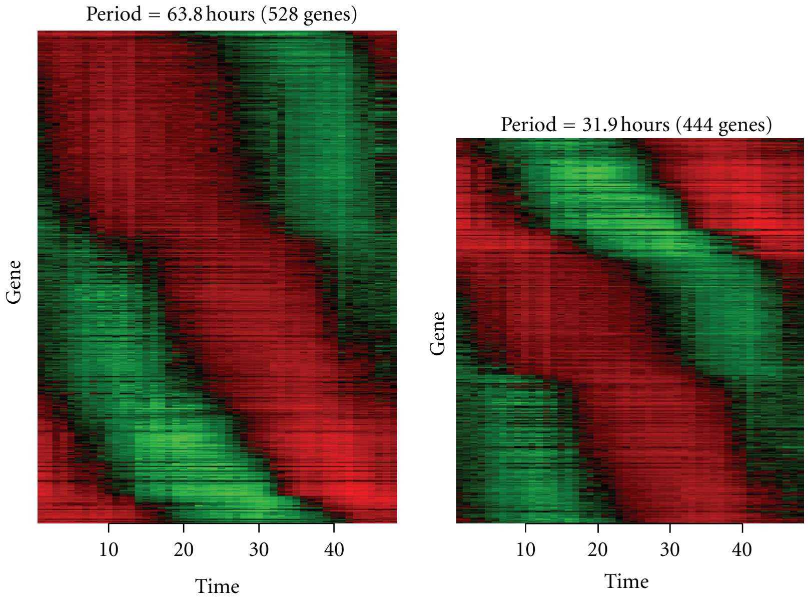

Figure 9: Expression profiles for the 530 *P. falciparum* genes [35] in the two different clusters that are found by dynamic clustering. The number of genes and their average period is at the top of each plot. The x-axis represents time (from 1 to 48 hours) and the y-axis represents genes that are ordered by their phases (from the signal decomposition). Red represents high expression values and green represents low expression values.

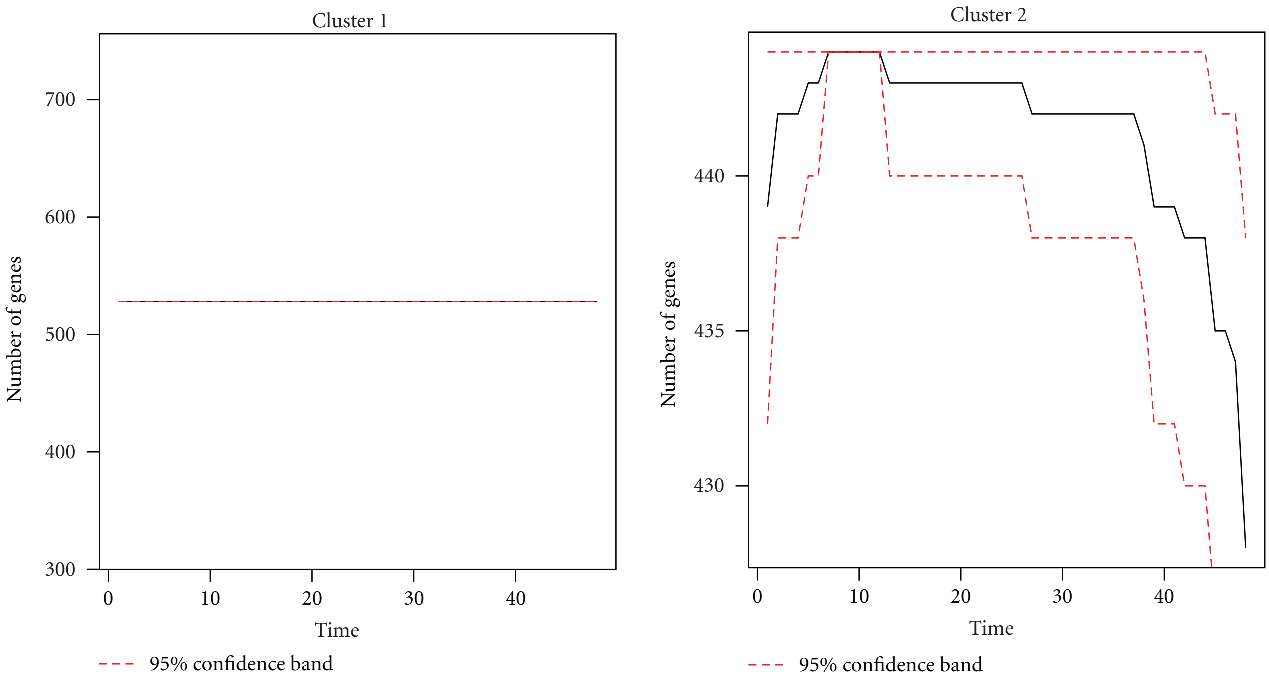

Figure 10: The number of genes varies across time in each of the two clusters found from dynamic clustering of 530 *P. falciparum* genes. The red dashed lines are the 95% bootstrap confidence band.

no information about time is contained in the clusters that are obtained by other methods. Time information is the critical component for both determining the clusters obtained by our method and calculating the "dynamic index" formula which measures the clustering performance.

The proposed method focuses on clustering periodic time series by considering the spectral frequencies that are decomposed and extracted from periodic data. Beyond the spectral frequency, phase information is obtained as well in the signal decomposition. In fact, clustering only the spectral

Table 2: *P. falciparum* genes appearing in both clusters and those having significant gene ontology (GO) terms.

GO id	GO name	FDR
GO:0006412	Translation	$1.55E-13$
GO:0006260	DNA replication	$1.12E-07$
GO:0006418	tRNA aminoacylation for protein translation	$7.87E-04$
GO:0043038	Amino acid activation	$7.87E-04$
GO:0043039	tRNA aminoacylation	$7.87E-04$
GO:0006399	tRNA metabolic process	0.00962

frequency of time series may not be sufficient to understand very complicated biological processes. In other words, even though two genes involved in the same biological process have the same spectral frequencies, they may play different roles in the process. For example, one gene may serve as a regulator for the other. Studying the phase relationship between genes may help understand such regulation and is a point of future research. Further, consider two genes that participate in different phases of a cellular process. A phase study is certainly necessary after clustering the spectral frequencies. It is worth noting that we cannot change the order of studies, (i.e., perform a phase study first followed by a frequency study) since genes with different spectral frequency must belong to different biological processes, and genes with different phases may or may not belong to the same process. Therefore, within each cluster of spectral frequency, genes can be subclustered according to their component phases so that the gene relationship may be revealed in greater detail [34].

Genes involved in multiple biological processes (simultaneously) may play a major role in one process while playing a minor role in another process. The importance of a gene in multiple processes has potential for further investigation. Since the energy of a time series is proportional to its amplitude squared, the importance of a gene in a process can be measured using the squared amplitude of its corresponding component. Based on this, and due to the fact that genes may participate in different processes at different time, the dynamic importance of genes in biological processes can be established. As a point of future research, if three features of periodic time series, namely, spectral frequency, phase shift, and amplitude, are all included, as well as the time information of components, the complex dynamic biological processes may be better understood [34].

Because the proposed dynamic clustering process is time dependent, an appreciation for the number of time points that are recommended for the method is a necessary discussion. Since the main feature of a periodic time series is spectral frequency, frequency detection is highly reliant on the sampling rate. Therefore, the minimal or recommended number of time points is related to the nature of biological processes/clusters in which the genes are involved. According to the Sampling Theorem [52], a signal can be exactly recovered if the sampling rate is greater than twice the signal frequency. Thus, if a signal has 5 cycles in one hour duration, the sampling rate must be more than 20 points in one hour. In other words, if the sampling rate is small, then a signal with high frequency cannot be detected.

Finally, since spectral frequencies are extracted from periodic time series, the time points occur at equally spaced intervals. For periodic time series with unevenly spaced points, evenly spaced time points can be artificially created by imputing missing data. Thus, our proposed approach is applicable, but some information may be lost due to data imputation. Further, since the proposed two-stage approach (i.e., data preparation and dynamic clustering) is designed for periodic data, for data that are not periodic, the signal decomposition approach is not applicable in the data preparation step. However, if some other characteristic can be defined and extracted from the nonperiodic time series, the second step of the proposed approach remains applicable.

Dynamic clustering is a two-step cluster validation that is able to differentiate meaningful clusters from noisy clusters. The results from our approach provide insight into the dynamic association among time-limited coexpressed genes that might otherwise go undetected by current clustering approaches. Clustering and gene network inference are both known to help in predicting the biological functions of genes or unraveling the mechanisms involved in biological processes. Usually clustering and gene network methods are developed independently. Specifically, in the gene networks, the challenge is to deal with a large number of genes, but when clustering, the clusters are assumed independent. In actuality, gene network procedures and gene clustering procedures cover each other's shortcomings [53]. As such, the proposed dynamic clustering has great potential for inferring gene networks, in particular, for exploring networks at the level of gene clusters. Although the proposed method is motivated by and explained in the context of time series microarray data, it is a general method that is applicable to any periodic phenomena, including but not limited to seasonal data in marketing research, meteorology, and astronomy.

Conflict of Interests

The authors declare that they have no conflict of interests.

Acknowledgment

This work was supported by the National Science Foundation [DBI 0501712 and 0733857] to RWD and DMS-1043080 to L. An.

References

[1] H. Yu, N. M. Luscombe, J. Qian, and M. Gerstein, "Genomic analysis of gene expression relationships in transcriptional regulatory networks," *Trends in Genetics*, vol. 19, no. 8, pp. 422–427, 2003.

[2] Y. Luan and H. Li, "Clustering of time-course gene expression data using a mixed-effects model with B-splines," *Bioinformatics*, vol. 19, no. 4, pp. 474–482, 2003.

[3] J. Qian, M. Dolled-Filhart, J. Lin, H. Yu, and M. Gerstein, "Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions," *Journal of Molecular Biology*, vol. 314, no. 5, pp. 1053–1066, 2001.

[4] Y. Cheng and G. Church, "Biclustering of expression data," in *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*, pp. 93–103, 2000.

[5] L. Lazzeroni and A. Owen, "Plaid models for gene expression data," *Statistica Sinica*, vol. 12, no. 1, pp. 61–86, 2002.

[6] L. Ji and K. L. Tan, "Identifying time-lagged gene clusters using gene expression data," *Bioinformatics*, vol. 21, no. 4, pp. 509–516, 2005.

[7] J. Z. Song, K. M. Duan, T. Ware, and M. Surette, "The wavelet-based cluster analysis for temporal gene expression data," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2007, Article ID 39382, 2007.

[8] S. Madeira and A. Oliveira, "An efficient biclustering algorithm for finding genes with similar patterns in time-series expression data," in *Proceedings of 5th Asia Pacific Bioinformatics*, pp. 67–80, 2007.

[9] Y. Zhang, H. Zha, and C. H. Chu, "A time-series biclustering algorithm for revealing co-regulated genes," in *Proceedings of the International Conference on Information Technology: coding and Computing (ITCC '05)*, pp. 32–37, April 2005.

[10] G. Palla, A. L. Barabási, and T. Vicsek, "Quantifying social group evolution," *Nature*, vol. 446, no. 7136, pp. 664–667, 2007.

[11] E. R. Dougherty, I. Shmulevich, and M. L. Bittner, "Genomic signal processing: the salient issues," *Eurasip Journal on Applied Signal Processing*, vol. 2004, no. 1, pp. 146–153, 2004.

[12] R. Carmona, W. Hwang, and B. Torresani, *Practical Time-frequency Analysis: Gabor and Wavelet Transforms with an Implementation in S Wavelet Analysis and Its Applications*, Academic Press, 1998.

[13] S. Qian, *Introduction to Time-Frequency and Wavelet Transforms*, Prentice-Hall, 2002.

[14] P. Addison, *The Illustrated Wavelet Transform Handbook*, Taylor & Francis, 2002.

[15] C. Furlanello, S. Merler, and G. Jurman, "Combining feature selection and DTW for time-varying functional genomics," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 2436–2443, 2006.

[16] P. Goupillaud, A. Grossmann, and J. Morlet, "Cycle-octave and related transforms in seismic signal analysis," *Geoexploration*, vol. 23, no. 1, pp. 85–102, 1984.

[17] A. J. Butte, L. Bao, B. Y. Reis, T. W. Watkins, and I. S. Kohane, "Comparing the similarity of time-series gene expression using signal processing metrics," *Journal of Biomedical Informatics*, vol. 34, no. 6, pp. 396–405, 2001.

[18] U. Grenander, *The NyguIst Frequency Is That Frequency Whose Period Is Two Sampling Intervals. Probability and StatIstics: The Harald Cramer Volume*, Wiley, 1959.

[19] G. J. Szekely and M. L. Rizzo, "Brownian distance variance," *The Annals of Applied Statistics*, vol. 3, no. 4, pp. 1236–1265, 2009.

[20] L. An and R. W. Doerge, "Dynamic clustering of cell-cycle gene expression data," in *Proceedings of the Kansas State University Conference on Applied Statistics in Agriculture*, pp. 18–36, Manhattan, NY, USA, 2008.

[21] J. Ward, "Hierarchical grouping to optimize an objective function," *Journal of American Statistical Association*, vol. 58, pp. 236–244, 1963.

[22] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On clustering validation techniques," *Journal of Intelligent Information Systems*, vol. 17, no. 2-3, pp. 107–145, 2001.

[23] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, no. 2, pp. 159–179, 1985.

[24] K. Y. Yeung, D. R. Haynor, and W. L. Ruzzo, "Validating clustering for gene expression data," *Bioinformatics*, vol. 17, no. 4, pp. 309–318, 2001.

[25] S. Salvador and P. Chan, "Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms," in *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI '04)*, pp. 576–584, November 2004.

[26] H. Chipman, T. Hastie, and R. Tibshirani, "Clustering microarray data," in *Statistical Analysis of Gene Expression Microarray Data*, T. Speed, Ed., pp. 159–201, Chapman & Hall/CRC Press, 2003.

[27] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society. Series B*, vol. 63, no. 2, pp. 411–423, 2001.

[28] A. Brondsted, *Introduction to Convex PolyTopes*, Springer, New York, NY, USA, 1983.

[29] B. Munneke, K. A. Schlauch, K. L. Simonsen, W. D. Beavis, and R. W. Doerge, "Adding confidence to gene expression clustering," *Genetics*, vol. 170, no. 4, pp. 2003–2011, 2005.

[30] G. C. Tseng and W. H. Wong, "Tight clustering: a resampling-based approach for identifying stable and tight patterns in data," *Biometrics*, vol. 61, no. 1, pp. 10–16, 2005.

[31] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[32] L. Collins and C. Dent, "Omega: a general formulation of the rand index of cluster recovery suitable for non-disjoint solutions," *Multivariate Behavioral Research*, vol. 23, pp. 231–242, 1988.

[33] R. J. G. B. Campello, "A fuzzy extension of the Rand index and other related indexes for clustering and classification assessment," *Pattern Recognition Letters*, vol. 28, no. 7, pp. 833–841, 2007.

[34] L. An, *Dynamic clustering of time series gene expression [Ph.D. thesis]*, Purdue University, West Lafayette, Ind, USA, 2008.

[35] Z. Bozdech, M. Llinás, B. L. Pulliam, E. D. Wong, J. Zhu, and J. L. DeRisi, "The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*," *PLoS Biology*, vol. 1, no. 1, article e5, 2003.

[36] R. S. Istepanian, A. Sungoor, and J. C. Nebel, "Comparative analysis of genomic signal processing for microarray data clustering," *IEEE Trans Nanobioscience*, vol. 10, no. 4, pp. 225–238, 2011.

[37] M. L. Whitfield, G. Sherlock, A. J. Saldanha et al., "Identification of genes periodically expressed in the human cell cycle and their expression in tumors," *Molecular Biology of the Cell*, vol. 13, no. 6, pp. 1977–2000, 2002.

[38] P. T. Spellman, G. Sherlock, M. Q. Zhang et al., "Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization," *Molecular Biology of the Cell*, vol. 9, no. 12, pp. 3273–3297, 1998.

[39] R. J. Cho, M. J. Campbell, E. A. Winzeler et al., "A genome-wide transcriptional analysis of the mitotic cell cycle," *Molecular Cell*, vol. 2, no. 1, pp. 65–73, 1998.

[40] M. J. Gardner, N. Hall, E. Fung et al., "Genome sequence of the human malaria parasite *Plasmodium falciparum*," *Nature*, vol. 419, no. 6906, pp. 498–511, 2002.

[41] L. Du, S. Wu, A. W. C. Liew, D. K. Smith, and H. Yan, "Spectral analysis of microarray gene expression time series data of *Plasmodium falciparum*," *International Journal of Bioinformatics Research and Applications*, vol. 4, no. 3, pp. 337–349, 2008.

[42] A. Kallio, N. Vuokko, M. Ojala, N. Haiminen, and H. Mannila, "Randomization techniques for assessing the significance of gene periodicity results," *BMC Bioinformatics*, vol. 12, article 330, 2011.

[43] R. Jurgelenaite, T. M. H. Dijkstra, C. H. M. Kocken, and T. Heskes, "Gene regulation in the intraerythrocytic cycle of *Plasmodium falciparum*," *Bioinformatics*, vol. 25, no. 12, pp. 1484–1491, 2009.

[44] M. Hirsch, S. Swift, and X. Liu, "Optimal search space for clustering gene expression data via consensus," *Journal of Computational Biology*, vol. 14, no. 10, pp. 1327–1341, 2007.

[45] O. Troyanskaya, M. Cantor, G. Sherlock et al., "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.

[46] B. Efron and R. Tibshirani, *Introduction to the Bootstrap*, Chapman and Hall, New York, NY, USA, 1993.

[47] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature Protocols*, vol. 4, no. 1, pp. 44–57, 2009.

[48] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Research*, vol. 37, no. 1, pp. 1–13, 2009.

[49] C. Aurrecoechea, J. Brestelli, B. P. Brunk et al., "PlasmoDB: a functional genomic database for malaria parasites," *Nucleic Acids Research*, vol. 37, no. 1, pp. D539–D543, 2009.

[50] http://david.abcc.ncifcrf.gov.

[51] L. C. Wu, J. L. Huang, J. T. Horng, and H. D. Huang, "An expert system to identify co-regulated gene groups from time-lagged gene clusters using cell cycle expression data," *Expert Systems with Applications*, vol. 37, no. 3, pp. 2202–2213, 2010.

[52] J. Proakis and D. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*, Macmillan, 1992.

[53] Y. Shiraishi, S. Kimura, and M. Okada, "Inferring cluster-based networks from differently stimulated multiple time-course gene expression data," *Bioinformatics*, vol. 26, no. 8, Article ID btq094, pp. 1073–1081, 2010.