JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 17, Number 8, 2010

Pp. 1067–1080
DOI: 10.1089/cmb.2009.0270

Wavelet-Based Functional Clustering for Patterns of High-Dimensional Dynamic Gene Expression

BONG-RAE KIM,[1] TIMOTHY MCMURRY,[2] WEI ZHAO,[3]
RONGLING WU,[4] and ARTHUR BERG[4]

ABSTRACT

Functional gene clustering is a statistical approach for identifying the temporal patterns of gene expression measured at a series of time points. By integrating wavelet transformations, a power dimension-reduction technique, noisy gene expression data is smoothed and clustered allowing for new patterns of functional gene expression profiles to be identified. We implement the idea of wavelet dimension reduction into the mixture model for gene clustering, aimed to de-noise the data by transforming an inherently high-dimensional biological problem to its tractable low-dimensional representation. As a first attempt of its kind, we capitalize on the simplest Haar wavelet shrinkage technique to break an original signal down into its spectrum by taking its averages and differences and, subsequently, detect gene expression patterns that differ in the smooth coefficients extracted from noisy time series gene expression data. The method is shown to be effective on simulated data and and on recent time course gene expression data. Supplementary Material is available at www .liebertonline.com.

Key words: algorithms, statistics.

1. INTRODUCTION

ALTHOUGH HIGH-THROUGHPUT TECHNOLOGIES, such as DNA microarrays and proteomics platforms, have provided researchers with a set of unprecedented tools to ask and address various fundamental questions in developmental biology and biomedicine, the use of these technologies that generate enormous amounts of gene or protein data from biological entities relies critically on statistical analysis and modeling of the data.

The past decade has witnessed an astonishing development of statistical methods for cataloguing the patterns of gene expression and using these distinct patterns to assessing developmental functions and mechanisms of a biological phenomena (Eisen et al., 1998; Ramoni et al., 2002; Ghosh and Chinnaiyan, 2002; McLachlan et al., 2002; Zapala and Schork, 2006). More recently, there has been a considerable body

[1]Department of Dentistry, Seoul National University, Seoul, Republic of Korea.
[2]Department of Mathematical Sciences, DePaul University, Chicago, Illinois.
[3]Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, Tennessee.
[4]Department of Biostatistics, Pennsylvania State University, Hershey, Pennsylvania.

of literature about the derivations of statistical methods for clustering time-dependent gene expression (Qian et al., 2001; Holter et al., 2001; Zhao et al., 2001; Park et al., 2003; Bar-Joseph et al., 2003; Luan and Li, 2003; Ernst et al., 2005; Storey et al., 2005; Ma et al., 2006; Ng et al., 2006; Inoue et al., 2007; Kim et al., 2008; Wang et al., 2009).

The central idea of functional gene clustering is to mathematically model the mean vectors for each gene pattern within the mixture model context incorporating the structure of covariance of the gene expressions measured at discrete time points. Such mathematical modeling has two major advantages. First, instead of estimating every mean at each time point and every element in the covariance matrix, functional clustering only needs to estimate a reduced number of mathematical parameters that model the mean-covariance structures. This provides greater power to detect significantly differentiated patterns during a time course. Second, gene expression profiles related to many biological processes have a certain pattern, which can be described robustly by mathematical functions. By estimating the parameters that determine mathematical functions, the genetic differentiation over time course can be estimated and tested. The results from these biologically justified models are, therefore, more closer to biological reality.

Despite its statistical and biological relevances, functional clustering has two significant limitations that may prevent its broad and deep uses in some particular situations. First, it does not allow the number of repeated measurements (defined as the dimensionality of observation) to unlimitedly increase for robust parameter estimation. While increased dimensionality possesses richer information, structural modeling of high-dimensional variances and correlations will be computationally expensive. With increasing dimension, the computation of inverse covariance matrix will tend to be unstable. Second, in practice, the sparsity of a data set increases exponentially with its dimensionality. Functional gene clustering based on a multivariate normal density function will be affected for high-dimensional data as measurement error will become increasingly problematic in parameter estimation of the classical mixture models.

An efficient treatment of high-dimensional microarray data is through dimensionality reduction, i.e., the transformation that brings data from a high- to low-order dimension. It has been shown that models with low dimension are not only computationally efficient, but also more robust than high dimensional models. Wavelet transforms that preserve signal pattern and yield better or comparative classification accuracy provide a powerful tool for dimensionality reduction (Donoho, 1995; Donoho and Johnstone, 1994).

In this article, we derive a wavelet-based de-noising method for functional clustering of time-dependent microarray gene expression data. By reducing the dimensionality of data, this method improves the accuracy and power of gene cluster detection in many situations.

2. METHODS

2.1. Wavelet Transform

According to wavelet transform methodology, an original signal is divided into two sequences each with a length equal to a half of the original signal length (Mallat et al., 1989; Vidakovic, 1999; Jensen and la Cour-Harbo, 2001). The first sequence, denoted as the *smooth coefficients* (or *approximation coefficients*), corresponds to an approximation process of the original signal, whereas the second sequence, denoted as the *detail coefficients,* corresponds to the detail information (subtleties) that is complementary to the approximation process. We use c^{-r} and d^{-r} to denote the smooth and detail coefficients, respectively, where superscript $-r$ indicates the resolution level at which the initial sequence is split into smooth and detail coefficients. Since detail coefficients are contaminated severely by random errors, shrinking them to zero will be helpful for reducing the overall noise level of the signal (Donoho, 1995; Donoho and Johnstone, 1994).

As the simplest wavelet transform, discrete Haar transform calculates detail coefficients by subtracting successive values in the sequence (Walker, 1999). The data of expression profile for gene i measured at T time points can be expressed as

$$\boldsymbol{y}_i = \{y_i(1), y_i(2), \ldots, y_i(T-1), y_i(T)\}. \quad (1)$$

The smooth and detail coefficients of the original signal after the first-resolution Haar wavelet transform are arrayed by

$$
\begin{aligned}
z_i^{-1} &= \left\{ \frac{y_i(1)+y_i(2)}{\sqrt{2}}, \ldots, \frac{y_i(T-1)+y_i(T)}{\sqrt{2}}, \right. \\
&\qquad \left. \frac{y_i(1)-y_i(2)}{\sqrt{2}}, \ldots, \frac{y_i(T-1)-y_i(T)}{\sqrt{2}} \right\} \\
&= \left\{ c_i^{-1}(1), \ldots, c_i^{-1}\left(\frac{T}{2}\right), d_i^{-1}(1), \ldots, d_i^{-1}\left(\frac{T}{2}\right) \right\} \\
&= \left\{ \{c_i^{-1}(\tau)\}_{\tau=1}^{\frac{T}{2}}, \{d_i^{-1}(\tau)\}_{\tau=1}^{\frac{T}{2}} \right\}
\end{aligned}
\tag{2}
$$

where τ is the new sequence index used after applying the Haar wavelet transform. It can be seen that variation in detail coefficients at resolution 1 only reflects local fluctuations between the nearest neighbors in the sequence. Similarly, smooth coefficients at resolution 2 are obtained by summing pairs of resolution 1 smooth coefficients. At each resolution, the number of smooth and detail coefficients obtained drops by 1/2. The process can be repeated, each time reducing the dimension of the smooth and detail coefficients (Walker, 1999).

The pattern of the smooth coefficients in the wavelet space resembles the signal pattern in the time space. In Figure 1, $\mathbf{c^{-0}}$ represents a sample of repeated measurements of a pattern at 24 time points. The smooth coefficients of the first and second Haar wavelet transformation are plotted as $\mathbf{c^{-1}}$ and $\mathbf{c^{-2}}$, respectively. The pattern of $\mathbf{c^{-1}}$ and $\mathbf{c^{-2}}$ coefficients conform to the signal pattern although they are in two different resolution levels. Because of the similarity, it is reasonable to model the original pattern based on low-dimensional smooth coefficients.

2.2. *Thresholding*

In wavelet transform, we need to find an approximation of the original signal which is smooth and can also adequately represent the input signal. Such an approximation can be detected by two thresholding approaches—the *hard threshold filter* (H_h) and the *soft threshold filter* (H_s). The hard threshold filter, also known as the "keep or kill" method (Aboufadel and Schlicker, 1999), removes coefficients below a threshold value determined by the estimated noise variance. The soft threshold shrinks large coefficients towards zero but also completely removes the smaller coefficients (Ghael et al., 1997). Since the de-noised signal is irreversibly different than the noisy signal, thresholding induces a loss of information.

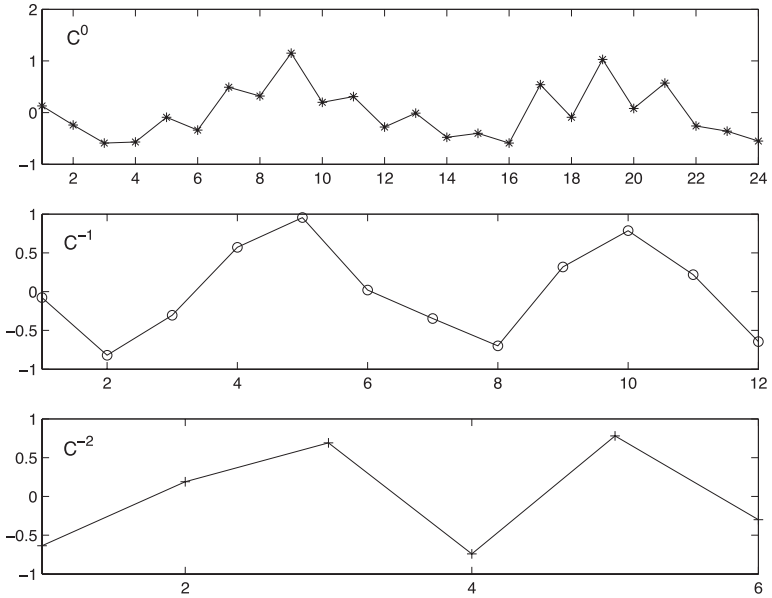

FIG. 1. The original periodic profile of a gene ($\mathbf{c}^0$), subject to Haar wavelet transform at the first ($\mathbf{c}^{-1}$) and second resolutions ($\mathbf{c}^{-2}$). The transformed data preserve most information of the original signal, although the lower-order resolution tends to be close to the original signal than does the higher-order resolution.

The two thresholding approaches will produce different results. To make the resulting signal smoother, the soft threshold filter should be used, whereas, to make computation fasters, the hard threshold filter may be used. In practice, it is difficult to choose a threshold value because a small threshold value may not be able to remove a noise while a large threshold value introduces a bias. Many approaches have been available to determine an optimal threshold value. One universal method is to assign a threshold value given by

$$\lambda_T = \hat{\sigma}\sqrt{2\log T}, \tag{3}$$

where $\hat{\sigma}$ is the estimator of standard deviation of the noise, and T is the length of the input vector (Donoho and Johnstone, 1994). The hard thresholding rule is defined as

$$\delta(y, \lambda_T) = \begin{cases} y, & \text{if } |y| > \lambda_T \\ 0, & \text{if} |y| < \lambda_T \end{cases} \tag{4}$$

As pointed out in Donoho and Johnstone (1994) and Johnstone and Silverman (1997), for a sequence of normal distributed random variables $z(t) \sim N(0, \sigma^2(t))\,(t = 1, \ldots, T)$, we have

$$P(\max_t \left| \frac{z(t)}{\sigma(t)} \right| > \sqrt{2\log T}) \to 0 \quad \text{as } T \to \infty. \tag{5}$$

Hence, if a detail coefficient is truly zero, then with a high probability it is estimated as zero in terms of the hard thresholding rule. The expected number of $\left|\frac{z(t)}{\sigma(t)}\right|$ greater than the threshold tends to zero. For most applications, the hard thresholding rule only keeps those detail coefficients that are significantly greater than zero. Here, the hard thresholding rule is used to either keep or kill the whole level of detail coefficients.

The following procedure is proposed to perform data dimensionality reduction through wavelet transforms:

(1) Select proper orthogonal wavelet filters;
(2) Calculate empirical variances for the detail coefficients;
(3) Apply the hard thresholding rule to the detail coefficients;
(4) Truncate the whole level of the detail coefficients if they are all set to zero by (3), and keep the whole level of the detail coefficients otherwise;
(5) Repeat procedures (1) to (4) for user-prescribed j times.

Different wavelet filters vary in filter length. A longer filter length wavelet tends to ''average'' over more signal points. The purpose of hard thresholding is to reduce the dimensionality of the data by truncating certain levels of detail coefficients. The variance estimator $\sigma^2_{-\mathrm{r}}(\tau)$ for each detail coefficient $\mathrm{d}^{-\mathrm{r}}(\tau)$ is suggested in Donoho and Johnstone (1994) and Johnstone and Silverman (1997), i.e.,

$$\sigma^2_{-\mathrm{r}}(\tau) = \frac{\mathrm{MAD}\{d_t^{-\mathrm{r}}(\tau), t = 1, \ldots, T\}}{0.6745}, \tag{6}$$

where MAD denotes the median absolute deviation and 0.6745 is chosen to adjust for a normal distribution.

2.3. *Wavelet-based functional clustering*

2.3.1. Likelihood. Suppose there are n genes each measured at T equally-spaced time points. Let $\boldsymbol{y}_i = (y_i(1), \ldots, y_i(T))$ be the gene expression data for gene i. If these genes are clustered into J patterns, this means that any one of the genes (i) is assumed to arise from one (and only one) of the J possible expression patterns. Thus, the distribution of gene expression data is expressed as the J-component mixture probability density function, i.e.,

$$\boldsymbol{y}_i \sim f(\boldsymbol{y}_i; \omega, \boldsymbol{u}_i, \Sigma) = \sum_{j=1}^{J} \omega_j f_j(\boldsymbol{y}_i; \boldsymbol{u}_{j(i)}, \Sigma), \tag{7}$$

where $\omega = (\omega_1, \cdots, \omega_J)$ is the mixture proportions which are non-negative and sum to unity; $\boldsymbol{u}_i = (\boldsymbol{u}_{1(i)}, \cdots, \boldsymbol{u}_{J(i)})$ contains the component- (or pattern) specific mean vector for gene i; and Σ contains residual variances and covariances among the T time points for gene i which are common for all gene

expression patterns. For a given gene i, the probability density function of the jth gene expression pattern or cluster, $f_j(y_i; \mathbf{u}_{j(i)}, \Sigma)$, is assumed to be multivariate normally distributed with mean vector

$$\boldsymbol{u}_{j(i)} = \boldsymbol{u}_j = (u_j(1), u_j(2), \ldots, u_j(T-1), u_j(T)) \tag{8}$$

and common covariance matrix Σ. For simplicity of notation, we drop the understood index i from the mean vector $\boldsymbol{u}_{j(i)}$ as it is understood to depend on gene i.

As shown in equation (2), the original signal (1) is subject to wavelet transform at the first resolution. Correspondingly, the smooth and detail coefficients of gene pattern-specific mean signals $\boldsymbol{u}_j$ after the first-resolution Haar wavelet transform can be arrayed as

$$\begin{aligned}\boldsymbol{w}_j^{-1} &= \left\{\frac{u_j(1)+u_j(2)}{\sqrt{2}}, \ldots, \frac{u_j(T-1)+u_j(T)}{\sqrt{2}}, \right. \\ &\quad \left. \frac{u_j(1)-u_j(2)}{\sqrt{2}}, \ldots, \frac{u_j(T-1)-u_j(T)}{\sqrt{2}}\right\} \\ &= \left\{w_{c_j}^{-1}(1), \ldots, w_{c_j}^{-1}\left(\frac{T}{2}\right), w_{d_j}^{-1}(1), \ldots, w_{d_j}^{-1}\left(\frac{T}{2}\right)\right\} \\ &= \left\{\left\{w_{c_j}^{-1}(\tau)\right\}_{\tau=1}^{\frac{T}{2}}, \left\{w_{d_j}^{-1}(\tau)\right\}_{\tau=1}^{\frac{T}{2}}\right\},\end{aligned} \tag{9}$$

Now, let $z^{-\mathrm{r}}$ be the new variable with a reduced dimension $T_{-\mathrm{r}}$ ($T_{-\mathrm{r}} < T$) transformed from the rth resolution Haar wavelet. The likelihood function based on a mixture model containing J gene expression patterns can be rewritten, in terms of $z^{-\mathrm{r}}$, as

$$L\left(\Omega^{-\mathrm{r}} | z^{-\mathrm{r}}\right) = \prod_{i=1}^{n} \sum_{j=1}^{J} \left[\omega_j f_j\left(\boldsymbol{z}_i^{-\mathrm{r}}; \boldsymbol{w}_{\mathrm{c_j}}^{-\mathrm{r}}, \Sigma_{-\mathrm{r}}\right)\right], \tag{10}$$

where $\Omega^{-\mathrm{r}} = (\{\omega_j, \boldsymbol{w}_{c_j}^{-\mathrm{r}}\}_{j=1}^{J}, \Sigma_{-\mathrm{r}})$ contains unknown parameters, $(\omega_1, \ldots, \omega_J)$ are the mixture proportions of J different gene expression patterns, as shown in equation (7), and $f_j(\mathbf{z}_i^{-\mathrm{r}}; \boldsymbol{w}_{c_j}^{-\mathrm{r}}, \Sigma_{-\mathrm{r}})$ is the multivariate normal distribution of gene i that belongs to gene expression pattern j, in which $\boldsymbol{z}_i^{-\mathrm{r}} = \{c_i^{-\mathrm{r}}(1), \ldots, c_i^{-\mathrm{r}}(T_{-\mathrm{r}})\}$ is a vector of smooth coefficients for gene i, $\boldsymbol{w}_{c_j}^{-\mathrm{r}} = \{w_{c_j}^{-\mathrm{r}}(1), \ldots, w_{c_j}^{-\mathrm{r}}(T_{-\mathrm{r}})\}$ is a vector of expected smooth coefficients for gene expression pattern j and $\Sigma_{-\mathrm{r}}$ is the $(T_{-\mathrm{r}} \times T_{-\mathrm{r}})$ residual covariance matrix for the smooth coefficients.

2.3.2. Modeling wavelet-based mean vectors. It is well known that the transcript levels of many DNA microarrays in terms of the amount of mRNAs vary with a particular pattern in time course. For example, the amount of mRNAs within the cell division cycle may change periodically (Spellman et al., 1998; De Lichtenberg et al., 2005). The regulation of these genes in a periodic manner coincident with the cell cycle may help maintain proper order during cell division and may also aid in conserving limited resources. The oscillation of cell cycle-regulated genes can be mathematically described by a simple periodic Fourier function expressed as a linear combination of cosine and sine waves. Thus, by estimating the parameters that define the periodic curves for individual genes, we can determine the differences in the temporal pattern of gene expression.

For periodically regulated genes, they can be approximated by Fourier series (Lasser, 1996). Fourier series approximation can assess periodicity. So, by applying a Fourier series approximation, we can identify the genes whose RNA levels varied periodically within the cell cycle and further find the associated amplitudes and phases. For a given gene expression pattern, let $u_j(t)$ denote the expected gene intensity ratio value at time point $t (t = 1, \ldots, T)$. Note that the ratio values are log transformed (base 2 for simplicity, so that $\log_2$ ($Cy5/Cy3$)) to treat inductions or repressions of identical magnitude as numerically equal but with opposite sign. The mean vector for a given gene expression pattern, $\boldsymbol{u}_j = (u_j(1), \ldots, u_j(T))$, can be modeled by a Fourier series approximation of order one. Thus, the log ratio gene expression value of gene expression pattern j at time point t can be expressed as

$$u_j(t) = \frac{1}{2}a_{j0} + \left[a_{j1}\cos\left(\frac{2\pi t}{\tau_j}\right) + b_{j1}\sin\left(\frac{2\pi t}{\tau_j}\right)\right], \tag{11}$$

where a_{j0} is the gene-specific fundamental frequency, a_{j1} and b_{j1} are the pattern-specific amplitude coefficients, which determine the times at which the gene achieves peak and trough expression levels, respectively, and τ_j is the gene-specific period of the cell cycle.

In general, the gene expression value of pattern j in time course can be mathematically fitted in form $u_j(t; \Omega_{u_j})$ by a set of curve parameters Ω_{u_j}. The mean vector transformed at the first resolution transformation is expressed as

$$\boldsymbol{w}_{c_j}^{-\mathbf{1}} = \left(\frac{u_j(1; \Omega_{u_j}) + u_j(2; \Omega_{u_j})}{\sqrt{2}}, \ldots, \frac{u_j(T-1; \Omega_{u_j}) + u_j(T; \Omega_{u_j})}{\sqrt{2}} \right) \tag{12}$$

Thus, by estimating Ω_{u_j} with transformed data at an appropriate transformation resolution $-$r, a gene expression curve in time course can be elucidated for individual patterns. Differences of the curves ($\boldsymbol{w}_{c_j}^{-r}$) can be compared and tested for the statistical significance of time-dependent gene expression patterns.

2.3.3. Modeling the covariance structure. It is not parsimonious to estimate all the elements in the covariance matrix among different time points because some structure exists for time-dependent variances and correlations. The covariance structure in the wavelet-domain can be modeled by a stationary first-order autoregressive (AR(1)) model (Diggle et al., 2002), expressed as

$$\begin{cases} \sigma^2(1) = \cdots = \sigma^2(T) = \sigma^2 & \text{for variance} \\ \sigma(t_1, t_2) = \sigma^2 \rho^{|t_2 - t_1|} & \text{for covariance,} \end{cases} \tag{13}$$

where $0 < \rho < 1$ is the proportion parameter with which the correlation decays with time lag. The parameters for the covariance structure are arrayed in $\Omega_v = (\rho, \sigma^2)$.

2.3.4. Estimation and tests. The standard EM algorithm is derived to estimate the parameters contained in the likelihood (10). Since the actual number of gene expression patterns is unknown, we will employ the commonly used model selection methods, AIC or BIC, to estimate the optimal number of components in the mixture model (10). After the optimal number of gene expression patterns is determined, a variety of biologically meaningful hypotheses can be formulated and tested. The most important hypothesis about overall differences in transcriptional expression profile among different patterns of microarray genes is formulated as

$$\begin{aligned} &H_0 : \Theta_{u_j} \equiv \Theta_u, \text{ for } j = 1, \ldots, J \\ &H_1 : \text{ At least one of the equalities above does not hold.} \end{aligned} \tag{14}$$

The log-likelihood ratio (LR) test statistic is then calculated by

$$\mathrm{LR} = -2[\ln L(\tilde{\Omega}|\mathbf{y}) - \ln L(\hat{\Omega}|\mathbf{y})],$$

where the tildes and hats stand for the MLEs of the unknown parameters under the null and alternative hypotheses, respectively. The null hypothesis means that no different patterns of temporal expression exist among the genes studied, whereas the alternative hypothesis states that at least two different patterns can be identified. The critical threshold for claiming distinguishable expression patterns can be determined on the basis of simulation studies.

3. IMPLEMENTATION

3.1. Simulated data application

Time course gene expressions for 5000 genes were simulated over 40 equally spaced time points with mean expression profiles generated from one of the eight curves pictured in Figure 2. Residual error on the simulated series was generated from a stationary Gaussian autoregressive process with autocorrelation parameter $\rho = 0.5$ and standard deviation $\sigma = 0.3$. In the real data analysis performed below, we found the standard error of the clustered expression profiles to be around $\hat{\sigma} = 0.15$. Therefore the simulated data

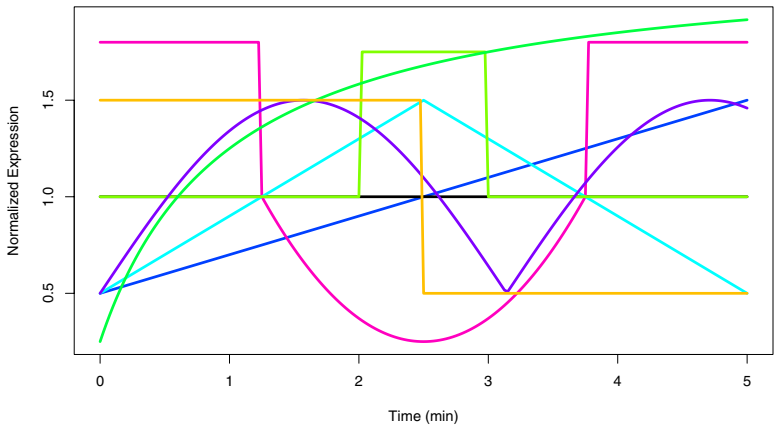

FIG. 2. True mean curves (eight in total) from which gene expression data was simulated from. Eighty percent of the expression profiles were simulated from a constant expression of one.

presents a smaller signal-to-noise ratio as compared to the real data analyzed below. Out of the 5000 simulated genes, 4000 were simulated from a flat signal with a constant value of one. The number of genes simulated under the other mean expression profiles are listed in Table 1.

Without assuming the number of clusters is known (even though it is), we utilized the AIC and BIC empirically identify the optimal number of clusters. In Figure 3, the AIC and BIC values are graphed under three levels of Haar wavelet smoothing: no wavelet smoothing ($r = 0$), one level of smoothing ($r = 1$), and two levels of smoothing ($r = 2$). Without wavelet smoothing, AIC and BIC suggest five clusters, with one level of smoothing seven clusters are suggested, and eight clusters are suggested with two levels of smoothing. When two levels of smoothing are applied, a seemingly more robust number of clusters are selected by AIC/BIC, and the correct number of clusters (eight) were identified.

Although the AIC/BIC values without wavelet smoothing are rather erratic, it is interesting to note that the overall minimum of AIC and BIC across the three levels of smoothing is obtained under no smoothing. Without wavelet smoothing, however, the AIC/BIC identified only five clusters, and even when eight clusters are carefully considered under no smoothing (refer to Appendix A in Supplementary Materials; for online Supplementary Material, see www.liebertonline.com), not all eight clusters are correctly identified.

Looking more carefully at the eight clusters identified by two levels of smoothing, the eight estimated mean curves are graphed in Figure 4 which are shown to closely follow the true cluster means displayed in Figure 2. The dimension reduction induced by the Haar wavelet transformation is evident in the wavelet means. These eight fitted clusters are individually analyzed in Figure 5.

A gene will be classified to a specific cluster if it has at least a 90% estimated probability of belonging to that cluster. Some genes will not be classified to a specific cluster if the estimated cluster probabilities are all less than 90% (though the sum of estimated cluster probabilities is always 100%). For each identified

TABLE 1. ORIGINAL NUMBER OF GENES ALLOCATED TO THE EIGHT CLUSTERS ARE RECORDED HERE ALONG WITH THE NUMBER OF GENES THAT WERE CORRECTLY AND INCORRECTLY CLASSIFIED; A GENE IS SAID TO BE CLASSIFIED TO A GIVEN CLUSTER IF THE ESTIMATED CLUSTER PROBABILITY FOR THE CORRESPONDING CLUSTER IS AT LEAST 90%

Cluster	*Original*	*Correctly classified*	*Incorrectly classified*
Flat	4000	3840	25
Linear increase	150	107	1
Smile	50	50	0
Hat	150	99	1
Off-on-off	200	163	1
Abs(sin)	100	85	3
Curved increase	150	150	0
On-off	200	198	0
Total	**5000**	**4692**	**31**

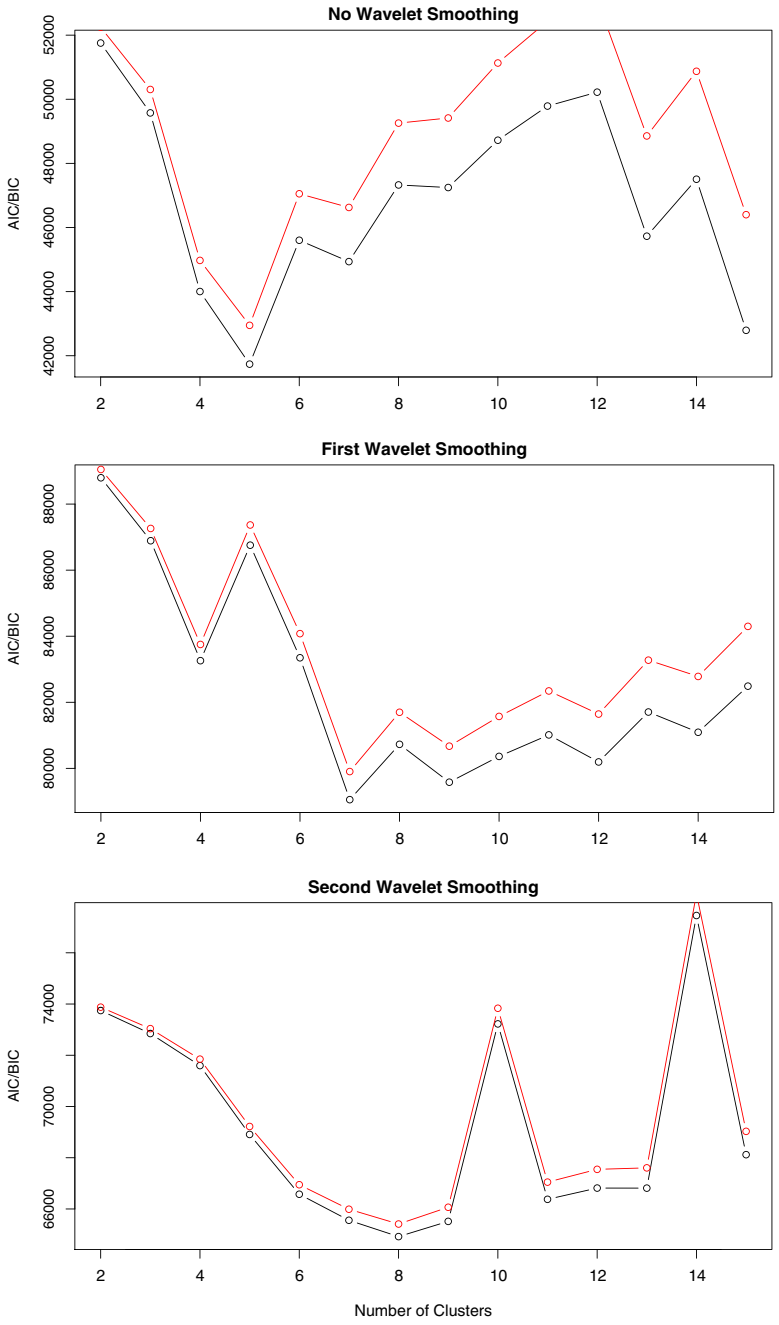

FIG. 3. AIC (black) and BIC (red) values under three levels of wavelet smoothing.

cluster, the classified genes (based on the 90% threshold) are graphed in Figure 5, along with the mean wavelet curve and the true mean curve that corresponds to the cluster. Gene expression profiles are colored according to their original cluster allowing inappropriately clustered genes to be easily identified. The total number of correctly and incorrectly classified are included listed in Table 1. This simulation analysis indicates that incorporating wavelets into functional clustering can be a powerful tool for optimally identifying nonparametric signals and clusters within time-course data.

3.2. Real data application

This methodology is applied to time time course gene expression data published in Rustici et al. (2004). In their research, a total of 8 time-course experiments were performed with expression data collected at

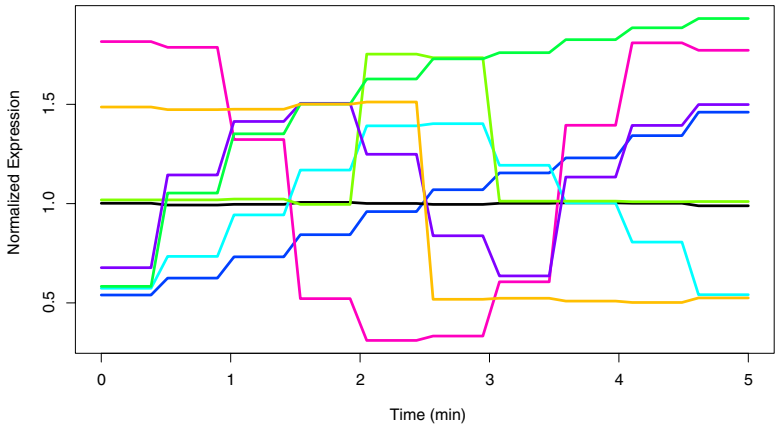

FIG. 4. Estimated wavelet mean curves of the eight clusters as selected by AIC and BIC under two levels of smoothing; all clusters were correctly identified.

18–22 times on 15-minute intervals. We analyzed data from one time-course experiment; the raw and processed datasets are accessible from ArrayExpress with accession number E-MEXP-54. After selecting genes with full time-course expression profiles, 2955 genes were used with 21 equally spaced time measurements taken for each gene.

As only 21 time measurements are available for this dataset, only one level of Haar wavelet smoothing ($r = 1$) was incorporated. The initial values of the parameters in the EM algorithm were randomly selected from a normal distribution with mean $\sqrt{2}$ and variance 1. The AIC and BIC were used to identify the best number of clusters, and they both selected thirteen clusters. The AIC and BIC values are graphed in Figure 6.

The estimated mean curves for these clusters are graphed in Figure 7. As before in the simulated data analysis, a gene is classified to a given cluster if it has an estimated probability of belonging of at least 90%. For each of the identified clusters, the number of genes classified to that cluster are tabulated in Table 2. For instance, only four genes are seen to be strongly clustered in the first cluster whereas 447 genes indicate a strong classification to the last cluster. A total of 1540 genes (52%) were not strongly classified to one of the thirteen clusters. Each mean curve is individually graphed with the expression profiles of the genes belonging to the cluster in in Figure 8. Several interesting clusters were identified in this dataset. By plotting the individual mean curves with the classified genes separately, one can get a better sense of the true nature of each of the identified clusters. Some clusters were identified with only a few genes but with very unique profiles and other clusters indicate multiple gene with expressions in sync with a periodic signal. For biological relevance from this clustering application, the reader is can refer to the original application of this data in Rustici et al. (2004).

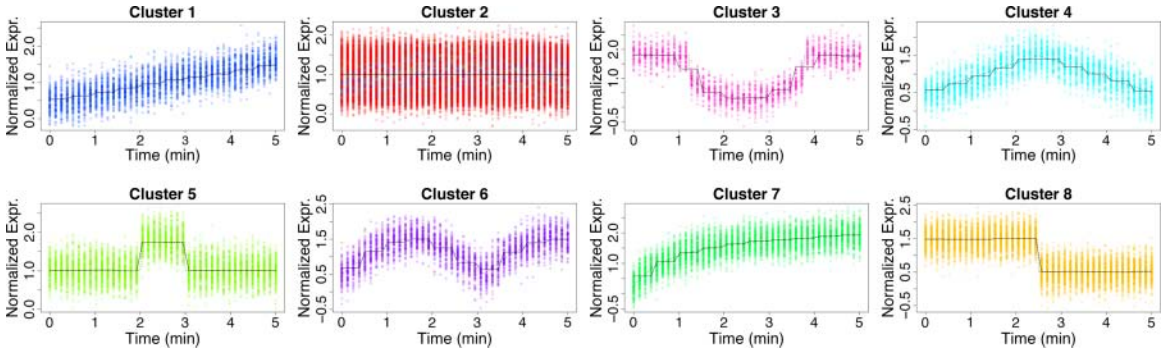

FIG. 5. Mean curves for each of the eight clusters identified by AIC and BIC under two levels of wavelet smoothing ($r = 2$) are individually graphed together with the time-course gene expressions for genes with greater than a 90% probability of belonging to the cluster. Gene expression profiles are colored according to their original cluster allowing inappropriately clustered genes to be easily identified. All of the clusters were correctly identified and true mean curves are graphed along with wavelet-estimated mean curves.

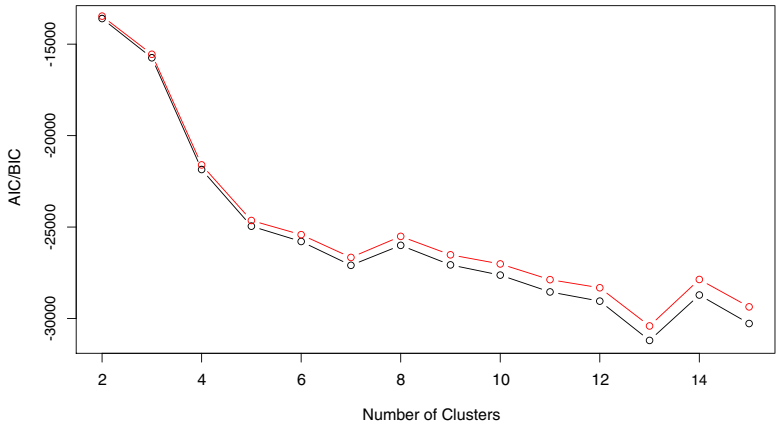

FIG. 6. AIC (black) and BIC (red) values for functional clustering under one level of wavelet smoothing on the time course gene expression data considered in Rustici et al. (2004) indicates the optimal number of clusters to be thirteen.

This dataset was also analyzed in Ning et al. (2010) utilizing a periodic Fourier series approximation in the mixture model to cluster the gene expression profiles. In this application, the AIC and BIC criteria selected nine distinct profiles, and these clusters are depicted in Figure 9. To quantitatively compare the 13 "wavelet clusters" with the 9 "Fourier clusters," the proportion of overlap index was used. For each method, a gene was classified to a given cluster if it had greater than 90% posterior probability of belonging to the cluster. Given two clusters, the number of genes that were classified to both clusters forms the numerator of the index, and the total number of between the two clusters forms the denominator of the index. Therefore if two clusters perfectly match, the proportion of overlap is one, and if two clusters are disjoint, the proportion of overlap is zero. Some of the clusters matched up fairly closely, and matches and partial matches are provided in Table 3.

The computational requirements were not enormous. The computations were performed on a single desktop with a 4GHz (overclocked) Intel i7 quad-core processor. The R package multicore (Urbanek, 2009) was utilized to fully utilize the multiple cores. The source code to perform the real data analysis with corresponding dataset is available online at http://statgen.psu.edu.

4. DISCUSSION

The studies of gene expression profiles in time course can help to understand the developmental machinery of gene regulation related to a biological process. A considerable body of literature has been

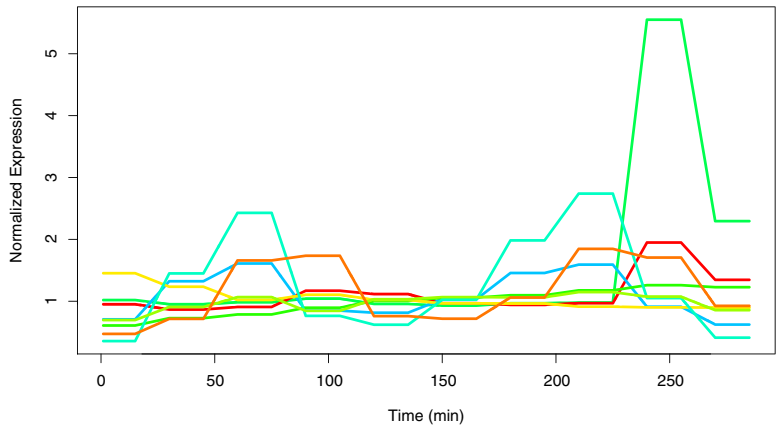

FIG. 7. The thirteen estimated mean curves are graphed together.

TABLE 2. THE NUMBER OF GENES CLASSIFIED TO EACH OF THE THIRTEEN CLUSTERS IS PRESENTED HERE

Cluster 1	*Cluster 2*	*Cluster 3*	*Cluster 4*	*Cluster 5*	*Cluster 6*	*Cluster 7*	*Cluster 8*	*Cluster 9*	*Cluster 10*	*Cluster 11*	*Cluster 12*	*Cluster 13*	*Unclassified*
4	12	22	53	105	7	144	10	6	301	146	158	447	1540

available on statistical methods for characterizing different patterns of gene expression (Qian et al., 2001; Holter et al., 2001; Zhao et al., 2001; Park et al., 2003; Bar-Joseph et al., 2003; Luan and Li, 2003; Ernst et al., 2005; Storey et al., 2005; Ma et al., 2006; Ng et al., 2006; Inoue et al., 2007). When gene expression is measured at a long series of time points, it is possible that response data are contaminated by noises and, thus, the detection of patterns suffers from the so-called "curse of dimensionality." As an increasingly popular means for data compression and de-noising in the context of signal and image processing (Donoho and Johnstone, 1994; Johnstone and Silverman, 1997), wavelet shrinkage has been used here to catalogue gene expression dynamics. This wavelet-based model projects higher dimensional data to a manageable lower dimensional subspace.

We have implemented the idea of wavelet dimension reduction into the mixture model for gene clustering, aimed to de-noise the data by transforming an inherently high-dimensional biological problem to its tractable low-dimensional representation. As a first attempt of its kind, we capitalize on the simplest Haar wavelet shrinkage technique to break an original signal down into spectrum by taking its averages and differences and, subsequently, to detect gene clusters that differ in the smooth coefficients extracting from noisy time series gene expression data. The wavelet thresholding approach that we utilized in this manuscript was constructed for equally spaced longitudinal data, however its extension to non-equally spaced data can be made possible through the development of second-generation wavelets (Pensky and Vidakovic, 2001; Jansen, 2003; Vanraes et al., 2002).

It is noted that the clusters identified with this method are meant to be exploratory. Identification of these clusters can help frame and guide biological investigations. Once data analysis has proceeded to the end

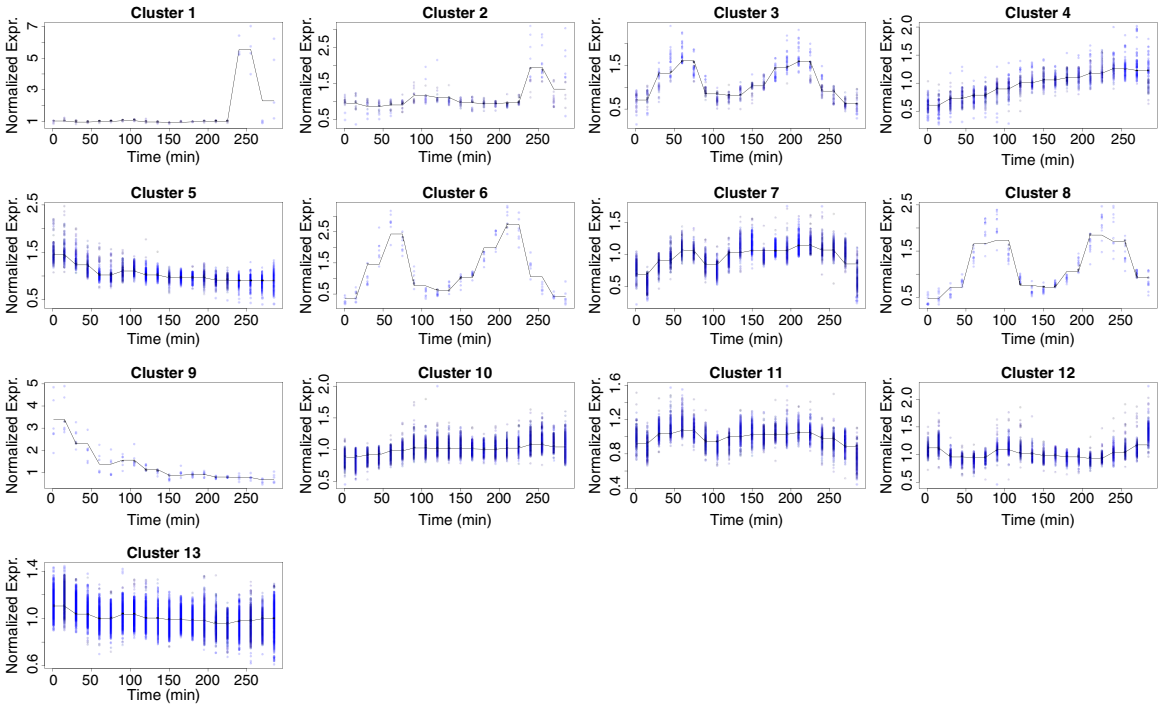

FIG. 8. Mean curves for each of the thirteen clusters identified are individually graphed along with individuals gene expressions for genes with at least 90% probability of belonging to such a cluster.

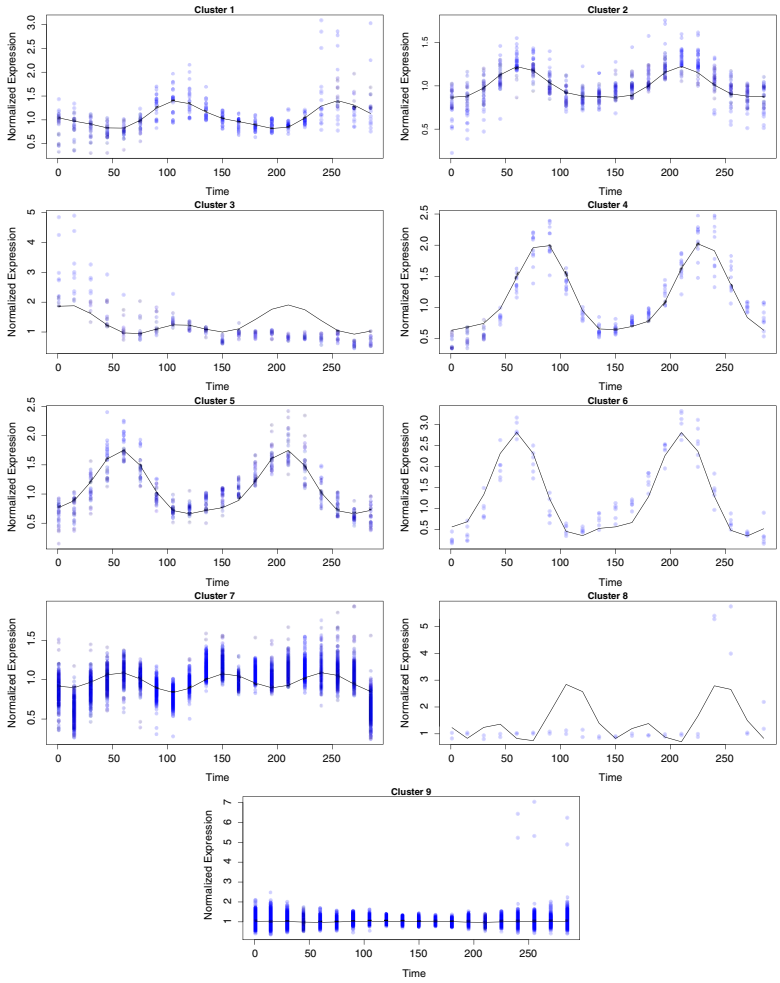

FIG. 9. Mean curves for each of the nine clusters identified in Ning et al. (2010), which were produced by embedding a periodic Fourier series approximation into the mean expressions of the mixture components.

point of our investigations, we are at a point where we can discuss mechanisms with biologists and hopefully work with them to understand the physical mechanisms.

This wavelet-based model will have many implications for addressing biologically meaningful hypotheses at the interplay between gene actions (or interactions) and developmental pathways in various complex biological processes or networks. Although our main application in this article is with time-course gene expression data, the techniques we have developed are generally applicable to other time-course

TABLE 3. THIS TABLE COMPARES OUR CLUSTERING RESULTS THAT UTILIZE WAVELETS WITH THE RESULTS PUBLISHED IN NING ET AL. (2010)

Wavelet cluster no.	*Fourier cluster no.*	*Proportion of overlap*
Cluster 3	Cluster 5	.870
Cluster 6	Cluster 6	.857
Cluster 9	Cluster 3	.600
Cluster 7	Cluster 7	.467
Cluster 2	Cluster 1	.348
Cluster 13	Cluster 9	.203
Cluster 10	Cluster 9	.109

A wavelet cluster is highly matches with a fourier cluster if it has a large proportion of overlap.

datasets including applications to financial data where high-dimensional characteristics are ubiquitous. We hope that our method described within can provide a starting point for further exploration in the functional clustering of high-dimensional data.

ACKNOWLEDGMENTS

This work was partially supported by NSF/NIH (joint grant DMS/NIGMS-0540745).

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Aboufadel, E., and Schlicker, S. 1999. *Discovering wavelets.* Wiley.

Bar-Joseph, Z., Gerber, G.K., Gifford, D.K., et al. 2003. Continuous representations of time-series gene expression data. *J. Comput. Biol.* 10, 341–356.

De Lichtenberg, U., Jensen, L.J., Fausboll, A., et al. 2005. Comparison of computational methods for the identification of cell cycle-regulated genes. *Bioinformatics* 21, 1164.

Diggle, P., Heagerty, P., Liang, K.Y., et al. 2002. *Analysis of Longitudinal Data.* Oxford University Press, New York.

Donoho, D.L. 1995. De-noising by soft-thresholding. *IEEE Trans. Inform. Theory* 41, 613–627.

Donoho, D.L., and Johnstone, J.M. 1994. Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 81, 425.

Eisen, M.B., Spellman, P.T., Brown, P.O., et al. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95, 14863.

Ernst, J., Nau, G.J., and Bar-Joseph, Z. 2005. Clustering short time series gene expression data. *Bioinformatics* 21, 159.

Ghael, S.P., Sayeed, A.M., and Baraniuk, R.G. 1997. Improved wavelet denoising via empirical wiener filtering. *Proc. SPIE* 3169, 389–399.

Ghosh, D., and Chinnaiyan, A.M. 2002. Mixture modelling of gene expression data from microarray experiments. *Bioinformatics* 18, 275.

Holter, N.S., Maritan, A., Cieplak, M., et al. 2001. Dynamic modeling of gene expression data. *Proc. Natl. Acad. Sci. USA* 98, 1693.

Inoue, L.Y.T., Neira, M., Nelson, C., et al. 2007. Cluster-based network model for time-course gene expression data. *Biostatistics* 8, 507–525.

Jansen, M. 2003. Wavelet thresholding on non-equispaced data. *Nonlinear Estimation Classification* 261.

Jensen, A., and la Cour-Harbo, A. 2001. *Ripples in Mathematics: The Discrete Wavelet Transform.* Springer Verlag, New York.

Johnstone, I.M., and Silverman, B.W. 1997. Wavelet threshold estimators for data with correlated noise. *J. R. Statist. Soc. Ser. B* 59, 319–351.

Kim, B.R., Zhang, L., Berg, A., et al. 2008. A computational approach to the functional clustering of periodic gene expression profiles. *Genetics* 180, 821–834.

Lasser, R. 1996. *Introduction to Fourier Series.* CRC, Boca Raton, FL.

Luan, Y., and Li, H. 2003. Clustering of time-course gene expression data using a mixed-effects model with b-splines. *Bioinformatics* 19, 474.

Ma, P., Castillo-Davis, C.I., Zhong, W., et al. 2006. A data-driven clustering method for time course gene expression data. *Nucleic Acids Res.* 34, 1261.

Mallat, S.G., et al. 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 11, 674–693.

McLachlan, G.J., Bean, R.W., and Peel, D. 2002. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* 18, 413.

Ng, S.K., McLachlan, G.J., Wang, K., et al. 2006. A mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics* 22, 1745.

Ning, L., McMurry, T., Berg, A., et al. 2010. Functional clustering of periodic transcriptional profiles through arma(p,q). *PLoS ONE* (in press).

Park, T., Yi, S.G., Lee, S., et al. 2003. Statistical tests for identifying differentially expressed genes in time-course microarray experiments. *Bioinformatics* 19, 694.

Pensky, M., and Vidakovic, B. 2001. On non-equally spaced wavelet regression. *Ann. Instit. Statist. Math.* 53, 681–690.

Qian, J., Stenger, B., Wilson, C.A., et al. 2001. Partslist: a web-based system for dynamically ranking protein folds based on disparate attributes, including whole-genome expression and interaction information. *Nucleic Acids Res.* 29, 1750.

Ramoni, M.F., Sebastiani, P., and Kohane, I.S. 2002. Cluster analysis of gene expression dynamics. *Proc. Natl. Acad. Sci. USA* 99, 9121.

Rustici, G., Mata, J., Kivinen, K., et al. 2004. Periodic gene expression program of the fission yeast cell cycle. *Nat. Genet.* 36, 809–817.

Spellman, P.T., Sherlock, G., Zhang, M.Q., et al. 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. cell* 9, 3273.

Storey, J.D., Xiao, W., Leek, J.T., et al. 2005. Significance analysis of time course microarray experiments. *Proc. Natl. Acad. Sci. USA* 102, 12837.

Urbanek, S. 2009. *multicore: Parallel processing of R code on machines with multiple cores or CPUs*. R package version 0.1–3. Available at: www.rforge.net/multicore/. Accessed June 1, 2010.

Vanraes, E., Jansen, M., and Bultheel, A. 2002. Stabilised wavelet transforms for non-equispaced data smoothing. *Signal Process.* 82, 1979–1990.

Vidakovic, B. 1999. *Statistical Modeling by Wavelets.* Wiley, New York.

Walker, J.S. 1999. *A Primer on Wavelets and Their Scientific Applications.* CRC Press, Boca Raton, FL.

Wang, L., Chen, X., Wolfinger, R.D., et al. 2009. A unified mixed effects model for gene set analysis of time course microarray experiments. *Statist. Appl. Genet. Mol. Biol.* 8, 47.

Zapala, M.A., and Schork, N.J. 2006. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proc. Natl. Acad. Sci. USA* 103, 19430.

Zhao, L.P., Prentice, R., and Breeden, L. 2001. Statistical modeling of large microarray data sets to identify stimulus-response profiles. *Proc. Natl. Acad. Sci. USA* 98, 5631.

Address correspondence to:
Dr. Arthur Berg
Department of Biostatistics
Pennsylvania State University
500 University Drive, Mail Code CH69
Hershey, PA 17033

E-mail: berg@psu.edu