OMICS A Journal of Integrative Biology
Volume 14, Number 3, 2010

DOI: 10.1089/omi.2010.0005

A New Symbolic Representation for the Identification of Informative Genes in Replicated Microarray Experiments

Jeremy D. Scheff,[1] Richard R. Almon,[2,3] Debra C. DuBois,[2,3] William J. Jusko,[3] and Ioannis P. Androulakis[1,4,5]

Abstract

Microarray experiments generate massive amounts of data, necessitating innovative algorithms to distinguish biologically relevant information from noise. Because the variability of gene expression data is an important factor in determining which genes are differentially expressed, analysis techniques that take into account repeated measurements are critically important. Additionally, the selection of informative genes is typically done by searching for the individual genes that vary the most across conditions. Yet because genes tend to act in groups rather than individually, it may be possible to glean more information from the data by searching specifically for concerted behavior in a set of genes. Applying a symbolic transformation to the gene expression data allows the detection overrepresented patterns in the data, in contrast to looking only for genes that exhibit maximal differential expression. These challenges are approached by introducing an algorithm based on a new symbolic representation that searches for concerted gene expression patterns; furthermore, the symbolic representation takes into account the variance in multiple replicates and can be applied to long time series data. The proposed algorithm's ability to discover biologically relevant signals in gene expression data is exhibited by applying it to three datasets that measure gene expression in the rat liver.

Introduction

In recent years, microarrays have become indispensable tools in molecular biology because of their ability to quantitatively measure the expression of thousands of genes at the same time (Brown and Botstein, 1999). Their prominence and utility has grown with the increased computational power available to researchers and the development of the field of bioinformatics. The analysis of gene expression data from microarray experiments is a dynamic field; diversity in experimental design and in statistical methods has produced myriad different computational algorithms to make sense of the extremely high dimensional data. As microarray technology becomes more common and cost-effective (Bryant et al., 2004), experiments have gotten larger. This means not only that sampling is done at higher frequencies and for longer periods of time; but also, researchers are including more repeated measurements in their experiments to enhance the reliability of their results. These trends necessitate new innovations from the computational domain so that we can fully exploit the added information from more thorough experiments. To meet this challenge, existing algorithms need to be modified and new approaches must be developed.

The primary goal in the analysis of gene expression data is separating biologically relevant signals from the underlying biological and experimental noise inherent in microarray experiments. Multiple biological replicates are necessary to produce reproducible, statistically significant results in microarray experiments (Churchill, 2002). Although averaging together multiple measurements does greatly improve accuracy relative to using just a single measurement (Lee et al., 2000), all information about the variance in the replicates is lost in the averaging. With this in mind, many methods have been proposed for analyzing gene expression data, typically by assigning each gene a score and setting a cutoff at an acceptable error rate (Androulakis et al., 2007; Storey and Tibshirani, 2003; Tusher, 2001). However, these types of methods do not account for the fact that genes do not act as independent features; rather, their behaviors are often highly correlated (Storey et al., 2007; Wolfe et al., 2005). Incorporating this knowledge into the analysis of gene expression data may lead to more biologically relevant insights. Thus, clustering methods are commonly applied when studying gene expression data.

[1]Biomedical Engineering Department, Rutgers University, Piscataway, New Jersey.
[2]Department of Biological Sciences, State University of New York at Buffalo, Buffalo, New York.
[3]Department of Pharmaceutical Sciences, State University of New York at Buffalo, Buffalo, New York.
[4]Chemical and Biochemical Engineering Department, Rutgers University, Piscataway, New Jersey.
[5]Department of Surgery, UMDNJ–Robert Wood Johnson Medical School, New Brunswick, New Jersey.

Traditional clustering methods, such as k-means or hierarchical clustering (Eisen et al., 1998), can be used on gene expression data. But for time course data, they are not ideal because they ignore the sequential nature of the data collection (Ernst et al., 2005). For this reason, there is interest in methods of searching for temporal patterns in the data; this type of analysis is particularly well suited for time course gene expression data because it searches for groups of genes with similar dynamics over time, which are likely biologically relevant.

Analysis on the level of expression patterns rather than individual genes can be accomplished by assigning the genes into a large yet finite number of categories, depending on the gene's trajectory; this process is called discretization, and the result is a symbolic representation of each gene. A symbolic representation is desirable because it allows the discovery of patterns in the data (genes with the same or similar symbolic representations) instead of limiting the analysis to looking only at differential expression in individual genes. By transforming each point in the time series into a discrete symbol, the statistical analysis becomes more straightforward. Other advantages of symbolic representations include noise reduction and computational efficiency.

The discretization of time series data has been thoroughly discussed in the literature, with applications in virtually all fields of science and engineering (Daw et al., 2003). The procedure generally consists of setting a number of cutoffs and assigning different symbols to values falling in different partitions. The symbols can then be temporally ordered, resulting in a sequence of symbols. Alternatively, one symbol can be used to represent more than one time point, further temporally discretizing the data.

A popular example of a symbolic representation is the Symbolic Aggregate approXimation (SAX) (Lin et al., 2007). SAX has been applied to gene expression data through SLINGSHOTS (Yang et al., 2007), which selects informative motifs from gene expression data based on the symbolic representation. However, because of preprocessing steps required before the symbolic transformation, it does not take into account the magnitude of change in gene expression and the variance in multiple replicates. These limitations are important because gene expression data is notoriously noisy, particularly at low expression values.

When studying time course gene expression data, it is natural to desire a symbolic representation that has only three possible symbols, reflecting the most intuitive possible responses of genes to stimuli: upregulation, downregulation, and no regulation. Unlike SLINGSHOTS , Trajectory Clustering (Phang et al., 2003) transforms time series microarray data into a symbolic representation that takes into account multiple replicates and the magnitude of gene expression changes. However, it can only be applied to short time series data (five time points or less) before the number of clusters explodes exponentially and becomes unmanageable.

Symbolic discretization is similar to the idea behind Short Time-series Expression Miner (STEM) (Ernst and Bar-Joseph, 2006), which groups genes into a predefined set of clusters and selects the informative clusters based on their p-values. However, STEM was designed to only function on short time series (approximately eight time points or less), whereas all of the datasets considered in this article contain between 11 and 18 time points.

Bayesian approaches have also been proposed toward significance testing in gene expression data (Angelini et al., 2007). In the context of clustering, Bayesian clustering of curves approaches gene expression data with the goal of searching for patterns in the data rather than searching for individual genes, similar to symbolic methods. In Heard et al. (2006), this technique is applied to gene expression data to find underlying patterns in the data. However, other than discarding large clusters that do not vary greatly, they do not extend their method to quantitatively determining which clusters are most significant, which is of importance for studying biological data. Furthermore, the discrete nature of symbolic representations is appealing because of the high levels of noise inherent in gene expression data (Androulakis et al., 2007).

To overcome these issues with existing methods, this article proposes a new symbolic representation that takes advantage of all of the available experimental data, rather than averaging measurements together; this symbolic representation is presented in conjunction with a procedure to identify statistically significant patterns in the data. This new symbolic representation is simple, intuitive, and effective. The method's ability to discover biologically relevant signals is illustrated by running it on three different datasets, all of which are from time course experiments measuring gene expression in the rat liver. Two datasets concerning corticosteroid treatment are considered: one follows the response to an acute dosage (Jin et al., 2003), whereas the other dataset contains the response to a constant drug infusion (Almon et al., 2007). The third dataset explores the normal circadian rhythm in rats that are exposed only to regular light/dark cycling (Almon et al., 2008).

Materials and Methods

Clustering and selection algorithm

The proposed algorithm consists of a very fine-grained clustering on the data followed by the selection of statistically significant clusters, as shown in Figure 1. This is accomplished by first filtering the data with a permissive ANOVA (p-value = 0.05) so that further steps are only con-

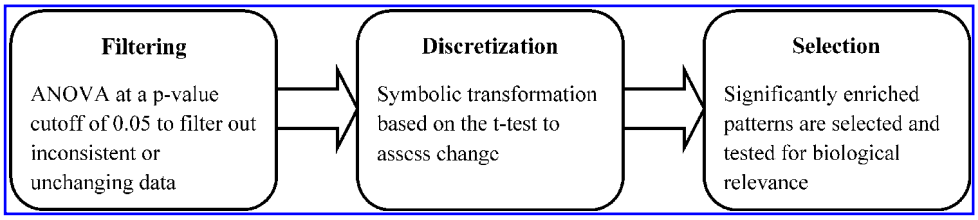

FIG. 1. Flowchart showing the steps involved in the proposed algorithm.

sidering varying data with relatively consistent measurements. Then, the time course expression values for each probe set are discretized into a sequence of symbols, with each unique sequence representing a different motif. The motifs that represent significantly more probe sets than expected by random chance are selected and called informative.

Symbolic representation

A critically important component of the proposed algorithm is the initial step of transforming the data into a symbolic representation. A symbolic transformation of time course data is generally performed by breaking up the domain into regions and assigning a different symbol to points that are in different regions. The data points that are used to represent one symbol are called a *word*, the range of possible symbols is known as the *alphabet*, a complete sequence of symbols for a feature (probe set) is called a *motif*, and the number of probe sets represented by each distinct motif is that motif's *population*.

In SLINGSHOTS , the original symbolic transformation is based on the SAX algorithm (Lin et al., 2007). Briefly, it is performed by averaging the replicates together; normalizing those average values so each probe set has the same standard deviation; setting breakpoints such that each symbol has an equal probability of being selected for random data; and ordering the symbols temporally to create motifs (Yang et al., 2007, 2008). This symbolic representation is not ideal for gene expression data because when the replicates are averaged, all knowledge about the accuracy of those measurements is lost. Furthermore, when the expression values are normalized, the magnitude of the differential gene expression is ignored, potentially amplifying small changes in gene expression. Ideally, all of this information should be retained and used to create more refined motifs.

The ultimate goal of a symbolic representation is to assign different symbols to points with very different values. The variance in the repeated measurements is a key factor in determining how different two sets of measurements really are, so it makes sense that an ideal symbolic representation would take this information into account. Our proposed symbolic representation accomplishes this while retaining the simplicity of the original method.

Each dataset contains measurements of thousands of genes at several different time points with several repeated measurements. Thus, consider $g_{i,j}$ as a vector that contains the expression values for all of the replicates of probe set i at time point j; then, the length of the vector $g_{i,j}$ is the number of repeated measurements. The proposed symbolic representation transforms each adjacent pair of $g_{i,j}$ and $g_{i,j+1}$ into a discrete symbol $s_{i,[j,j+1]}$, where $[j, j+1]$ represents the time interval between two adjacent points j and $j+1$.

To achieve this symbolic transformation, for each probe set a t-test is taken between the replicates at all adjacent time points. This is possible because gene expression data is very close to log-normal (Hoyle et al., 2002), so taking the log of the data before applying the t-test is sufficient to obtain valid results. The t-tests give t-values that correlate with the direction of differential gene expression, as shown in Figure 2; for instance, an increase in gene expression across two time points results in a positive t-value, whereas a decrease results in a negative t-value.

These t-values are discretized into equiprobable symbols by setting appropriate breakpoints. These breakpoints can be determined by using a table of the CDF for the t-distribution with the appropriate degrees of freedom. Or, more generally, the breakpoints can be found by solving the following equation for b, the breakpoint values, where $n = 1,2,\ldots,a-1$ is the breakpoint number, a is the alphabet size, Γ is the gamma function, ${}_2F_1$ is the hypergeometric function, and ν is the number of degrees of freedom, which is a function of how many repeated measurements there are.

$$\frac{n}{a} = \frac{1}{2} + b\Gamma\left(\frac{\nu+1}{2}\right)\frac{{}_2F_1\left(\frac{1}{2}, \frac{\nu+1}{2}; \frac{3}{2}; \frac{b^2}{\nu}\right)}{\sqrt{\pi\nu}\Gamma(\nu/2)}$$

As with the original method, when random data is used, all sequences of symbols are equiprobable. The symbols $s_{i,[j,j+1]}$

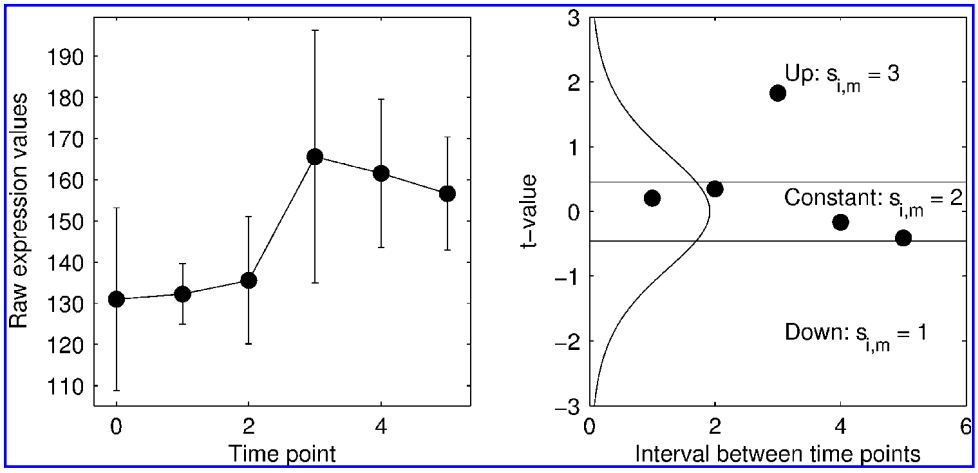

FIG. 2. Symbolic representation of a gene. On the left are time course gene expression values. The error bars show the standard deviations of repeated measurements. The t-tests are taken between adjacent time points, resulting in the t-values shown on the right. The curve is the t-distribution, which is used to discretize the t-values into three equiprobable symbols labeled Up, Constant, and Down. Note that the six time points are represented by only five symbols because there is one symbol for each interval between adjacent time points.

are defined such that 1 is the lowest symbol, and each subsequent higher symbol is increased by 1.

The proposed symbolic representation depends on two parameters, w and a. w, the word size, defines how many adjacent time points to pool together when taking a t-test. Using a word size of anything besides 1 results in a loss of resolution, so it should only be increased when motifs are sparsely populated. a is the alphabet size, which is a small integer defining how many different symbols to use (or, how many breakpoints to set). The selection of optimal parameter values is discussed in the following section. After the symbolic transformation is complete, each motif can be hashed to a unique integer by effectively converting it from base a to base 10 (Lin et al., 2002):

$$h_i = \sum_{j=1}^{T-1} [S_{i,[j,j+1]} - 1]a^{T-j-1}$$

In this equation, h_i is the hash value for a probe set i, T is the total number of time points, $s_{i,[j,j+1]}$ is the symbol for a probe set i at interval m, and a is the alphabet size. Then, each unique hash value corresponds to a distinct motif.

The symbolic transformation and hashing for $a = 3$ and $w = 1$ is presented below in Matlab pseudocode:

```
for i = 1:N % Iterate through probe sets
  h(i) = 0; % Initialize the hash value

  for j = 1:T-1 % Iterate through time points
    t_value = result of t-test between vectors
              g(i,j) and g(i,j+1)

    % Discretize the t-value (s(i,j) is
       s_{i,[j,j+1]})
    if t < first breakpoint
      s(i,j) = 1;
    elseif t < second breakpoint
      s(i,j) = 2;
    else
      s(i,j) = 3;
    end
    h(i) = h(i) + (s(i,j)-1) * a^(T-j-1);
          % Update hash with each symbol
  end
end
```

After the symbolic transformation has been performed, the final step in the analysis is to determine which patterns in the data are overrepresented. This is done by selecting motifs that are significantly enriched in population.

Selection of informative motifs

The goal is to discover overrepresented patterns in the data, because those likely correspond with concerted changes in the expression of related genes. Because multiple genes are typically regulated together, the motifs with the most informative genes should be more highly populated than the motifs found in random data. For both of the symbolic representations described here, random data results in a uniform distribution of motif populations because each motif is equiprobable. This fact is used to determine which motifs are significantly larger than what we would expect by chance.

Random data is generated by randomly resampling the original gene expression data. This random resampling is repeated 1,000 times to ensure that it is an accurate approximation of the null distribution. The distribution of motif populations is fit to an exponential distribution and the size cutoff is determined by finding the corresponding motif size for a given p-value. This is done at a p-value of 0.01 for all of the results shown in this article. Typically, applying this cutoff to the real data gives several large motifs.

The datasets that were analyzed in this study are all long time series; the shortest is the chronic corticosteroid data with 11 time points. Because of this, some of the selected motifs that differ only at one or two symbols may actually be representing the same biological phenomena. To compensate for this possibility, the motifs are combined to maximize homogeneity (H) and separation (S). This process combines only the most similar motifs that have the same qualitative profiles.

$$H(M_i) = \frac{1}{\binom{|M_i|}{2}} \sum_{x,y \in M_i} sim(x,y)$$

$$S(M_i, M_j) = \frac{1}{|M_i||M_j|} \sum_{x \in M_i,\, y \in M_j} 1 - sim(x,y)$$

In these formulas, each unique motif is represented by an M_i, which is the set of genes with that motif; the term $|M_i|$ gives the number of genes in the motif M_i. The similarity matrix sim(x, y) is calculated by finding the Pearson correlation coefficient between the genes x and y. The homogeneity is calculated for each motif and the separation is calculated for each combination of two motif. Then, each possible combination of two clusters is exhaustively evaluated to find the one combination that maximizes the sum of homogeneity and separation. These two clusters are combined, and the process is repeated until the homogeneity and separation is maximized. After this procedure, the combined significant motifs are called clusters.

To optimize the accuracy of the results, the word size w should be minimized. This is because increasing w simply pools adjacent points together, effectively smoothing the signal and resulting in a decreased temporal resolution. However, finding the most descriptive symbolic representation for a gene may require a word size greater than 1. Therefore, in all cases, the goal is to find the minimum word size that produces significantly populated motifs as defined above. For the acute and chronic datasets, this procedure resulted in a word size of 2; for the circadian data, a word size of 3 was required. This is likely caused by two factors: the circadian data has more time points than the other datasets, and it is just observing the normal circadian rhythms rather than some specific powerful stimulus that produces large, coordinated responses.

An alphabet size (a) of 3 was used in all of the analysis performed in this article. There is no technical limitation forcing this choice of a, but the symbolic representation functions in the space of transitions, not in the space of raw data. In other words, the symbolic representation is searching for significant changes in gene expression between time points, so an alphabet size of 3 is selected to represent the three broad classes of potential gene expression patterns: upregulation, downregulation, and no regulation. Therefore, in general an alphabet size of 3 should be used.

Synthetic data

It is difficult to quantitatively assess the differences between various gene expression clustering algorithms using real data because of the high level of noise inherent in microarray data. Compounding this problem is the fact that it is typically unknown what the "correct" clusters are, or if they even exist. For these reasons, it is essential to test algorithms on synthetic data, in addition to real data, so some objective comparisons can be made.

Synthetic data is used to show that the proposed symbolic representation performs better than the original SAX-based symbolic representation. Because the ultimate goal of gene expression analysis is to find the underlying signals and patterns hidden in noise, the performance of the two symbolic representations is assessed as the noise level of some synthetic data is varied. To do this, synthetic data is generated with and without noise. For both the new and original symbolic representation, the symbolic transformation is applied to both the clean and noisy data. Then, the adjusted Rand index (Hubert and Arabie, 1985) is computed to assess how effective each symbolic representation is at determining the true temporal patterns of expression in the noisy data. For various noise levels, this test is repeated 1,000 times on different sets of random synthetic data so that the adjusted Rand index converges.

The synthetic data used in this study is similar to the data generated in other studies on gene expression clustering (Yeung et al., 2003; Yao et al., 2008). Six clusters are created, each with 66 genes, for a total of $n = 396$ synthetic genes. Four of the clusters are sinusoidal and two are linear. The synthetic experiment is comprised of five time points with four repeated measurements at each time point. The subscripts i, j, k, and m represent the gene number, the time point, the replicate number, and the cluster number, respectively. Then, the sinusoidal and linear clusters are defined by the following two equations, respectively:

$$g_{i,j,k} = \sin\left(\frac{j\omega_m}{n} + \phi_m\right) + \alpha\sigma_i\hat{\sigma}_j x_{i,j,k}$$

$$g_{i,j,k} = \pm\frac{j}{n} + \alpha\sigma_i\hat{\sigma}_j x_{i,j,k}$$

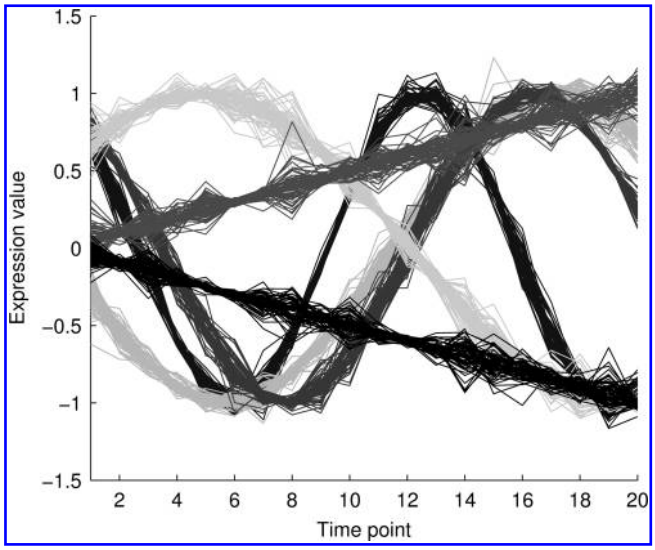

FIG. 3. Synthetic data, generated with an error level of 0.2.

In these equations, $g_{i,j,k}$ is an expression value for a specific gene, time point, and replicate; ω_m is the wavelength; φ_m is the phase; α is the noise level; σ_i is the error for the ith gene; and $\hat{\sigma}_j$ is the error for the jth time point.

As in Yao et al. (2008), the random errors σ_i and $\hat{\sigma}_j$ are drawn from uniform random distributions on [0.2, 1.2]; $x_{i,j,k}$ is drawn from an N(0, 1) normal distribution; the parameters ω_m and φ_m are drawn from uniform random distributions on $[\pi/2, 5\pi]$ and $[0, 2\pi]$, respectively; and the noise level α is varied from 0 to 2 to assess how the proposed symbolic representation responds to noise. An example of the synthetic data used in this study is shown in Figure 3. The synthetic data used in this study contains only genes with a true underlying pattern; that is, there are no synthetic genes generated with no specific time dependence. This is done because genes that do not have a coordinated temporal behavior would have very low populated motifs that would not be chosen by the selection algorithm.

Gene expression data

All of the data analyzed in this study is publicly available in the Gene Expression Omnibus (GEO) database (Barrett et al., 2009). The first dataset, available with accession number GDS253 (Jin et al., 2003), measures the transcriptional response of the liver to a bolus dose of 50 mg/kg methylprednisolone (MPL). Forty-three male adrenalectomized Wistar rats were sacrificed at 16 time points: 0.25, 0.5, 0.75, 1, 2, 4, 5, 5.5, 6, 7, 8, 12, 18, 30, 48, and 72 h after dosing; in addition, four more rats were used as a control group (untreated). Isolated RNA from each rat liver was hybridized to Affymetrix Rat Genome U34A microarrays, which measure the expression value of 8,799 probe sets.

The second dataset (GDS972) (Almon et al., 2007) is from a similar study, but instead of an acute dosing of MPL, a constant infusion of 0.3 mg/kg/h is given to 40 male adrenalectomized Wistar rats. Rats are sacrificed over the course of 7 days at 10 time points: 6, 10, 13, 18, 24, 36, 48, 72, 96, and 168 h. Four control rats were treated with saline and sacrificed at 6, 18, 48, and 96 h. A total of 15,923 probe sets were then measured using the Affymetrix Rat Expression 230A microarray platform.

The final dataset, available with accession number GSE8988 (Almon et al., 2008), studies changes in liver gene expression during a normal 24-h circadian cycle. Two groups of 27 normal male Wistar rats were used: one for the light period, and one for the dark period. All rats were subjected to regular 12-h light/dark cycles. Rats were killed at 18 time points, measured from the time lights were turned on: 0.25, 1, 2, 4, 6, 8, 10, 11, 11.75, 12.25, 13, 14, 16, 18, 20, 22, 23, and 23.75 h. As in the previous dataset, Affymetrix Rat Expression 230A microarrays were used.

Results

Synthetic data

Figure 4 shows how the adjusted Rand index for the two symbolic representations changes as the noise level in the synthetic data is varied. A value of 1 indicates a perfect recovery of the underlying structures, whereas a value of 0 is given when performance is no better than random guessing. Both methods degrade in accuracy as the noise level increases,

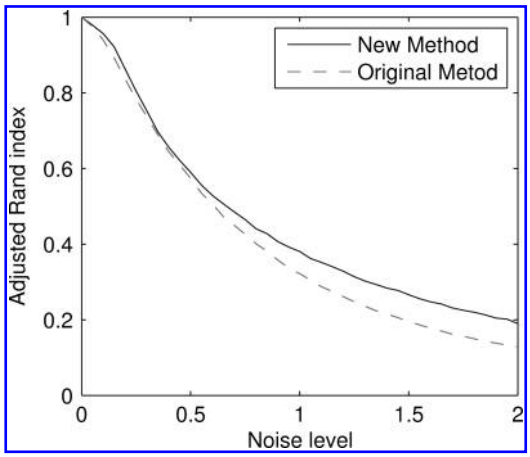

FIG. 4. The adjusted Rand index as noise increases for the two symbolic representations. The adjusted Rand index is calculated 1,000 times for each method at each error level. Error bars for the mean values are not present because they are too small to be shown at this scale.

but the proposed method is more resilient to noise than the previous method. This is important because real microarray data certainly contains some level of noise. The values in Figure 4 are the averages of 1,000 iterations, so the mean values of the adjusted Rand index have converged.

Gene expression data

Table 1 contains the size of the datasets along with the number of motifs and clusters found in each dataset. Figure 5 shows the distribution of motif populations for each of the three datasets. The large dots represent the significantly populated motifs that are selected and combined into the final clusters shown in Figures 6 through 8. The list of genes comprising these clusters, their expression values, and their ontologies are provided as supplementary material at http://rci.rutgers.edu/~yannis/supp/omics_symbolic/clusters.xls.

The acute corticosteroid dataset shows the response to a concentrated perturbation followed by an eventual return to homeostasis. Figure 6 shows the four most informative clusters that were selected by the algorithm. All four clusters exhibit an early deviation followed by an eventual return to baseline conditions. Initially, clusters 1 and 2 are downregulated, whereas clusters 3 and 4 are upregulated.

As in the acute corticosteroid results, there are clusters found in the chronic corticosteroid dataset that contain only an early perturbation followed by an eventual return to baseline. Yet, in this dataset, there is a richer set of responses to the drug. In addition to the early and up- and downregulation, several clusters shown in Figure 7 display a common distinct pattern: early upregulation/downregulation, a brief return toward homeostasis, and finally late upregulation/downregulation as the system is overwhelmed by the chronic drug treatment and reaches its new steady state.

In the analysis of the circadian rhythm data, four clusters are selected and displayed in Figure 8, each with approximately 100 genes. It is important to note that the first 12 h is the light period and the last 12 h is the dark period, as indicated by the shading in Figure 8. The rats are active and feeding during the dark period. All four clusters are approximately sinusoidal with periods of 24 h. The difference between the clusters is that they are all out of phase by 90 degrees, suggesting that they represent four distinct molecular responses to the circadian rhythm.

Discussion

The value of the proposed algorithm is exhibited by assessing its ability to extract biologically relevant patterns from three different long time series gene expression datasets.

The importance of understanding the pharmacokinetic and pharmacogenomic properties of corticosteroids derives from their potent anti-inflammatory and immunosuppressive properties and their potentially harmful side effects (Schimmer et al., 1996). The results presented here are particularly interesting because they facilitate the comparison between acute and chronic dosing of corticosteroids. The four clusters found for the acute corticosteroid dataset are shown in Figure 6. Cluster 1 exhibits a downregulated profile and contains genes with ontologies and pathways highly enriched for metabolic processes. The pathways for the metabolism of several different amino acids are all enriched, in agreement with previous analyses on the effect of corticosteroids (Jin et al., 2003). Cluster 2 has a similar temporal profile, and its genes have similar molecular functions as the genes in cluster 1. The rapid transcriptional and translational response to the corticosteroid treatment is evident in cluster 3. Several regulatory pathways are enriched (aminoacyl-tRNA biosynthesis, proteasome, long-term potentiation, neurodegenerative diseases), and highly enriched ontologies are primarily related to transcription, and translation: translation initiation factor activity, translation regulator activity, nucleotide binding, nucleic acid binding, and GTP binding. These ontologies show the rapid, coordinated, and powerful cellular response to the stimulus. This becomes even more apparent when looking at

Table 1. Datasets

		Number of probe sets		*Number of motifs*		
Dataset	*Time points*	*Before filtering*	*After filtering*	*Total*	*Selected*	*Clusters*
Acute	17	8,799	2,920	1,224	12	4
Chronic	11	15,923	4,361	236	54	12
Circadian	18	15,923	2,468	192	5	4

The third and fourth columns show the number of probe sets before and after ANOVA filtering for each dataset. The next columns show the number of motifs in each dataset and the number of significantly large (selected) motifs. The final column shows the total number of clusters after the significantly large motifs have been combined.

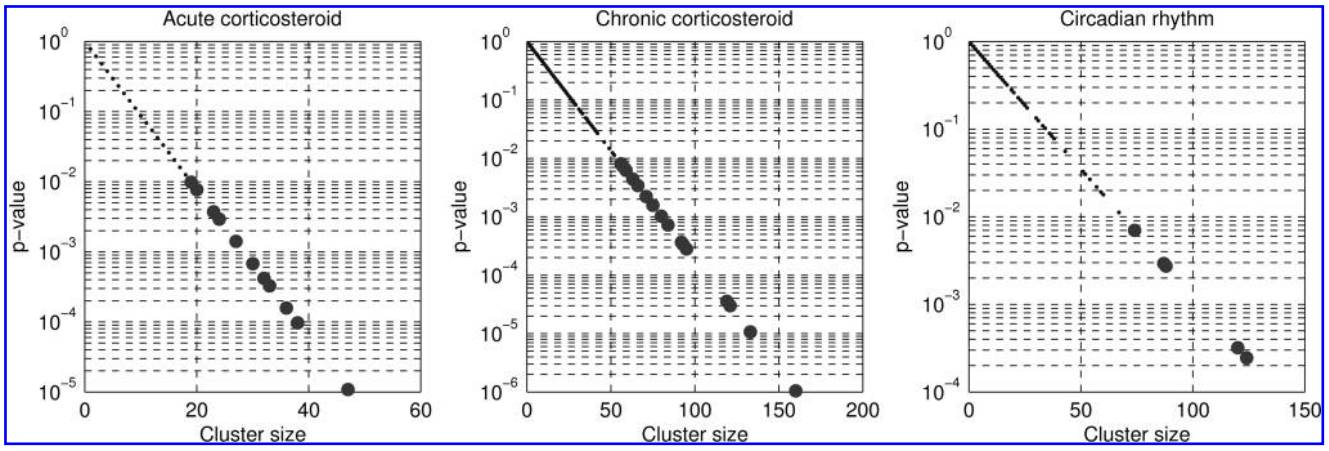

FIG. 5. For each of the three datasets, the significant clusters (large dots) are separated from the other clusters (small dots) by setting a cutoff that corresponds to the cluster size with a *p*-value of 0.01 for random data.

the cluster sizes; cluster 3 contains 265 genes, whereas the next-largest cluster, cluster 1, contains only 55 genes. Two of the genes in cluster 4, *Prmt1* and *Prmt3*, function in post-translational modification of proteins.

Figure 7 shows the clusters found in the chronic corticosteroid data. Clusters 1, 2, 4, 5, and 6 contain genes that function in fatty acid metabolism and other cellular metabolic processes, which have been shown to be downregulated in response to corticosteroid treatment (Das, 2000). Interestingly, there are two distinct patterns for these clusters. Clusters 5 and 6 are initially down-regulated and then return to their original expression values. But in clusters 1, 2, and 4, after the genes are downregulated early, there is a peak around 50 h where the expression returns near baseline before they are ultimately downregulated again throughout the end of the experiment. This shows how some transcriptional responses remain perturbed as long as the corticosteroid infusion persists. These results suggest that the metabolic response to chronic corticosteroid exposure cannot be simply described as "downregulated." Also in clusters 1, 4, 8, 9, and 11 are genes related to the metabolism of several amino acids. Corticosteroids are known to have an effect on amino acid metabolism

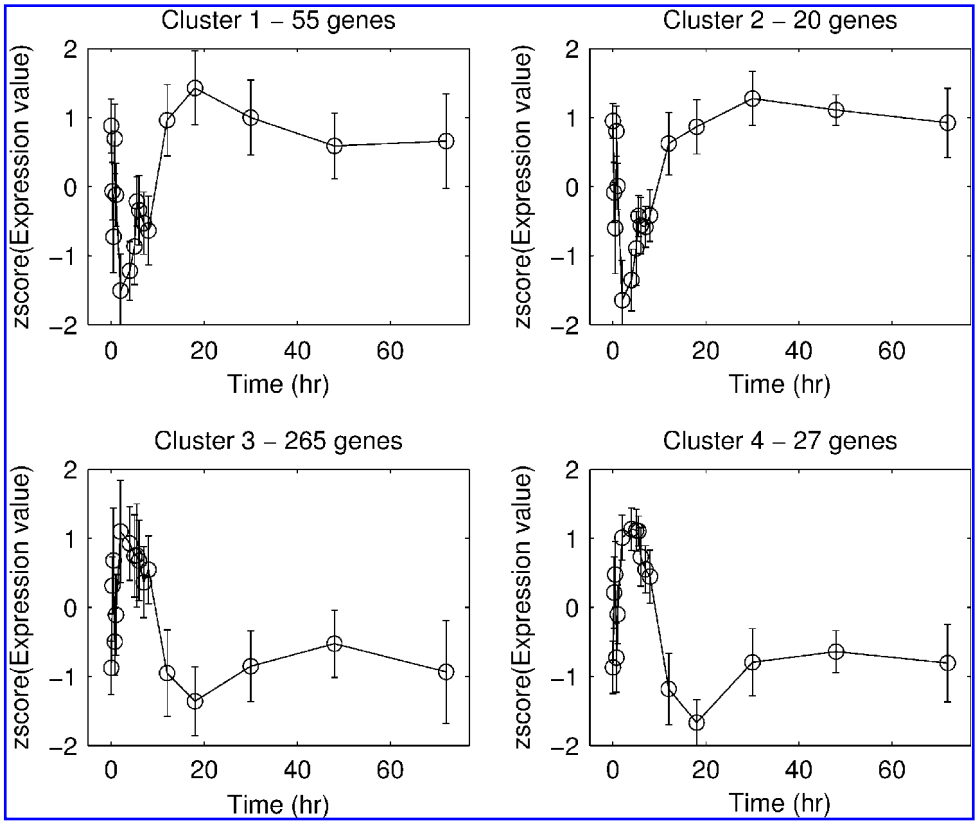

FIG. 6. Acute corticosteroid clusters. Error bars are the standard deviation of the replicates at each time point.

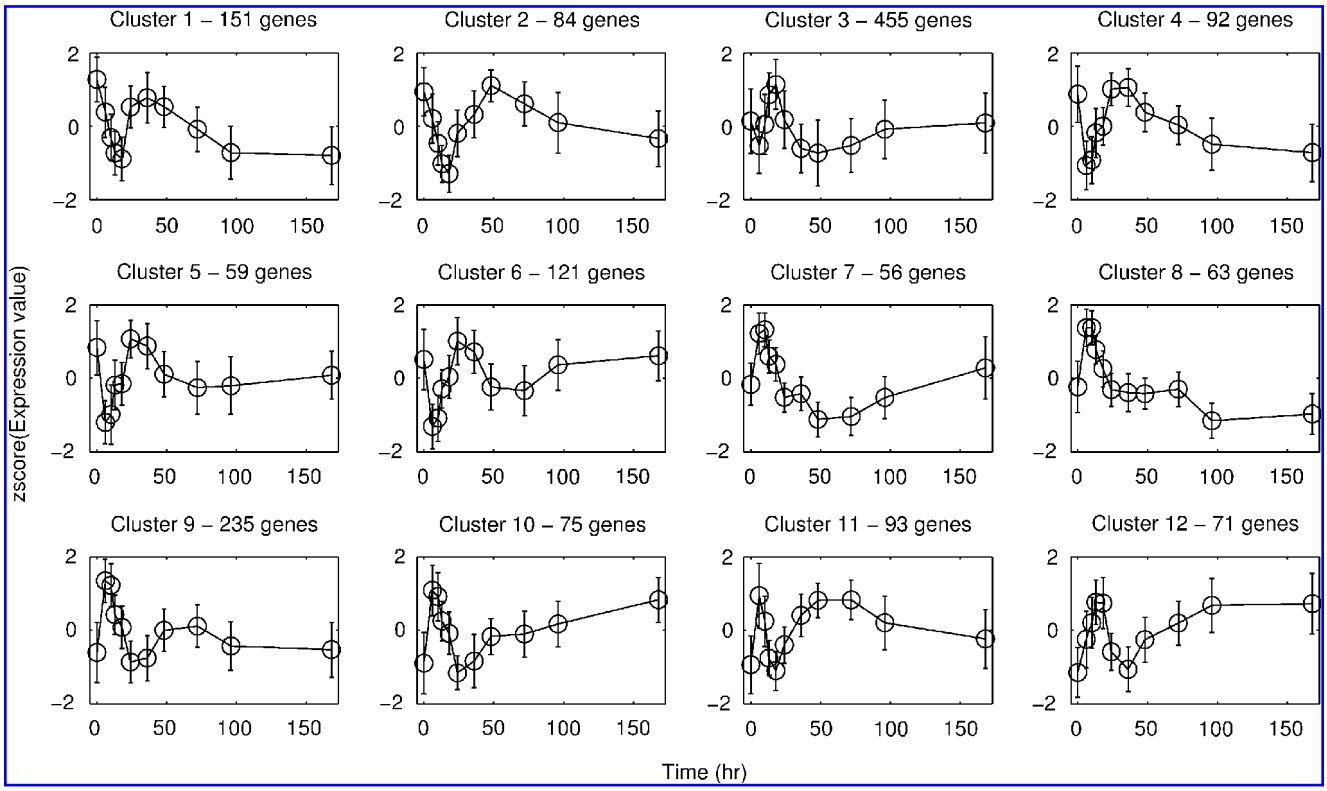

FIG. 7. Chronic corticosteroid clusters. Error bars are the standard deviation of the replicates at each time point.

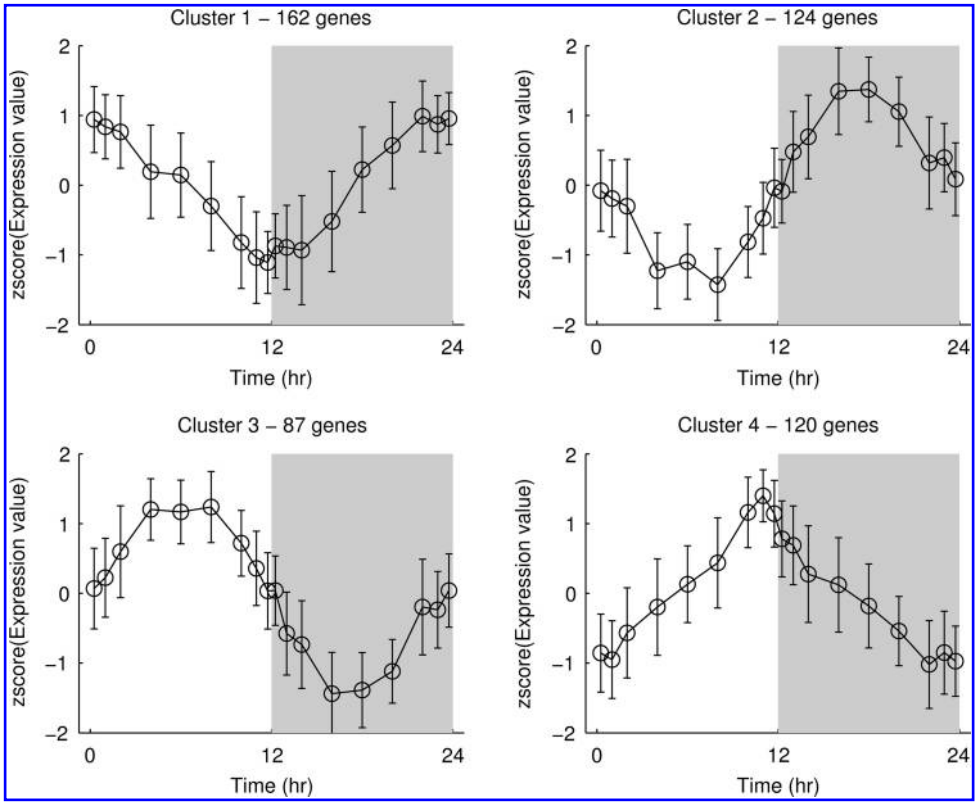

FIG. 8. Circadian rhythm clusters. Error bars are the standard deviation of the replicates at each time point.

(Engelen et al., 2000; May et al., 1996). The body's response to corticosteroids is evident because many of the genes in cluster 1 are annotated with the oxidoreductase ontology. The 11β-hydroxysteroid dehydrogenase enzymes catalyze the interconversion of active cortisol and inactive cortisone, which regulates the concentration of active corticosteroids (Draper et al., 2005). Cluster 3 appears to show a delayed peak several hours after the treatment begins, and then it returns to its original expression. The genes in cluster 3 belong to several fundamental cellular regulatory pathways related to the ribosome, the proteasome, and apoptosis.

Circadian rhythms are the result of internal timing mechanisms that measure periodic cycles of time. In mammals, this is typically a response to the daily pattern of light and dark. For this reason, genes regulated by the circadian rhythm typically have sinusoidal expression profiles with periods of 24 h. Several previous studies of circadian rhythms in gene expression data have been performed by explicitly searching for these sinusoidal patterns in the gene expression data (Almon et al., 2008; Ueda et al., 2005; Yan et al., 2008). The proposed method makes no *a priori* assumptions about the patterns in the data, yet it still finds four sinusoidal clusters with 24- periods and phases evenly spaced throughout the day, as shown in Figure 8.

Cluster 1 contains genes that decrease in expression during the light period and increase in expression during the dark period. *Arntl* (also known as *Mop3* or *Bmal1*) is a member of this cluster, and has been shown to be an essential component of the circadian pacemaker in mice (Bunger et al., 2000). The loss of *Arntl* can lead to a diminished circadian rhythm and can even result in a complete loss of circadian rhythmicity in certain conditions. *Arntl* is also important because it forms a heterodimer with *Clock,* which regulates the transcription of many circadian-controlled genes (Reppert and Weaver, 2001). Several pathways are significantly enriched in this cluster, mainly related to metabolism: glycolysis/gluconeogenesis, polyunsaturated fatty acid biosynthesis, aminosugars metabolism, and carbon fixation.

The second cluster is 90 degrees out of phase with the first one, reaching maximum expression in the middle of the dark period and minimum expression in the middle of the light period. A member of the Period family of genes, *Per2*, appears in this cluster. These genes operate in a feedback loop with genes like *Arntl* from cluster 1 to regulate the clock cycle (Reppert and Weaver, 2001). Mutations to *Per2* can lead to a dysregulation of the normal circadian period, illustrating its importance in this response (Steinlechner et al., 2002). The genes in cluster 2 belong to molecular pathways related to DNA replication and protein synthesis, including DNA polymerase, purine metabolism, and amino acid metabolism.

Cluster 3 is 180 degrees of phase with the previous cluster; the maximum expression is in the middle of the light period and the minimum expression is in the middle of the dark period. The genes in this cluster are highly enriched in the p53 signaling pathway, suggesting that they play a role in the regulation of the cell cycle.

Finally, cluster 4 is the opposite of cluster 1. Its genes increase in expression during the light period and decrease in expression during the dark period. One of the genes in this cluster is *RevErbA*α (also known as *Nr1d1*), which is another gene that has historically been implicated in circadian rhythms. It functions as a transcription factor for many circadian genes, including *Arntl* from cluster 1 (Yan et al., 2008).

Conclusions

This article introduces a novel symbolic representation that can be used to cluster gene expression data. In addition, we present a procedure for selecting a subset of biologically informative clusters by searching for overrepresented patterns in the data; these patterns likely correspond to coordinated cellular responses. The selection process is validated by running the algorithm on three different datasets and observing the correspondence between the results and our current biological understanding.

There are several features of the proposed method that make it an intriguing alternative to previous methods. It searches for patterns common in multiple genes in the data instead of individually assigning each gene a score representing how different its expression is between groups. Experiments with high temporal resolution can be analyzed because arbitrarily long time series data is accepted. Repeated measurements are used not only to more accurately determine a gene expression value; the variance of repeated measurements is also a key component in determining if there is a significant deviation between classes. These features are critical in discovering biologically relevant information in large sets of gene expression data, and as microarray experiments continue to increase in scope, redundancy, length, and temporal resolution, these issues will only gain importance.

Acknowledgments

J.D.S. and I.P.A. acknowledge financial support from the NIH under Grant GM082974, the EPA under Grant GAD R 832721-010. R.R.A., D.C.D., and W.J.J. acknowledge financial support from the NIH under Grant GM 24211.

Author Disclosure Statement

No competing financial interests exist.

References

Almon, R., DuBois, D., Yao, Z., Hoffman, E., Ghimbovschi, S., and Jusko, W. (2007). Microarray analysis of the temporal response of skeletal muscle to methylprednisolone: comparative analysis of two dosing regimens. Physiol Genomics 30, 282.

Almon, R.R., Yang, E., Lai, W., Androulakis, I.P., DuBois, D.C., and Jusko, W.J. (2008). Circadian variations in liver gene expression: relationships to drug actions. J Pharmacol Exp Ther 326, 700–716.

Androulakis, I.P., Yang, E., and Almon, R.R. (2007). Analysis of time-series gene expression data: methods, challenges, and opportunities. Annu Rev Biomed Eng 9, 205–228.

Angelini, C., De Canditiis, D., Mutarelli, M., and Pensky, M. (2007). A Bayesian approach to estimation and testing in time-course microarray experiments. Stat Appl Genet Mol Biol 6, 24.

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., et al. (2009). NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res 37, D885.

Brown, P.O., and Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. Nat Genet 21, 33–37.

Bryant, P.A., Venter, D., Robins-Browne, R., and Curtis, N. (2004). Chips with everything: DNA microarrays in infectious diseases. Lancet Infect Dis 4, 100–111.

Bunger, M.K., Wilsbacher, L.D., Moran, S.M., Clendenin, C., Radcliffe, L.A., Hogenesch, J.B., et al. (2000). Mop3 is an essential component of the master circadian pacemaker in mammals. Cell 103, 1009–1017.

Churchill, G.A. (2002). Fundamentals of experimental design for cDNA microarrays. Nat Genet 32, 490–495.

Das, U.N. (2000). Essential fatty acids and osteoporosis. Nutrition 16, 386–390.

Daw, C.S., Finney, C.E.A., and Tracy, E.R. (2003). A review of symbolic analysis of experimental data. Rev Sci Instrum 74, 915–930.

Draper, N., and Stewart, P.M. (2005). 11β-Hydroxysteroid dehydrogenase and the pre-receptor regulation of corticosteroid hormone action. J Endocrinol 186, 251–271.

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

Engelen, M.P.K.J., Wouters, E.F.M., Deutz, N.E.P., Menheere, P.P.C.A., and Schols, A.M.W.J. (2000). Factors contributing to alterations in skeletal muscle and plasma amino acid profiles in patients with chronic obstructive pulmonary disease. Am J Clin Nutr 72, 1480.

Ernst, J., and Bar-Joseph, Z. (2006). STEM: a tool for the analysis of short time series gene expression data. BMC Bioinformatics 7, 191.

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

Heard, N.A., Holmes, C.C., and Stephens, D.A. (2006). A quantitative study of gene regulation involved in the immune response of Anopheline mosquitoes: an application of Bayesian hierarchical clustering of curves. J Am Stat Assoc 101, 18–29.

Hoyle, D.C., Rattray, M., Jupp, R., and Brass, A. (2002). Making sense of microarray data distributions. Bioinformatics 18, 576–584.

Hubert, L., and Arabie, P. (1985). Comparing partitions. J Classif 2, 193–218.

Jin, J.Y., Almon, R.R., DuBois, D.C., and Jusko, W.J. (2003). Modeling of corticosteroid pharmacogenomics in rat liver using gene microarrays. J Pharmacol Exp Ther 307, 93–109.

Lee, M.L.T., Kuo, F.C., Whitmore, G.A., and Sklar, J. (2000). Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc Natl Acad Sci USA 97, 9834–9839.

Lin, J., Keogh, E., Lonardi, S., and Patel, P. (2002). Finding motifs in time series. *The 2nd Workshop on Temporal Data Mining, the 8th ACM Int'l Conference on Knowledge Discovery and Data Mining,* Edmonton, Alberta, Canada.

Lin, J., Keogh, E., Wei, L., and Lonardi, S. (2007). Experiencing SAX: a novel symbolic representation of time series. Data Mining Knowledge Discov 15, 107–144.

May, R.C., Bailey, J.L., Mitch, W.E., Masud, T., and England, B.K. (1996). Glucocorticoids and acidosis stimulate protein and amino acid catabolism in vivo. Kidney Int 49, 679.

Phang, T.L., Mc Neville, M.R., and Hunter, L. (2003). Trajectory clustering: a non-parametric method for grouping gene expression time courses, with applications to mammary development. Pac Symp Biocomput 351–362.

Reppert, S.M., and Weaver, D.R. (2001). Molecular analysis of mammalian circadian rhythms. Annu Rev Physiol 63, 647–676.

Schimmer, B.P., and Parker, K.L. (1996). Adrenocorticotropic hormone; adrenocortical steroids and their synthetic analogs; inhibitors of the synthesis and actions of adrenocortical hormones. In *Goodman and Gilman's The Pharmacological Basis of Therapeutics*, 9th (McGraw-Hill, New York), pp. 1459–1486.

Steinlechner, S., Jacobmeier, B., Scherbarth, F., Dernbach, H., Kruse, F., and Albrecht, U. (2002). Robust circadian rhythmicity of Per1 and Per2 mutant mice in constant light, and dynamics of Per1 and Per2 gene expression under long and short photoperiods. J Biol Rhythms 17, 202.

Storey, J.D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. Proc Natl Acad Sci USA 100, 9440.

Storey, J.D., Dai, J.Y., and Leek, J.T. (2007). The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. Biostatistics 8, 414.

Tusher, V.G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 98, 5116–5121.

Ueda, H.R., Hayashi, S., Chen, W., Sano, M., Machida, M., Shigeyoshi, Y., et al. (2005). System-level identification of transcriptional circuits underlying mammalian circadian clocks. Nat Genet 37, 187–192.

Wolfe, C.J., Kohane, I.S., and Butte, A. J. (2005). Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. BMC Bioinformatics 6, 227.

Yan, J., Wang, H., Liu, Y., and Shao, C. (2008). Analysis of gene regulatory networks in the mammalian circadian rhythm. PLoS Comput Biol 4, e1000193.

Yang, E., Maguire, T., Yarmush, M.L., Berthiaume, F., and Androulakis, I.P. (2007). Bioinformatics analysis of the early inflammatory response in a rat thermal injury model. BMC Bioinformatics 8, 10.

Yang, E., Almon, R., DuBois, D., Jusko, W., and Androulakis, I. (2008). Extracting global system dynamics of corticosteroid genomic effects in rat liver. J Pharmacol Exp Ther 324, 1243–1254.

Yao, J., Chang, C., Salmi, M.L., Hung, Y.S., Loraine, A., and Roux, S.J. (2008). Genome-scale cluster analysis of replicated microarrays using shrinkage correlation coefficient. BMC Bioinformatics 9, 288.

Yeung, K.Y., Medvedovic, M., and Bumgarner, R.E. (2003). Clustering gene-expression data with repeated measurements. Genome Biol 4, R34.

Address correspondence to:
Ioannis P. Androulakis
Biomedical Engineering Department
Rutgers University
599 Taylor Road
Piscataway, NJ 08854

E-mail: yannis@rci.rutgers.edu