PEM: A GENERAL STATISTICAL APPROACH FOR IDENTIFYING DIFFERENTIALLY EXPRESSED GENES IN TIME-COURSE CDNA MICROARRAY EXPERIMENT WITHOUT REPLICATE

Xu Han*
Genome Institute of Singapore, 60, Biopolis Street, Singapore 138672
*Email: hanxu@gis.a-star.edu.sg

Wing-Kin Sung
Genome Institute of Singapore, 60, Biopolis Street, Singapore 138672
School of Computing, National University of Singapore, Singapore 117543
Email:sungk@gis.a-star.edu.sg, ksung@comp.nus.edu.sg

Lin Feng
School of Computer Engineering, Nanyang Technological University, Singapore 637553
Email: asflin@ntu.edu.sg

Replication of time series in microarray experiments is costly. To analyze time series data with no replicate, many model-specific approaches have been proposed. However, they fail to identify the genes whose expression patterns do not fit the pre-defined models. Besides, modeling the temporal expression patterns is difficult when the dynamics of gene expression in the experiment is poorly understood. We propose a method called PEM (Partial Energy ratio for Microarray) for the analysis of time course cDNA microarray data. In the PEM method, we assume the gene expressions vary smoothly in the temporal domain. This assumption is comparatively weak and hence the method is general enough to identify genes expressed in unexpected patterns. To identify the differentially expressed genes, a new statistic is developed by comparing the energies of two convoluted profiles. We further improve the statistic for microarray analysis by introducing the concept of partial energy. The PEM statistic is incorporated into the permutation based SAM framework for significance analysis. We evaluated the PEM method with an artificial dataset and two published time course cDNA microarray datasets on yeast. The experimental results show the robustness and the generality of the PEM method. It outperforms the previous versions of SAM and the spline based EDGE approaches in identifying genes of interest, which are differentially expressed in various manner.

Keywords: Time course, cDNA microarray, differentially expressed gene, PEM.

1. INTRODUCTION

Time-course cDNA microarray experiments are widely used to study the cell dynamics from a genomic perspective and to discover the associated gene regulatory relationship. Identifying differentially expressed genes is an important step in time course microarray data analysis to select the biologically significant portion from the genes available in the dataset. A number of solutions have been proposed in the literature for this purpose.

When replicated time course microarray data is available, various statistical approaches, like ANOVA and its modifications, are employed (Lönnstedt & Speed, 2002; Park *et al.*, 2003; Smyth, 2004). This category of approaches has been extended to recent work on longitudinally sampled data, where the microarray measurements span in multi-dimensional space with the coordinates to be gene index, individual donor, and time point, etc. (Guo *et al.*, 2003; Storey *et al.*, 2005). However, replication of time series or longitudinal sampling is costly if the number of time points is comparatively large. For the sake of this, many published time course datasets have no replicate.

* Corresponding author.

When replicated time course is not available, clustering based approaches and model-specific approaches are widely used.

Clustering based approaches select genes whose patterns are similar to each other. A famous example of clustering software is the Eisen's Cluster (Eisen *et al.*, 1998). Clustering based approaches are advantageous in finding co-expressed genes. The drawback is that clustering does not provide a ranking for the individual genes, and it is difficult to determine a cut-off threshold based on confidence analysis. Additionally, cluster analysis may fail to detect changing genes that belong to clusters for which most genes do not change (Bar-Joseph *et al.*, 2003).

Model-specific approaches identify differentially expressed genes based on prior knowledge of their temporal patterns. For instance, Spellman *et al*. (1998) used Fourier transform to identify cell-cycle regulated genes; Peddada *et al.* (2003) proposed an order-restricted model to select responsive genes; Xu *et al.* (2002) developed a regression-based approach to identify the genes induced in Huntington's disease transgenic model; in the recent versions of SAM (Tusher *et al.*, 2001), two alternative methods, slope based and signed area based, are provided for analyzing single time course data. However, the assumption underlying the model-specific approaches is too strong and some biologically informative genes that do not fit the pre-defined model may be ignored. Bar-Joseph *et al.* (2002) proposed a spline based approach, which is established on comparatively weaker assumptions. The software of EDGE (Storey *et al.*, 2005) implemented natural cubic spline and polynomial spline for testing the statistical significance of genes. In spline based approaches, the dimension of spline needs to be chosen carefully to balance the robustness and the diversity of gene patterns, and an empirical setting of dimension may not be applicable for some applications.

The goal of this paper is to propose a new statistical method called PEM (Partial Energy ratio for Microarray) for the analysis of time course cDNA microarray data. In time-course experiments, the measurements are sampled from continuously varying gene expressions. Thus it is often observed that the log-ratio expression profiles of the differentially expressed genes are featured with "smooth" patterns, of which the energies mainly concentrate in low frequency. To utilize this feature, we employ two simple convolution kernels that function as a low-pass filter and a high-pass filter, namely smoothing kernel and differential kernel, respectively. The basic statistic for testing the smoothness of a temporal pattern is represented by the energy ratio of the convoluted profiles. We further improve the performance of the statistic for microarray anlaysis by introducing a concept called partial energy to solve the problem caused by "steep edge", which refers to rapid increasing or decreasing of gene expression level. The proposed ratio statistic is incorporated into the permutation based SAM (Tusher *et al.*, 2001) framework for determining confidence interval and false discovery rate (Benjamini and Hochberg, 1995). In the SAM framework, a small positive constant called "relative difference" is added to the denominator of the ratio, which efficiently stabilizes the variance of the proposed statistic.

An artificial dataset and two published cDNA microarray datasets are employed to evaluate our approach. The published datasets include the yeast environment response dataset (Gasch *et al*, 2000) and the yeast cell cycle dataset (Spellman *et al*, 1998). The experiment results showed the robustness and generality of the proposed PEM method. It outperforms previous versions of SAM and spline based EDGE in identifying genes differentially expressed in various manner. In the experiment with yeast cell cycle dataset, the PEM method not only identified the periodically expressed genes, but also identified a set of non-periodically expressed genes, which are verified to be biologically informative.

2. METHOD

2.1 Signal/noise model for cDNA microarray data

Consider a two-channel cDNA time-course microarray experiment over m genes: $g_1, g_2, \ldots, g_m$, and n time points: $t_1, t_2, \ldots, t_n$. the log-ratio expression profile of the gene g_i (i = 1 to m) can be represented by $X_i = [X_i(t_1), X_i(t_2), \ldots X_i(t_n)]^T$, where $X_i(t_j)$ (j = 1 to n) represents the log-ratio expression value of g_i at the jth time point.

We model the log-ratio expression profile X_i as the sum of its signal component $S_i = [S_i(t_1), S_i(t_2), \ldots S_i(t_n)]^T$ and its noise component $\varepsilon_i = [\varepsilon_i(t_1), \varepsilon_i(t_2), \ldots \varepsilon_i(t_n)]^T$, i.e. $X_i = S_i + \varepsilon_i$. We have the following assumption on the noise component:

Assumption of noise: $\varepsilon_i(t_1)$, $\varepsilon_i(t_2)$, …, $\varepsilon_i(t_n)$ are independent random variables following a symmetric distribution with the mean equal to zero.

Note that the noise distribution in our assumption is not necessarily normal so that this gives a better model of the heavily tailed symmetrical noise distribution that is often observed in microarray log-ratio data.

For a non-differentially expressed gene g_i, we assume its expression signals in two channels are identical at all the time points. In this case, the signal component S_i is constantly zero, and the log-ratio expression profile X_i only consists of the noise component. Thus the null hypothesis is defined as follow:

$$H_0: \quad X_i = \varepsilon_i$$

Due to the variation of populations in cDNA microarray experiments, there is bias between the expression signals in two channels. Thus the assumption underlying the null hypothesis may not be established if the log-ratios are calculated directly from the raw data. We suggest using pre-processing approaches such as Lowess regression to compensate the global bias (Yang *et al.*, 2002). To further overcome the influence of the gene-specific bias, we adopted the SAM framework, in which a small positive constant called "relative difference" was introduced to stabilize the variance of the statistic (Tusher *et al.*, 2001). Nevertheless, the null hypothesis provides a mathematical foundation for demonstration of our method.

2.2 Smoothing convolution and differential convolution

In time-course experiments, the measurements are sampled from continuously varying gene expressions. If there is adequate number of sampled time points, the temporal pattern of the signal S_i will be comparatively smooth so that the energy of S_i will concentrate in low frequency. To utilize this feature, we introduce two simple convolution kernels for time series data analysis, namely the smoothing kernel and the differential kernel. The smoothing kernel is represented by a sliding-window W_s = [1, 1], and the differential kernel is represented by W_d = [-1, 1]. In signal processing, the smoothing kernel and the differential kernel function as a low-pass filter and a high-pass filter, respectively. to detect the edges.

Given a vector $V = [V(t_1), V(t_2), \ldots, V(t_n)]^T$ representing a time-series, the smoothed profile and the differential profile of V are represented by $V*W_s = [V(t_1) + V(t_2), V(t_2) + V(t_3), \ldots, V(t_{n-1}) + V(t_n)]^T$, and $V*W_d = [V(t_1) - V(t_2), V(t_2) - V(t_3), \ldots, V(t_{n-1}) - V(t_n)]^T$, respectively, where * is the convolution operator.

Since the energy of the signal component S_i is likely to concentrate in low frequency, we have:

Assumption of signal: If S_i is a non-zero signal vector, then

$$E(|S_i * W_s|^2) > E(|S_i * W_d|^2)$$

where $E(|S_i * W_s|^2)$ and $E(|S_i * W_d|^2)$ represent the expected energies of the corresponding smoothed profile and differential profile.

Next, we derive two propositions from the *Assumption of noise* and the *Assumption of signal*, as follows:

***Proposition 1*:** If the noise component ε_i satisfies the *Assumption of noise*, then

$$E(|\varepsilon_i * W_s|^2) = E(|\varepsilon_i * W_d|^2) \quad (1)$$

Proposition 2: If the signal component S_i satisfies *Assumption of signal*, and the noise component ε_i satisfies the *Assumption of noise*, then

$$E(|(S_i + \varepsilon_i) * W_s|^2) > E(|(S_i + \varepsilon_i) * W_d|^2) \quad (2)$$

Propositions 1 and *2* can be proven based on the symmetry of noise distribution and the linear decomposability of convolution operation.

Note that the log-ratio expression profile $X_i = S_i + \varepsilon_i$. According to Eq. (1) and Eq. (2), we define a statistic called energy ratio (ER) for testing the null hypothesis, as follow:

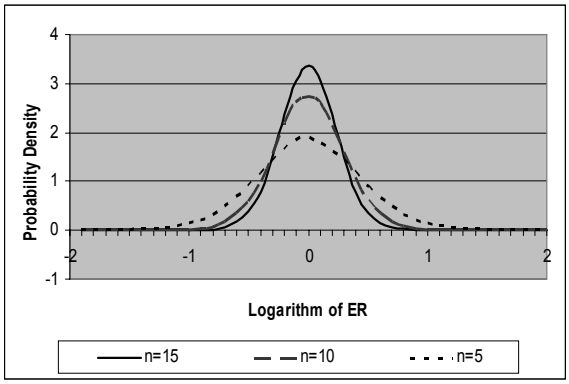


Fig. 1. The numerically estimated distribution of logarithm of ER(ε_i) , where n is the number of time points.

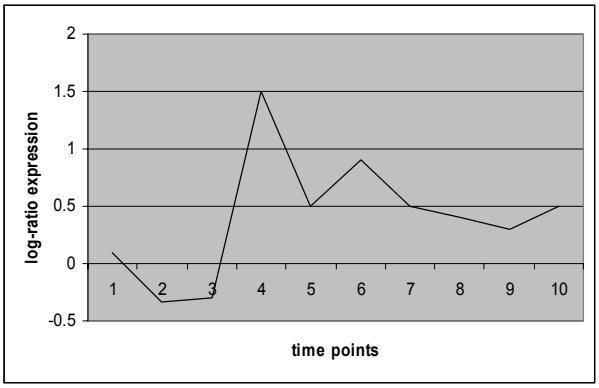


Fig. 2. An example of responsive gene expression profile where a "steep edge" occurs between the 3rd and the 4th time points.

$$ER(X_i) = \frac{|X_i * W_s|^2}{|X_i * W_d|^2} \quad (3)$$

The distributions of logarithm of $ER(\varepsilon_i)$ are shown in Fig. 1, where the number of time points varies from 5 to 15 and the distribution of ε_i is multivariate normal. We take logarithm simply for the convenience of visualization. Obviously, the logarithm of $ER(\varepsilon_i)$ follows a symmetric distribution highly peaked around zero mean. The distribution is two-tailed, but we are only interested in the positive tail when testing the null hypothesis. This is because the negative tail implies the energy concentrates in the high frequency. According to Nyquist sampling criterion, the high frequency component is not adequately sampled thus the expression profile may not be reliable. When $n \to \infty$, $ER(\varepsilon_i)$ is asymptotically independent on the distribution of ε_i, which can be easily proven based on central limit theorem.

2.3 Partial energy

In most time-course microarray experiments, the number of time points is limited. Due to insufficient sampling, the smoothness of the signal component S_i is not guaranteed at all the time points. We call this a "steep edge" problem. A steep edge refers to rapid increasing or decreasing of gene expression level at certain time points. Fig. 2 shows an example of responsive gene expression profile in which a steep up-slope edge occurs between the 3rd and the 4th time points. When the number of time points is limited, the steep edge adds a large value to the denominator in Eq. (3), hence reduces the statistical significance of the ER score.

To solve the "steep edge" problem, we propose a new concept called partial energy. The basic idea of partial energy is to exclude the steep edges in calculating the energy of a differential profile. Denote $Y = [Y_1, Y_2, \ldots Y_n]^T$ be a vector representing a profile, the k-order partial energy of Y is defined as:

$$PE_k(Y) = \sum_{i=1}^{n} Y_i^2 - \sum_{i=1}^{k} Y_{(i)}^2$$

where $k<n$, and $Y_{(i)}^2$ represents the ith biggest value of $Y_1^2, Y_2^2, \ldots, Y_n^2$. For example, let $Y = [1, -1, -4, 3, 2]^T$, its 2-order partial energy $PE_2(Y) = 1^2 + (-1)^2 + 2^2 = 6$, where -4 and 3 are excluded in calculating the partial energy.

For most responsive patterns in microarray data, the number of steep edges is much smaller than the number of time points. We assume there are no more than 2 steep edges in the gene expression profile, and modify the statistic to be the ratio of the 2-order partial energies (PER_2) of the smoothed profile and the differential profile:

$$PER_2(X_i) = \frac{PE_2(X_i * W_s)}{PE_2(X_i * W_d)} \quad (4)$$

where * is the convolution operator, and W_s and W_d are the smoothing kernel and the differential kernel defined in section 2.2.

2.4 Significance analysis

Since the PER_2 statistic defined in Eq. (4) takes the form of a ratio, it can be easily incorporated into the SAM framework (Tusher *et al.*, 2001) for significance analysis.

In the first step, a "relative difference" s_0 is added to the denominator in Eq. (4), as follow:

$$PEM(X_i) = \frac{PE_2(X_i * W_s)}{PE_2(X_i * W_d) + s_0} \quad (5)$$

For the sake of simplicity, the constant s_0 is chosen to be the 5 percentile of PE_2 (X_i*W_d) for all the genes (i = 1 to m). By adding introducing the relative difference, the genes with small fold-changes are excluded from the top-ranking list. This efficiently reduces the influence of channel bias and stabilizes the variance of statistic. The statistic defined in Eq. (5) is called PEM (Partial Energy ratio for Microarray).

Secondly, we employ the algorithm of SAM for determining the confidence interval and the false discovery rate (sometimes called q-value). For the detail of the algorithm, one can refer to the SAM manual available at the website: http://www-stat.stanford.edu/~tibs/SAM/. Here, we briefly describe our strategy of randomized permutation. With the PEM statistic, the procedure of permutation consists of two steps. In the first step, the order of the log-ratio measurements in the expression profile are randomly permutated for each gene; in the second step, the signs of the measurements are randomly flipped. The second step is based on the *Assumption of noise*, where the distributions of the measurements of non-differentially expressed genes are assumed to be symmetric with zero mean.

3. EXPERIMENTS

The robustness and generality of the proposed PEM method are evaluated with both simulation dataset and published microarray datasets, which include the yeast environment response dataset (Gasch *et al*, 2000) and the yeast cell cycle dataset (Spellman *et al*, 1998). The missing values in the published datasets are filled in using KNN-Impute (Troyanskaya et al., 2001). The evaluation is based on relative operating characteristic (ROC) score, which gives a reasonable measurement of sensitivity vs. specificity. (McNeil and Hanley, 1984)

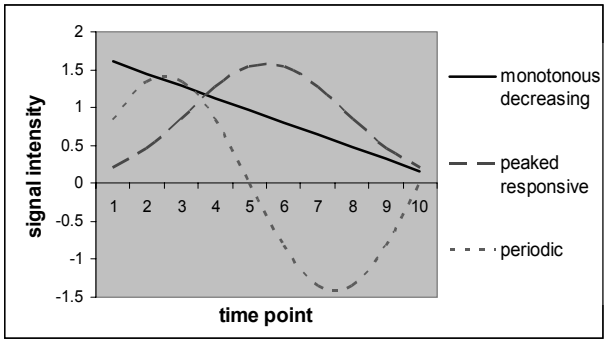
For comparison, our evaluation also includes the approaches employed in two of the most popular microarray analysis software, which are the SAM (Tusher, 2001) and the spline based EDGE (Storey *et al.*, 2005).

Recent version of SAM provides two alternative approaches for the analysis of single time course data. They are slope based and signed area based, respectively. The slope based SAM is designed for identifying the genes with monotonous increasing or decreasing patterns, and the signed area based SAM is an improved version of paired t-test.

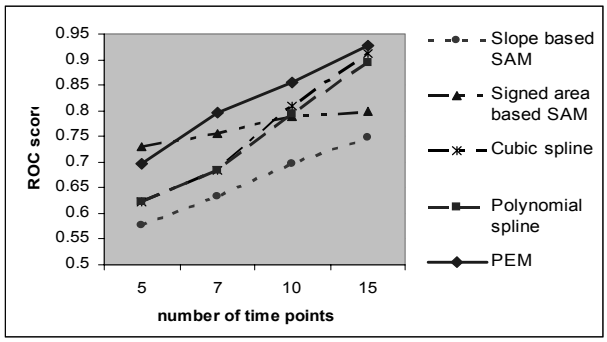
In the EDGE software, both the natural cubic spline and the polynomial spline based approaches are included in our evaluation. The dimension of splines is empirically optimized to be 4 in the simulation and in the yeast environment response experiments. In yeast cell cycle experiment, the dimension of splines is optimized to be 8 for cubic spline, and is set to be 5 for polynomial spline to avoid singular matrix in calculation.

3.1 Simulation

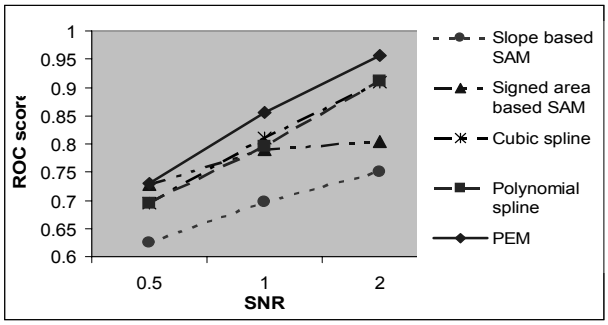
In the simulation experiment, each log-ratio expression profile is generated by summing its signal component and its noise component. The noise component follows normal distribution with zero mean. For non-differentially expressed genes, the intensity of the signal component is constantly zero. For differentially expressed genes, the signal component is one of the three frequently observed signal patterns in time course microarray data, as shown in Fig. 3(a). They are monotonous decreasing pattern defined by linear function, peaked responsive pattern defined by Gaussian function, and periodic pattern defined by sine function. There are two free parameters in our simulation test: the number of time points and the signal- noise ratio (SNR). For each parameter setting, we generate 6000 artificial time course gene expression profiles, of which 5400



(a) Three basic patterns of gene expression profile defined in the artificial dataset;



(b) ROC scores under variant number of time points;



c) ROC score vs. S/N ratio

Fig. 3. The numerically estimated distribution of logarithm of ER(ε_i) , where n is the number of time points.

(90%) belong to non-differentially expressed genes and 600 (10%) belong to differentially expressed genes. The 600 profiles of differentially expressed genes are equally divided into 3 portions, corresponding to monotonous decreasing pattern, peaked responsive pattern, and periodic pattern.

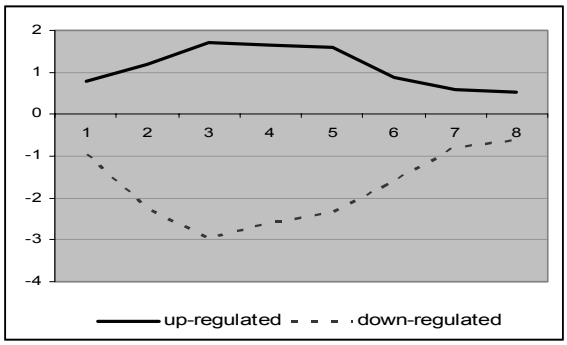
First, we generate artificial datasets by setting the SNR to be 1.0 and the numbers of time points to be 5, 7, 10, or 15. The ROC scores are plotted against the number of time points in Fig. 3(b). Next, we fix the number of time points to be 10 and set the SNR to be 0.5, 1.0, and 2.0. The ROC scores are plotted against SNR in Fig. 3(c).

The result of simulation experiment demonstrates that PEM achieves the best overall performance among the methods in evaluation. The signed area based SAM is the most robust when the number of time points is 5. However, as the number of time points increases or the SNR becomes larger, the PEM and EDGE approaches achieve much higher ROC score than SAM. This is because the SAM approaches are modeled base on specific patterns, while the models underlying PEM and EDGE are more general.

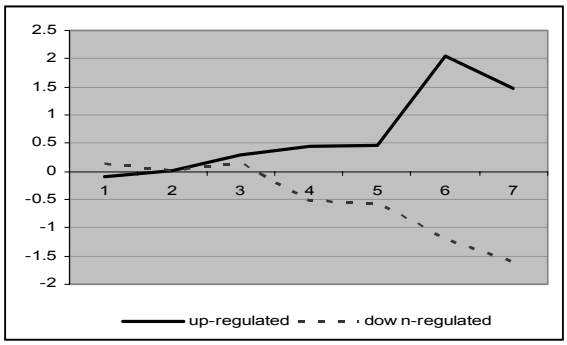
3.2 Evaluation with yeast environment response dataset

The yeast environment response dataset consists of measurements in 173 arrays published in (Gasch *et al.*, 2000) and (Derisi *et al.*, 1997). The dataset is used to discover the way in which the budding yeast *S. Cerevisiae* cells adapt to variant changing environments. Among the arrays available in the dataset, we selected 79 arrays based on two criteria: (i) population from wild-type cells; (ii) at least 7 time points sampled under each condition. These arrays fall into 10 individual experiments:

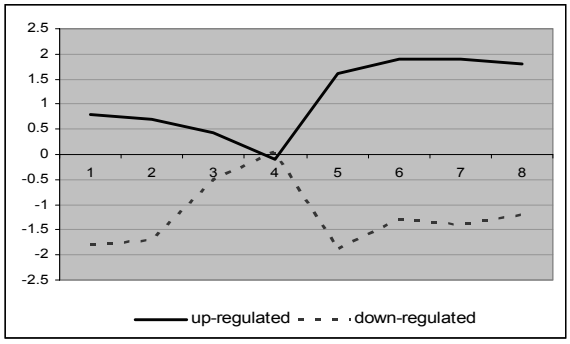
- Heat shock from 25°C to 37°C, consisting of 8 time points;
- Hydrogen peroxide treatment, consisting of 10 time points;
- Menadione exposure, consisting of 9 time points;
- DTT exposure, consisting of 8 time points;
- Diamide treatment, consisting of 8 time points;
- Hyper-osmotic shock, consisting of 7 time points;
- Nitrogen source depletion, consisting of 10 time points;
- Diauxic shift, consisting of 7 time points;



(a) heat shock



(b) diauxic shift



c) nitrogen depletion

Fig. 4. Average expression patterns of ERS genes in variant experiments.

• Stationary phase, including two nearly-identical experiments consisting of 10 and 12 time points, respectively.

We assess the approaches by applying them to the 10 time course experiments individually. To evaluate the sensitivity and specificity of the methods, we use a list of 270 genes available at the website of Chen et al. (2003). This list is the intersection of (i) around 800 Environment Stress Response (ESR) genes which were identified by Gasch *et al.* (2000) using hierarchical clustering on multiple experiments, and (ii) a list of ortholog genes in fusion yeast *S. Pombe* which are differentially expressed under environment stress. Figure 4 shows that these evolutionarily conserved ERS genes are expressed with various expression patterns in different experiments so that they provide a good test-bed to evaluate the robustness and the generality of the methods.

Table 1. ROC scores of the evaluated methods on environment response experiments. The bold fonts correspond to the highest scores in the rows.

Experiment	Slope based SAM	Signed area based SAM	Cubic Spline based EDGE	Poly. Spline based EDGE	PEM
Heat shock	0.501	0.753	0.841	0.848	**0.889**
Hydrogen peroxide	0.626	0.782	0.775	0.789	**0.792**
Menadione	0.557	0.552	0.588	**0.604**	0.565
DTT	**0.743**	0.522	0.722	0.723	0.722
Diamide	0.498	0.667	0.808	0.800	**0.838**
Hyper-osmotic shock	0.639	0.593	0.651	0.688	**0.736**
Nitrogen depletion	0.459	0.699	0.628	0.605	**0.715**
Diauxic shift	**0.758**	0.632	0.710	0.705	0.728
Stationary phase, expr. 1	0.494	0.841	0.702	0.673	**0.860**
Stationary phase, expr. 2	0.745	0.786	0.739	0.732	**0.889**
Average	0.602	0.683	0.716	0.717	**0.773**
P-value of paired t-test	4.0E-3	1.6E-3	8.4E-3	1.8E-2	__

The ROC scores for methods are summarized in Table 1. The PEM method outperforms other methods in 7 out of 10 experiments. It achieves reasonably good ROC scores (>0.7) in most experiments, except for the Menadione exposure experiment in which all the methods do not perform well. To further show the

superiority of PEM, we averaged the ROC scores over all experiments for each method, and used paired t-test for comparison of the performance of PEM and the other methods. The p-values of the paired t-test demonstrate the significance of the improvement made by PEM.

3.3 Evaluation with Yeast Cell Cycle Dataset

The yeast cell cycle dataset (Spellman *et al.*, 1998) consists of the measurements in three experiments (Alpha factor, CDC15, CDC28) on cell cycle synchronized yeast *S. Cerevisiae* cells. We employed a reference list containing 104 cell cycle regulated genes determined by traditional biological experiments, as mentioned in the original paper. In addition to SAM and EDGE, we also include the method of Fourier transform (Spellman *et al.*, 1998) in our evaluation. The Fourier transform (FT) method was introduced specifically for identifying periodically expressed genes.

The ROC scores are shown in Table 2. The PEM method outperforms the SAM approaches and the spline based EDGE approaches in all experiments. The FT method performs slightly better than PEM in identifying periodically expressed genes. However, the PEM method also identified a number of non-periodically expressed genes, which account for considerable false positives in calculating ROC scores. To show this, we clustered the top 706 differentially expressed genes identified by the PEM in the alpha factor experiment. These genes are selected based on a false discovery rate equal to 0.1. We applied K-mean clustering using Eisen's Cluster software (Eisen *et al.*, 1998) and came up with eight clusters, as shown in Fig. 5. Five of the clusters are periodic and the remaining three are non-periodic. Note that the non-periodic portion of the differentially expressed genes is not significant with the Fourier transform approach. The non-periodic clusters are mapped to the gene ontology clusters using GO Term Finder in SGD database (http://db.yeastgenome.org/cgi-bin/GO/goTermFinder/). We selected four significant gene ontology terms corresponding to the non-periodic clusters, as listed in Table 3. The Bonferroni correlated hypergeometric P-values show

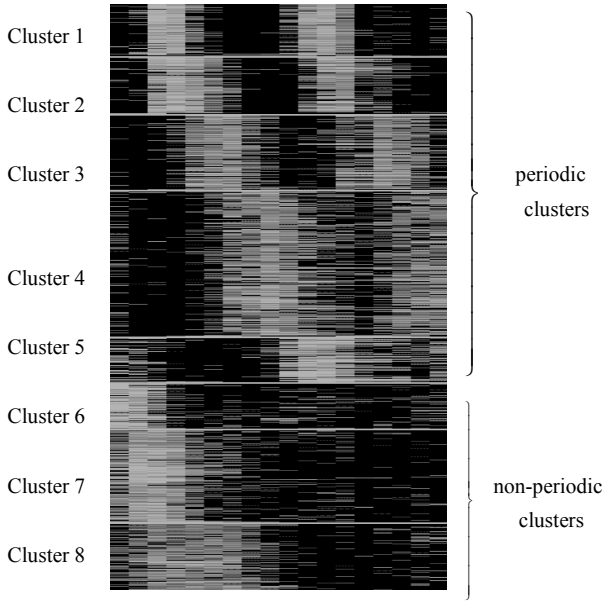


Fig. 5. Clustering result shows periodic and non-periodic patterns of differentially expressed genes identified by PEM in alpha factor experiment.

Table 2. ROC scores for evaluation of the methods in identifying periodically expressed cell cycle regulated genes.

	Slope based SAM	Signed area based SAM	Cubic Spline based EDGE	Poly. Spline based EDGE	FT	PEM
alpha	0.579	0.679	0.854	0.777	**0.917**	0.883
cdc15	0.402	0.386	0.804	0.590	**0.811**	0.808
cdc28	0.485	0.526	0.747	0.705	**0.859**	0.763

Table 3. Selected significant gene ontology terms mapped to non-periodic clusters. The GO terms and cluster IDs are retrieved from SGD database.

	Significant gene ontology term	GO cluster ID	P-value
Cluster 6	Sexual reproduction	GO:0019953	4.14e-14
Cluster 7	Oxidoreductase activity	GO:0016491	1.32e-6
Cluster 8	Glutamate biosynthesis	GO:0006537	1.00e-8
	Energy derivation by oxidation of organic compund	GO:0015980	3.38e-7

that these non-periodic clusters are biologically meaningful.

The evaluation with the yeast *S. Cerevisiae* cell cycle dataset clearly indicates the ability of the PEM method in identifying genes with either periodic or non-periodic patterns. In comparison to model-specific approaches like Fourier transform, the PEM method is more general and leads to a better overview of the dynamics of gene expression changes.

4 CONCLUSION AND DISCUSSION

Replications in time course microarray experiments are costly. In selecting methods for analysis without replicates, people usually have to face the problem to make tradeoff between robustness and generality. If the assumption underlying the method is too strong, the method may fail to identify the genes whose expression patterns do not fit the pre-defined model. In this paper, we propose a general statistical method called PEM (Partial Energy ratio for Microarray), for identifying differentially expressed genes in time course cDNA microarray experiment without replicates. In the PEM method, we assume the gene expressions vary smoothly in time series. This assumption is comparatively weak hence the PEM method is more general in identifying genes expressed in unexpected patterns. To identify differentially expressed genes, we employed convolution kernels in our statistic and introduced the concept of partial energy. The proposed statistic can be easily incorporated into the SAM framework for significance analysis, in which the variance of the statistic is stabilized by introducing the "relative difference". Experimental results show the robustness and generality of PEM when the number of time points is comparatively large (>6). Another advantage of PEM is that the parameters of PEM can be fixed for applications of different experiments, although automatic determination of the optimal parameters may slightly improve the performance, which will be investigated in the future.

The main limitation of the PEM method is, the assumption of signal smoothness may not be satisfied if the measurements are not adequately sampled. In this case, replication of the time series is necessary. Thus, we will also explore the possibility of modifying the PEM method for the applications where replicates are available. For this problem, one possible solution is to integrate the PEM statistic and the ANOVA F-score using a permutation based strategy, which will be implemented and be tested in near future.

Acknowledgments

The authors thank Dr. Krishna, Karuturi Radha, Dr. Vladimir Andreevich Kuznetsov, Mr. Juntao, Li, and Mr. Vega, Vinsensius Berlian for the valuable discussions on the topics related to this paper.

References

1. Bar-Joseph Z, Gerber G, Simon I, Gifford D, and Jaakkola T. Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. *Proc. Natl Acad. Sci. USA* 2003; **100**: 10146-10151.
2. Bar-Joseph Z, Gerber G, Gifford D, Jaakkola T, and Simon I. A new approach to analyzing gene expression time series data. *RECOMB* 2002: 39-48.
3. Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B.* 1995; **57**: 289-300.
4. Chen D, Toone WM, Mata J, Lyne R, Burns G, Kivinen K, Brazma A, Jones N, and Bähler J. Global transcriptional responses of fission yeast to environmental stress, *Mol. Biol. Cell* 2003; **14**: 214-229.
5. DeRisi JL, Iyer VR, and Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 1997; **278**: 680-686.
6. Eisen MB, Spellman PT, Brown PO, and Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* 1998; **95**: 14863-14868.
7. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, and Brown PO. Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* 2000; **11**: 4241-4257.

8. Guo X, Qi H, Verfaillie CM, and Pan W. Statistical significance analysis of longitudinal gene expression data, *Bioinformatics* 2003; **19**: 1628-1635.
9. Lönnstedt I. and Speed TP. Replicated microarray data. *Statistica Sinica* 2002; **12**: 31-46.
10. McNeil BJ and Hanley JA. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. *Med. Decis. Mak.* 1984; **4**: 137-150.
11. Park T, Yi S, Lee S, Lee SY, Yoo D, Ahn J, and Lee Y. Statistical tests for identifying differentially expressed genes in time-course microarray experiments. *Bioinformatics* 2003; **19**: 694-703.
12. Peddada SD, Lobenhofer EK, Li L, Afshari CA, Weinberg CR, and Umbach DM. Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference, *Bioinformatics* 2003; **19**: 834-841.
13. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology* 2004; **3**: article 3.
14. Spellman PT, Sherlock G, Zhang MO, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, and Futcher B. Comprehensive identification of cell-cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 1998; **9**: 3273-3297.
15. Storey JD, Xiao W, Leek JT, Tompkins RG, and Davis RW. Significance analysis of time course microarray experiments. *Proc. Natl Acad. Sci. USA* 1005; **102**: 12837-12842.
16. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, and Altman RB. Missing value estimation methods for DNA microarrays. *Bioinformatics* 2001; **17**: 520-525.
17. Tusher V, Tibshirani R, and Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA* 2001; **98**: 5116-5121.
18. Xu XL, Olson JM, and Zhao LP. A regression-based method to identify differentially expressed genes in microarray time course studies and its application in an inducible Huntington's disease transgenic model. *Human Molecular Genetics* 2002; **11**: 1977-1985.
19. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, and Speed TP. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nuclceic Acids Res.* 2002; **30**: e15.