A Noise Removal Algorithm for Time Series Microarray Data

Naresh Doni Jayavelu and Nadav Bar

Systems Biology Group, Department of Chemical Engineering,
Norwegian University of Science and Technology (NTNU), Trondheim, NO-7491,
Norway
nareshd@chemeng.ntnu.no

Abstract. High-throughput technologies such as microarray data are a great resource for studying and understanding biological systems at a low cost. However noise present in the data makes it less reliable, and thus many computational methods and algorithms have been developed for removing the noise. We propose a novel noise removal algorithm based on Fourier transform functions. The algorithm optimizes the coefficients of the first and second order Fourier functions and selects the function which maximizes the Spearman correlation to the original data. To demonstrate the performance of this algorithm we compare the prediction accuracy of well known modelling tools, such as network component analysis (NCA), principal component analysis (PCA) and k-means clustering. We compared the performance of these tools on the original noisy data and the data treated with the algorithm. We performed the comparison analysis using three independent real biological data sets (each data set with two replicates). In all cases the proposed algorithm removes the noise in the data and substantially improves the predictions of modelling tools.

Keywords: Microarray time series data, Noise, Smoothing, Fourier, Network component analysis, Principal component analysis, clustering.

1 Introduction

High-throughput technologies such as microarray have emerged as promising tools for studying, modelling and understanding complex biological systems at a low cost. Most often these data sets are prone to noise. This noise arises from stochastic variations in the experiments and changes in the biological processes, for example during sample preparation and hybridization processes [14]. The major challenge in this microarray analysis is to separate the true biological signals from the noisy measurements. Prediction abilities (knowledge discovery) of modelling tools as network component analysis (NCA)[6], principal components analysis [8] and clustering [15] based on these data sets depend heavily on the amount of noise present [11].

There are attempts in the literature to quantify and remove the noise in the microarray [14]. One approach is to replicate the measurements several times,

L. Correia, L.P. Reis, and J. Cascalho (Eds.): EPIA 2013, LNAI 8154, pp. 152–162, 2013.

an expensive approach that requires manpower, time and resources. There are many mathematical models developed for noise removal and smoothing of data [7]. Most of these models are based on the assumption that data is approximately Gaussian distributed [5]. However, Hardin et al [4] reported that microarray data does not necessarily need to satisfy this assumption. There are several other models available. For example Tang et al [13] described a singular value decomposition combined with spectral analysis for noise filtering.

In this article we developed an algorithm based on the Fourier transform function for removing noise and smoothing of the gene expression data. The expression values of the genes in time series experiments are known to follow a specific trends depending on the type of treatment given to the cells. For example, stimuli-response experiments are characterized by transient responses and in cell cycle experiments cyclic patterns are observed. In general, gene expressions follow two simple response shapes: either short impulses or long sustained responses [1]. We fitted individual time series gene expression data using an optimization algorithm that estimate the parameters of first and second order Fourier series functions, and constrains the functions to these two shapes or their combinations. The fitted functions represent the smooth approximations to the original expression values and thus remove the noise. We have applied our algorithm on three real biological microarray data sets. The first data set is gastrin responsive transcriptome data measured at 11 time points. The other two are epidermal growth factor (EGF) and heregulin (HRG) stimuli-response experiments with 17 time points. The data sets can be downloaded from the Array Express Website with accession numbers: GSE32869 and GSE13009 respectively for gastrin and epidermal growth factor and heregulin. We showed that predictions of NCA and PCA on data with noise removed are more consistent and accurate, and the clusters from k-means are tighter and distinct from each other.

2 Methods

Prior to the expression data fitting, we performed the selection of the differentially expressed genes based on fold change and p-value including normalization. The noise reduction algorithm fits the data (expression values of each gene) with Fourier functions and smoothing splines. The function that maximizes the Spearman correlation between the fitted curve and the original noisy data at the given time points is chosen.

2.1 Fourier Series

The Fourier series is a sum of trigonometric functions that describes a periodic signal. The fitting procedure uses a nonlinear least squares regression to estimate the parameters of a Fourier function. This optimization algorithm is a robust Bi-Square fitting method that creates the curve by minimizing the summed square of the residuals, and reduces the weight of outliers using Bi-square weights.

The optimization is performed using a Trust-Region method with coefficient constraints. The Fourier function is represented here in its trigonometric form:

$$f(x) = a_0 + \sum_{i=1}^{n} a_i \cos(n\omega x) + b_i \sin(n\omega x) \tag{1}$$

Where a_0, a_i, and b_i are the parameters to be estimated, a_0 is the constant term of the data and is associated with the i = 0 cosine term, ω is the fundamental frequency of the signal and $1 \leq n \leq \infty$ is the number of harmonics in the series. Since the temporal pattern of gene expression is known to be wave like or impulse-like shapes, the only harmonics we expect during a time span of less than 72 hours are $n = 1$ or $n = 2$, corresponding to the superposition of one or two harmonics, respectively. The fundamental frequency of the function is not expected to be larger than the time span of the micro array experiment, so the fitted function is most likely to consist of one or two peaks. The reason behind this assumption is that not many genes experience cyclic behavior of more than two cycles during a short, several hours time span.

For practical reasons, we use discrete Fourier transform (DFT) to estimate the above parameters, since the time points are discrete and finite. To compute the DFT, we used the fast Fourier transform algorithm in Matlab, Mathworks

2.2 Smoothing Spline

When the Fourier series function is poorly correlated (Spearman correlation) to the data, we fitted the expression data with smoothing spline function in Matlab

2.3 Network Component Analysis

The NCA [6]is a computational method for reconstructing the hidden regulatory signals (activity profiles of transcription factors) from gene expression data with known connectivity information in terms of matrix decomposition. The NCA method can be represented in matrix form as follows:

$$[E] = [C][T] \tag{2}$$

where the matrix E represents the expression values of genes over time points, the matrix C is the control strength of each transcription factor on gene and matrix T represents the activity profiles of transcription factors. The dimensions of E, C and T are $N \times M$ (N is the number of genes and M is the number of time points or measurement conditions), $N \times L$ (L is the number of transcription factors), $L \times M$, respectively.

Based on above formulation, the decomposition of $[E]$ into $[C]$ and $[T]$ can be achieved by minimizing the following objective function:

$$min\|[E] - [C][T]\|^2, s.t. C \in Z_0$$

where Z_0 is the initial connectivity pattern. The estimation of $[C]$ and $[T]$ is performed by using a two-step least-squares algorithm. With NCA as reconstruction method, we predicted significant TFs and their activity profiles.

2.4 Principal Component Analysis

The PCA [9] is a model reduction method that reduces the dimensionality (number of variables) of a data set by preserving as much variance as possible. PCA rotates the original data space such that the axes of the new coordinate system point into the directions of highest variance of the data. The axes or new variables are termed principal components (PCs) and are ordered by variance. We start with the expression matrix, E, where each row corresponds to genes and each column corresponds to different measurement conditions or time points at which cells were treated with a stimuli. The E_{it} entry of the matrix contains the expression value of gene i at condition t. The principal components are computed as:

$$e'_{ij} = \sum_{t=1}^{n} e_{it} v_{tj} \tag{3}$$

Where E_{tj} is the i^{th} coefficient for the j^{th} principal component. e_{it} is the expression value for gene i under t^{th} condition. E' is the data in terms of principal components.

2.5 K-means Clustering

K-means clustering is a powerful technique often employed in gene expression analysis for elucidating a variety of biological inferences. We selected the Pearson correlation as similarity measure and repeated the algorithm with 10 different initial randomizations (initial cluster centroid positions) to avoid local minima and chose the one with smallest sum of point to centroid distances. Selecting the correct number of clusters for a given data is always critical in k-means clustering so we considered a range of 3 to 15 clusters. We chose sum of point to centroid distances within the cluster, 'sumd' (tightness within cluster) and silhouette measure (separation in-between clusters) as clustering evaluation criteria [10, 2, 15].

The efficiency of the noise reduction (NR) algorithm is evaluated by the ability of prediction tools to reproduce the same temporal behaviors, whether it is activity profile of transcription factors (for NCA) or gain in cumulative variance (for PCA) and tightness and separation (for k-means clustering) from original noisy data (N) and data treated with algorithm (NR) . All computations were performed using Matlab, Mathworks.

3 Results and Discussion

3.1 NR Algorithm

The performance of the proposed algorithm (we term it as NR algorithm) in terms of removing the noise from the time series expression data are presented in Figure 1. This algorithm fits the gene expression data with Fourier 1, (impulse response), Fourier 2 (sustained response) or smoothing spline (complex patterns) and chooses the best one based on high correlation between the original and fitted data. In the majority of the genes, noise is repressed by smoothing out strong fluctuations. For instance, the gene expression of ANKZF1 is noisy at late time samples and this noise is removed with NR algorithm (Figure 1). Most of the genes temporal patterns are approximated with either Fourier 1 (%14 of total) or Fourier 2 (%58 of total) in gastrin data set.

3.2 Application to Modelling Tools

In this section, we illustrate the importance of noise removal in the gene expression data and performance of the noise reduction (NR) algorithm on predictive abilities of several modelling tools such as network component analysis (NCA), principal component analysis (PCA) and k-means clustering.

3.3 Network Component Analysis

NCA is a computational method applied on gene expression data to reconstruct the activity profiles of important transcription factors involved in the gene regulatory network. We demonstrated the performance of the NR algorithm on prediction ability of NCA by comparing the results of NCA applied on original noisy data (N) and data treated with NR algorithm (NR). The NCA predicted AP1 activity from NR treated data (in gastrin regulated system) showed peak activation at 4 hours (6^{th} time sample) and is in accordance with the previous experiments [3, 12](Figure 2). In contrast, AP1 activity from original noisy data displayed peak activation at 2 (5^{th} time sample) hours. Another measure to assess the performance of the NR algorithm is the similar reconstructed activity profiles from two independent measurements (replicates, see Figure 2). The similarity measure considered here is the Pearson correlation coefficient between profiles from two replicates. We computed the Pearson correlation for all the transcription factor activity profiles from the noisy replicate data sets and the replicate data sets treated with NR algorithm (Table 1). NCA predicted similar activity profiles (from two replicates) for the noise treated data sets. In contrast NCA predicted dissimilar profiles for the original noisy replicate data sets. The low correlation (0.07) between the AP1 activities predicted from two noisy EGF replicates is increased by a factor of 7.5 (to 0.55) after applying the NR algorithm. Similar results are found for CEBPG, TFAP2A, USF1, NKX21 and PAX6. This correlation analysis depicted that at least 80% of all the predicted activity profiles of transcription factors displayed improved correlations (between replicates) after treating the data with our NR algorithm in all three data sets gastrin, EGF and HRG (Figure 3).

3.4 Principal Component Analysis

PCA is a multi variate data analysis method to reduce the dimensionality of data (i.e. number of variables) while maximizing the variance. We applied PCA on

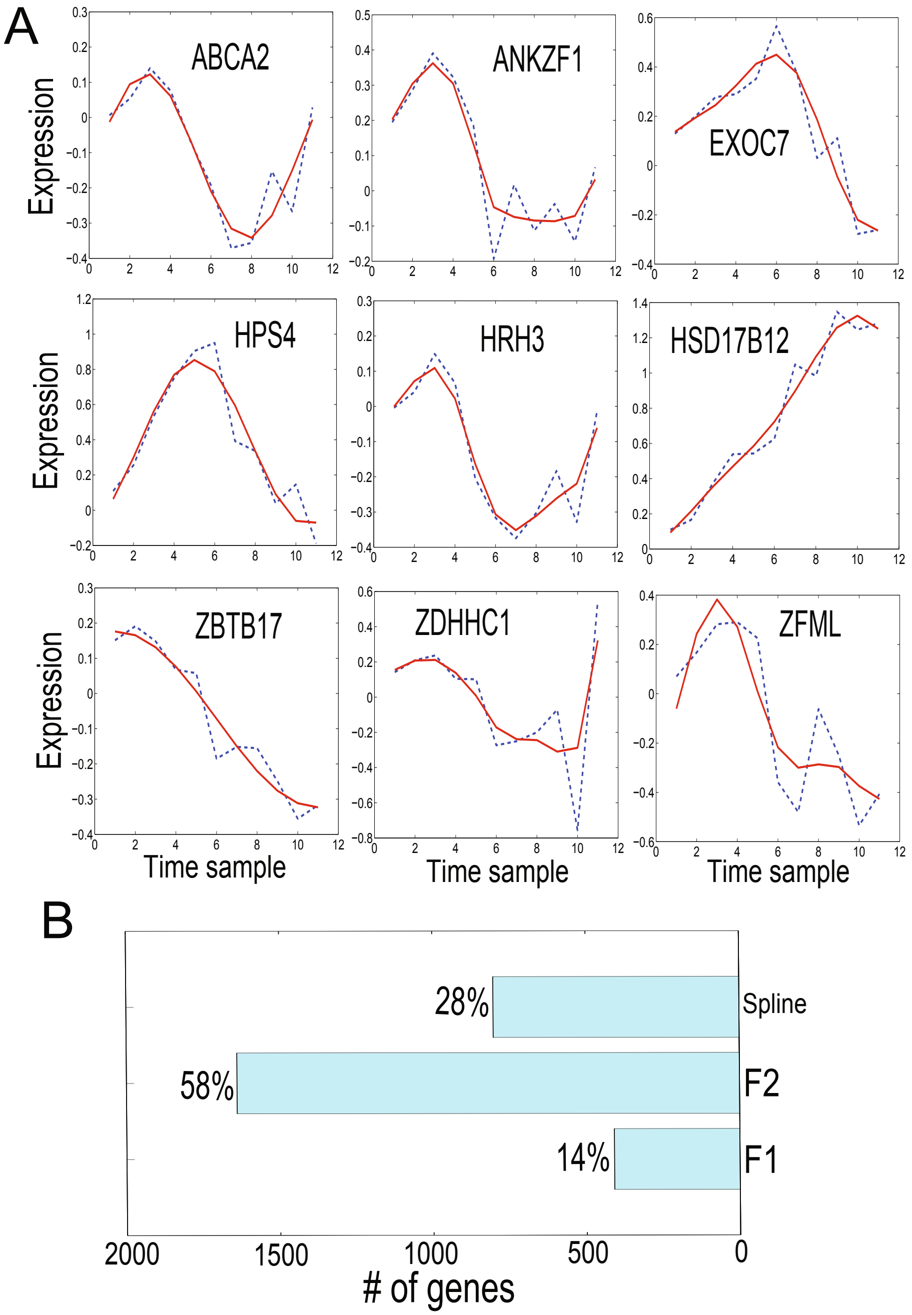

Fig. 1. (A) Fourier functions and smoothing spline approximations applied to gene expression data. Dotted blue lines denote original noisy data and solid red lines denote the fitted approximation. (B) Distribution of number of genes approximated by Fourier 1 (F1), Fourier 2 (F2) and smoothing spline.

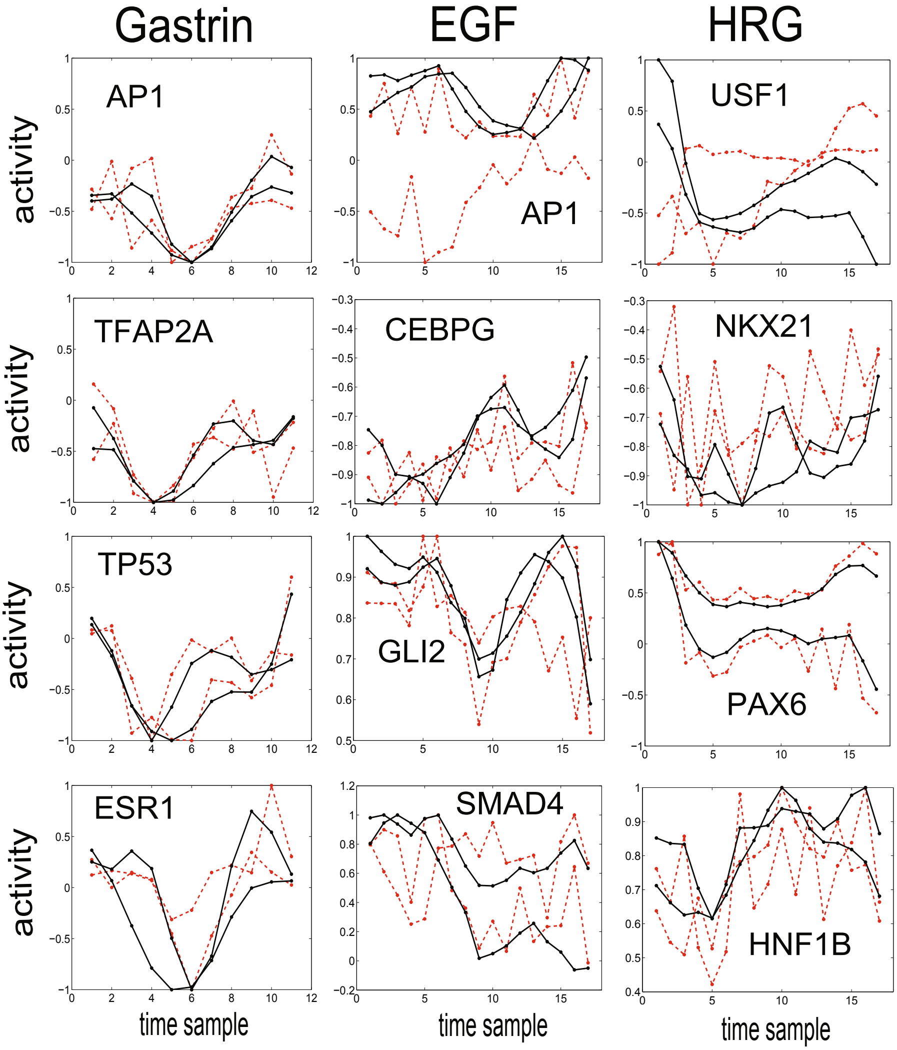

Fig. 2. NCA predicted activity profiles of TFs in three data sets (Gastrin, EGF and HRG) from two replicates. Black solid lines represent the noise reduced replicates and the dotted red lines represent the original noisy replicates.

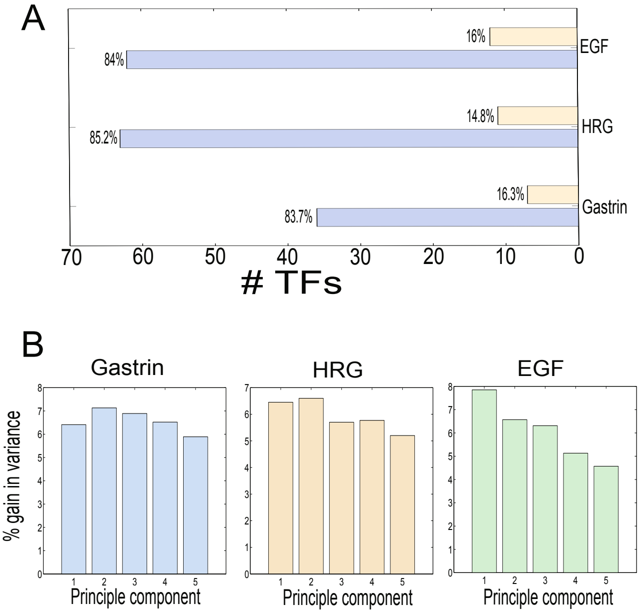

Fig. 3. (A) NCA results: Distribution of number of TFs with improved predictions (long bar) and unimproved (short bar) in three data sets. (B) PCA results: Gain in cumulative variance (difference in variance between noisy (N) and noise reduced data (NR)) associated with principal components in percent.

Table 1. Pearson correlation coefficient of the NCA predicted activity profiles of TFs between two replicates of the noise (N) and the noise reduced (NR) data sets

Data set	**Transcription**	**Noise data**	**Treated data**	**Improvement**
	factor	**(N)**	**NR**	**(fold)**
Gastrin	**AP1**	0.395	0.9	2.27
	TFAP2A	0.473	0.738	1.56
	TP53	0.441	0.648	1.47
	ESR1	0.528	0.742	1.4
EGF	**AP1**	0.073	0.552	7.52
	CEBPG	0.03	0.645	21.37
	GLI2	0.125	0.795	6.32
	SMAD4	0.143	0.827	5.77
HRG	**USF1**	0.163	0.863	5.29
	NKX21	0.108	0.451	4.15
	PAX6	0.253	0.54	2.13
	HNF1B	0.361	0.69	1.9

the original noisy data (N) and NR algorithm treated data (NR) and compared the results. PCA reduced the NR treated data with at least 5% more cumulative variance associated with up to five principal components (PCs) than the noisy data (Figure 3 and Table 2) and this gain in variance is even more with fewer PCs. Similar results are obtained with all the three data sets (each with two replicates), which demonstrates the improvement of PCA analysis.

Table 2. Results of PCA applied on noisy (N) and noise reduced (NR) data sets on three systems. The values in the table represent the cumulative variance associated with principal components (PCs).

Principal	**Gastrin**		**EGF**		**HRG**	
component	**N**	**NR**	**N**	**NR**	**N**	**NR**
1	60.9	67.3	72.3	80.1	61.1	67.5
2	74.6	81.7	82.6	89.2	75.6	82.2
3	86.9	93.7	88.3	94.6	84.3	90.0
4	90.1	96.6	91.5	96.6	89.9	95.6
5	92.1	98.0	93.3	97.9	92.6	97.8

3.5 K-means Clustering

K-means clustering is a powerful technique often employed in gene expression analysis for elucidating a variety of biological inferences such as shared regulation

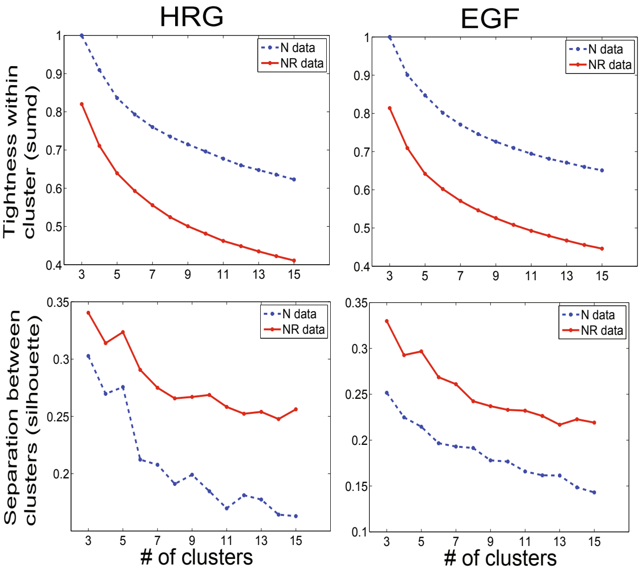

Fig. 4. K-means clustering: Statistical measures of sumd and silhouette are compared between noisy (N) and noise reduced(NR) data sets

or particular biological processes [2]. The k-means clustering applied in this analysis is generally sensitive to noisy data [11]. The Tightness within the cluster measure (sumd) in NR treated data is low compared to the noisy data for all cases of 3 to 15 clusters (Figure 4) in both EGF and HRG data sets. Smaller values of sumd corresponds to tighter clusters. This measure diminishes gradually from 3 clusters to 15 clusters (Figure 4) in both EGF and HRG systems. Another evaluation measure of silhouette is found to be higher in NR treated data sets (NR) than noise data (N). This indicates that NR treated data provides a clear separation of clusters in all 3 to 15 cluster cases, thus improving the performance of k-means clustering.

4 Conclusion

Noise is common in microarray gene expression data and reducing it is essential prior to the inference of knowledge from modelling which utilizes this data. The proposed NR algorithm fits the expression data with either Fourier 1, 2 or smoothing spline based on the temporal pattern and thus minimizes the noise. The performance of the algorithm is demonstrated based on improved prediction capabilities of several modelling tools. NCA reconstructed the very similar activity profiles of transcription factors from two replicate data sets after treating with the NR algorithm. PCA reduced the data with more cumulative variance in the NR treated data than noisy data. K-means clustering also performed better (in terms of statistical measures) with NR treated data.

References

[1] Bar-Joseph, Z., Gitter, A., Simon, I.: Studying and modelling dynamic biological processes using time-series gene expression data. Nat. Rev. Genet. 13(8), 552–564 (2012)

[2] D'haeseleer, P.: How does gene expression clustering work. Nat. Biotech. 23(12), 1499–1501 (2005), http://dx.doi.org/10.1038/nbt1205-1499

[3] Guo, Y.S., Cheng, J.Z., Jin, G.F., Gutkind, J.S., Hellmich, M.R., Townsend, C.M.: Gastrin stimulates cyclooxygenase-2 expression in intestinal epithelial cells through multiple signaling pathways: evidence for involvement of erk5 kinase transactivation of the epidermal growth factor. Journal of Biological Chemistry 277(50), 48755–48763 (2002)

[4] Hardin, J., Wilson, J.: A note on oligonucleotide expression values not being normally distributed. Biostatistics 10(3), 446–450 (2009), http://biostatistics.oxfordjournals.org/content/10/3/446.abstract

[5] Lewin, A., Bochkina, N., Richardson, S.: Fully bayesian mixture model for differential gene expression: simulations and model checks. Statistical Applications in Genetics and Molecular Biology 6 (2007)

[6] Liao, J.C., Boscolo, R., Yang, Y.L., Tran, L.M., Sabatti, C., Roychowdhury, V.P.: Network component analysis: reconstruction of regulatory signals in biological systems. Proc. Natl. Acad. Sci. U S A 100(26), 15522–15527 (2003)

[7] Posekany, A., Felsenstein, K., Sykacek, P.: Biological assessment of robust noise models in microarray data analysis. Bioinformatics (2011), `http://bioinformatics.oxfordjournals.org/content/early/2011/01/19/bioinformatics.btr018.abstract`

[8] Raychaudhuri, S., Stuart, J.M., Altman, R.B.: Principal components analysis to summarize microarray experiments: Application to sporulation time series. In: Pac. Symp. Biocomput., pp. 452–463 (2000)

[9] Ringner, M.: What is principal component analysis. Nat. Biotech. 26(3), 303–304 (2008), `http://dx.doi.org/10.1038/nbt0308-303`

[10] Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 20, 53–65 (1987), `http://www.sciencedirect.com/science/article/pii/0377042787901257`

[11] Sloutsky, R., Jimenez, N., Swamidass, S.J., Naegle, K.M.: Accounting for noise when clustering biological data. Brief Bioinform. (October 2012), `http://dx.doi.org/10.1093/bib/bbs057`

[12] Subramaniam, D., Ramalingam, S., May, R., Dieckgraefe, B.K., Berg, D.E., Pothoulakis, C., Houchen, C.W., Wang, T.C., Anant, S.: Gastrin-mediated interleukin-8 and cyclooxygenase-2 gene expression: Differential transcriptional and posttranscriptional mechanisms. Gastroenterology 134(4), 1070–1082 (2008)

[13] Tang, V., Yan, H.: Noise reduction in microarray gene expression data based on spectral analysis. International Journal of Machine Learning and Cybernetics 3, 51–57 (2012), `http://dx.doi.org/10.1007/s13042-011-0039-7`

[14] Tu, Y., Stolovitzky, G., Klein, U.: Quantitative noise analysis for gene expression microarray experiments. Proceedings of the National Academy of Sciences 99(22), 14031–14036 (2002), `http://www.pnas.org/content/99/22/14031.abstract`

[15] Warren Liao, T.: Clustering of time series data-a survey. Pattern Recogn. 38(11), 1857–1874 (2005), `http://dx.doi.org/10.1016/j.patcog.2005.01.025`