18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

Clustering Sleep Deprivation Effects On The Brain Of Drosophila Melanogaster

Loh W.P.[1], Y. Abu Hasan[1] and A.Talib[2]
[1] *School of Mathematical Sciences, Universiti Sains Malaysia, 11800 USM, Pulau Pinang*
[2] *School of Distance Education, Universiti Sains Malaysia, 11800 USM, Pulau Pinang*
Email: anita@usm.my

In almost all successful time-series clustering, feature selection methods are essential. Though lots of feature selection algorithms have being developed from time to time, issues concerning partially unclean data lead to controversies. In this paper, we introduce the idea of selecting interquartile range value (IQR) to set the minimum absolute expression change filtering boundary. The Short Time-Series Expression Miner (STEM) tool is employed to ensure that reliable data which pass the boundary support the STEM clustering method analyses. The main aim of this study is to access as to whether the idea of IQR value setting could ideally compress data and at the same time retrieve the most optimum information. Besides, analysis of clustering profiles generated enable good visualization effect. This is particularly needed in conditional studies of the effect and causality relations as determinant factors to judge the level of correlations among clusters. Our analysis is implemented on a fruit-fly species (*Drosophila melanogaster*) expression data available from GEO Datasets. The study data consist of 14010 gene profiles recorded in four time points; which subdivided further into three major conditions: unperturbed, perturbed and active.

Keywords: *Time Series Clustering, Visualization, Drosophila melanogaster*

1. INTRODUCTION

For some time now static microarray data analysis is an active field of study. Lately, the study is extended to dynamical microarray. Time-series expression data are mostly explored from the aspect of clustering analysis. Such clustering effort basically attempts to group data according to similar properties. Unlike static data, temporal microarray expression concerns an additional time point parameter to be considered. The majority of microarray clustering initiatives are based on the biological nature exhibited by organisms (Krebs et al., 2007).

In most successful time-series expression clustering, feature selection efforts play a significant role. The feature selection strategy is responsible in screening for reliable data that is free from noise. Several studies have in fact emphasized on feature selection technique in static microarrays (Bar-Joseph et al., 2003;Vingron 2001). Though there are numerous developed statistical techniques in microarray feature selections, few have made extensive reports on its impact in the dynamical microarray clustering.

How expression data is selected, filtered or compressed has important impact on clustering results. Partially unclean data brings upon cluster biasness while over clean data causes a great loss of information. The problem at this point is what are the most optimum criteria to discard data and maintaining the most reliable data for further study? In the novel STEM clustering method, a default value 1 is set as minimum absolute expression change parameter. Microarray data which expresses above the value will be subjected to subsequent analysis. This technique was carried out unambiguously by Ernst et al. (2005).

In this paper, we present the idea of tuning the minimum absolute expression change parameter of filtering boundary for STEM clustering method. The idea is performed and validated on the fruit-fly *Drosophila melanogaster* species expression data, which is publicly available from GEO Datasets (http://www.ncbi.nlm.ni.gov/geo/). This data is particularly chosen since the expression values of *Drosophila* show minimal differences under different experimental conditions and thus indicate minimal understanding. We primarily cluster expressions based on the default value of minimum absolute expression change facilitated by STEM tool and subsequently tuned this value by IQR (interquartile range) measures for a better effect.

2. RELATED WORK

In the efforts to extract reliable data, the technique of discretisation is applied by Friedman et al. (2000). It this effort, the author has simultaneously measured expression levels of thousands of genes. Thereby, models which are able to represent stochastic process and noisy data can be observed. However, the drawback of his technique is that the attempt is complicated and requires tedious work to be carried out. Diverse statistical techniques are adopted in feature selection of microarrays, including Chi-squared, F- ratio and T-test. Low level analysis is carried out in the studies of feature selection, data normalization and evaluation of expression indexes (Schadt et al., 2001a, b). The author presented algorithms to compute feature selection and normalizing two or more arrays. This idea could reduce replication variations in microarrays.

In earlier study of dynamical microarray, data organization and visualisation effect is employed (Ben-Dor et al.,1999; Tamayo et al.1999; Sharan & Shamir 2000). The effect of visualisation suggests appropriate clustering algorithm. Ernst et al. (2005), however, initiated the development of specific algorithm that accounts for short temporal microarrays. Strauch et al. (2007) employed the technique of two-step clustering process and carried out permutation analysis to evaluate the significances. A full space clustering method of multiple sclerosis (MS) is introduced by Jiang et al. (2004) while Zhao et al. (2005) utilised Tricluster algorithm application into cell cycle dataset. Both strategies search for gene groups that exhibit certain patterns across samples and time-series. Although, various existing ideas of problem-based clustering have been reported, they hardly reveal the impact of feature selection on results of cluster. .

3. IMPLEMENTATION

We initially give a brief discussion about the *Drosophila melanogaster* expression raw data and the experimental design of data collection. This could give an idea on the sorts of information to be extracted via performing cluster analysis. This is followed by clustering implementations whereby STEM clustering approach is performed and subsequently clustering strategy posterior to our idea of tuning value of minimum absolute expression change based on the genes interquartile range expression value. Finally, evaluations on the efficiencies our implemented technique is demonstrated.

3.1. Datasets

The study data involved in this study originates from microarray temporal expression data of *Drosophila melanogaster* species. This data records array-based short time-series, comprising a number of 14014 genotypes in four time points: 0, 2, 4 and 6 hours. This experimental expression values are carried out under six major conditional periods; normal controlled brain sleep, brain consolidated sleep, brain deprive, brain sleep, brain normal wake and brain during active period stimulation (APS). The normal sleep period and normal consolidated sleep period act as control experiments. Effects of external perturbations (brain deprivation) on the species during normal sleep period are studied to analyze the stimulated prolong wakening effects to the brain of *Drosophila*. Another aspect of analysis is the perturbation effect during the species active period stimulated (APS) response. Each of the conditions involves three-replication records. From these volumes of data, the target is to investigate as to whether sleep deprived condition has affected the brain of *Drosophila melanogaster* species. Out of the experimental studies, changes were accumulated during prolonged wakefulness and sleeping durations in array-based expressions.

3.2. Methodology

The Short Time-Series Expression Miner (STEM) tool, which is specifically designed for short temporal data (< 8 time points), is employed for this study. We study the expression changes effects under six experimental conditions, mainly from three aspects: normal sleep period, sleep deprivation and active period effects due to external perturbations. The correlations among the six conditions are also examined.

We have basically categorized the experimental conditions into perturbed (brainZT_0hr, brainDeprive_2hr, brain_Deprive_4hr,brain_Deprive_6hr), unperturbed (brain_0hr, brainSleep_2hr, brainSleep_4hr, brainSleep_6hr) and active periods (brainWake_4hr, brain_APS). These three categories are recorded under durations of 0, 2, 4 and 6 hours in three replications respectively. In the former phase, the experimental data undergo separated filtration and transformation efforts. Here, we tested on minimum absolute expression change. Gene expressions beyond this particular filtering boundary will be discarded while the remaining passes the filter. The passing expressions are assumed to be highly distinguished among the rests. The effort retrieves qualitative and informative data for further study. In this particular stage, replications significances and existing outliers are identified. At the same time, P-values evaluation using data permutations are also evaluated to ensure that probable existence of false positives are screened. Similar strategy is repeated by our idea to select the interquartile range (IQR) expression in each conditional time points to reset the minimum absolute expression change parameter.

Subsequent stage involves clustering analysis. The selected files of both strategies; default setting (partial-cleaned) and IQR setting (cleaned) are subjected to clustering whereby STEM clustering approach is practiced to group data according to similar profiles. Cluster profiles shall appear in the main experimental conditions; unperturbed, perturbed and active period. These conditions are further study in terms of whether different conditions have effect on the brain of *Drosophila*. Informative comparisons involve correlation analyses among three main categorical conditions. This is studied accordingly from time to time sequences to observe the causal and effect correlations. The latter evaluates the reliability of our IQR setting idea during data cleaning phase compared to default filter value.

4. RESULTS AND DISCUSSION

In the temporal data preparation phase, several data files undergo filtrations, according to their respective experimental conditions. The initial approach is filtration based on the default setting of minimum absolute expression change, 1. A subsequent attempt is carried out by tuning the value of minimum absolute expression change according to the IQR of genes expressions based on their respective experimental conditions. The former maintains a maximum absolute difference changes between temporal values of above the default parameter. This effort discards merely 0.03 % of genes leaving behind 14010 out of 14014 genes in all experimental conditions except for brainWake_4hr, which is about 8 % filtration.

In the latter, from brain_0hr, brainZT_0hr, brainSleep_2hr, brainSleep_4hr and brainSleep_6hr conditional files, 3% of genes are filtered, maintaining 13621 genes. Meanwhile the brainDeprive_2hr, brainDeprive_4hr and brainWake_4hr recorded a number of 2 %, 4 % and 9 % filtrations respectively. The remaining discards 1 % of overall gene expressions. The average percentage of filtering according to the default setting is 0.9 % while the IQR setting filters an average of 3 %. At this phase, the former technique of filtration by default value is considered as partial-cleaned as it discards very few data but maintains values that show little variations among study data. Thus, the filtered data might therefore be in a partially unclean state.

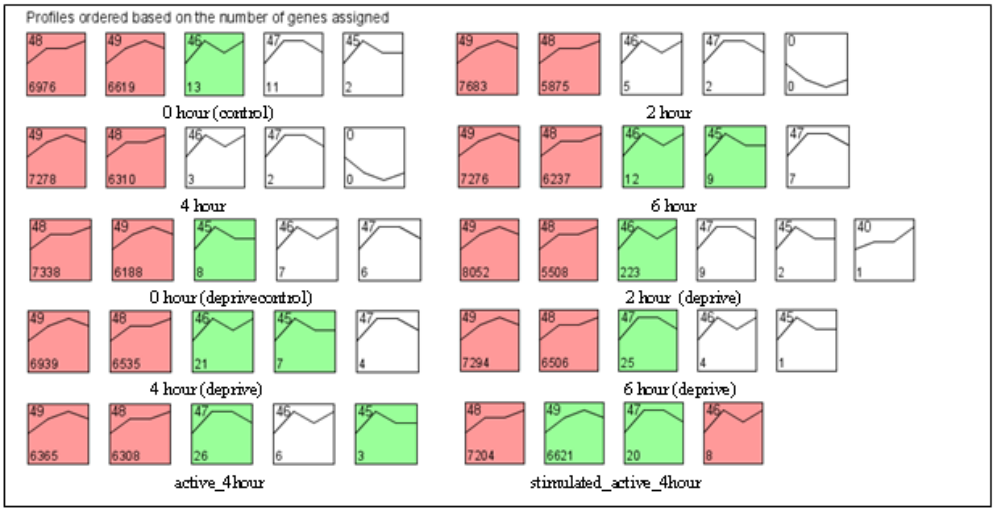

Figure 1. Cluster profiles of unperturbed, perturbed and active periods of Drosophila at different time points based on IQR setting filtration

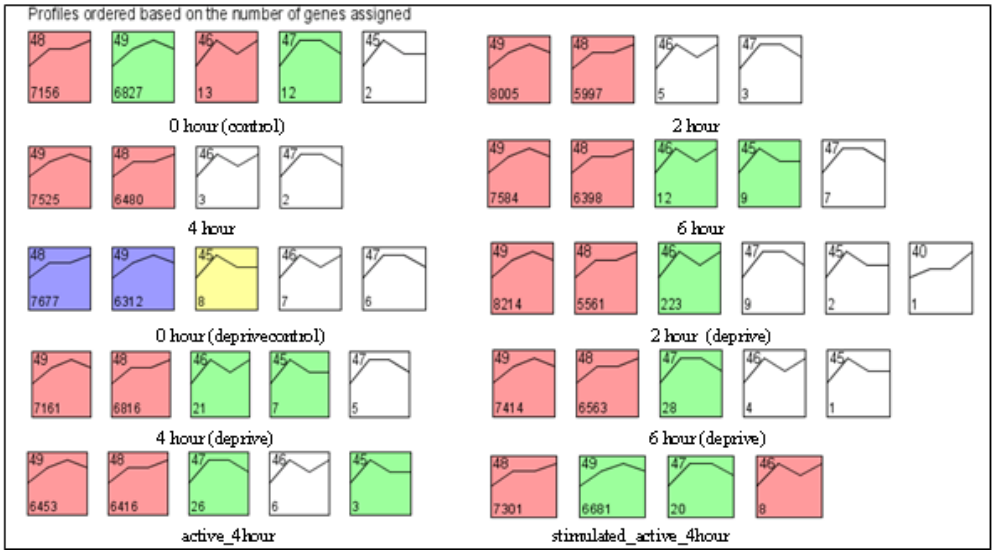

Figure 2. Cluster profiles of unperturbed, perturbed and active periods of Drosophila at different time points based on default value filtration

Both strategies of filtrations are subjected to clustering analysis for further evaluations whereby STEM clustering method is employed. The objective of the STEM clustering is to collect and group expression values according to similar exhibited patterns into profiles. Figure 1 depicts clusters of assigned gene numbers according to distinct characteristics expressed by genes of all experimental conditions at four time points using the idea of IQR setting, while Figure 2 shows similar clustering effort using the default value of minimum absolute expression change.

A small explanation is needed regarding the figures. The boxes visualise different expression profiles; coloured box denote significant number of genes assigned to a particular profile whilst uncoloured represent insignificancies. Statistically significant profiles that are similar form a cluster of profiles, and have the same colour. Figures on the upper left corner of the profile box represents profile ID while the lower left indicates number of genes clustered into the particular profile. Profiles originate from the same cluster observe similar graphical patterns.

Based on the figures, there is an approximately 15 % out of 27 cluster-profile sets difference generated by both attempts. Profiles numbered 48 to 49 are most significant as these four profiles appear in all major experimental conditions: unperturbed, perturbed and active. A further study of both attempts: IQR setting and default value filter is carried out in unperturbed condition whereby different time points are compared to the control profile as to whether the time parameter has effect on the original expressions. Similar effort is

applied to the perturbed conditions and subsequently a comparison of perturbed and unperturbed sleep duration. This aims to view whether these experimental conditions affect the normal active activities of *Drosophila*. Overall comparisons are illustrated in Figure 3 (by IQR setting) and Figure 4 (by default value).

Based on all cluster comparison analyses, the number of intersecting profiles and correlations among compared files are averaged and summarized in Table 1. Profile number 48 and 49 appear to be significant in all experimented conditions, showing high correlations among unperturbed vs. unperturbed, perturbed vs. perturbed, unperturbed vs. perturbed and wake vs. stimulated. Basically the correlation values of cluster intersections ranges from 0.73 (highly correlated) to 1.00 (perfect) using the two approaches. From the results shown in Table 2, majority number of profile intersections and correlation values between IQR and default setting strategy indicate similar results. Nevertheless, the weakness of point default value filtration is that it fails to correlate between compared unperturbed_0 hour vs. 4 hour and unperturbed_0 hour vs. brain wake period_4 hour. Besides, perturbed 0 hour vs. perturbed 2 hour shows a single insignificant intersection. All compared conditions using IQR setting indicate strong correlations. Though there is hardly a difference between overall perturbed via unperturbed conditions, if we further examine the correlations of IQR tuning method from the aspect of time change effect, there is a slight difference in the observed pattern, as illustrated in Figure 5. In the unperturbed vs. unperturbed, the correlation values show a sharp increase from 0.79 to 1.00. In the case of perturbed vs. perturbed indicate a drop from 0.75 to 0.73 followed by an increase until 1.00, while the unperturbed vs. perturbed condition shows the reverse; a decrease of correlation from 1.00 to 0.79. We therefore make an analysis based on two-hourly time steps beginning from time point 0 hour until 6 hours.

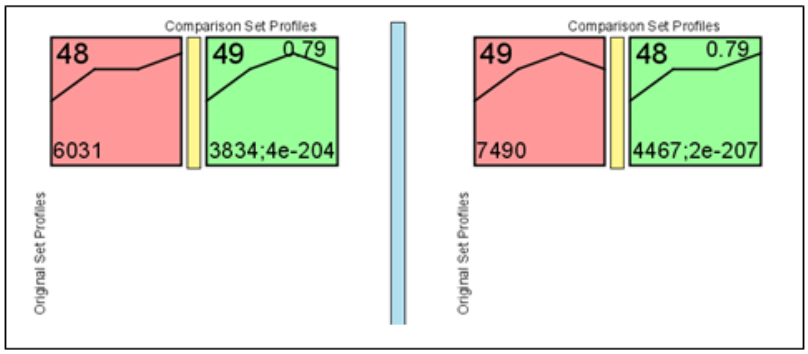

Figure 3. Cluster profiles comparisons between overall perturbed and unperturbed conditions using the IQR setting. Profile box to the right of yellow bar illustrate intersection clusters of profiles on the left. Values indicated on top right corner of profile boxes represent correlation values between compared profiles

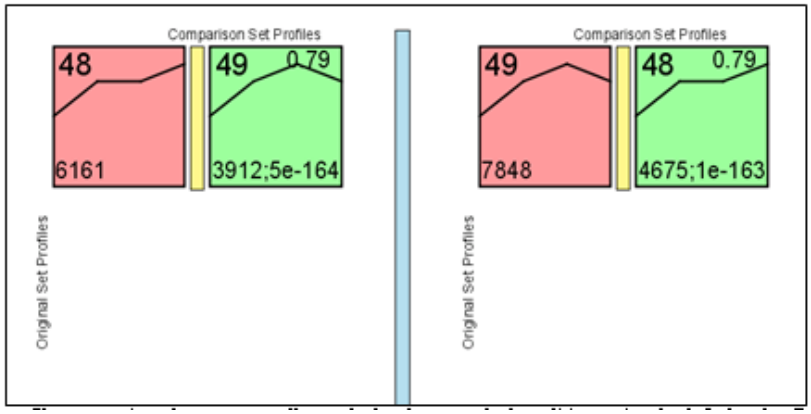

Figure 4. Cluster profiles comparisons between overall perturbed and unperturbed conditions using the default value. Profile box to the right of yellow bar illustrate intersection clusters of profiles on the left. Values indicated on top right corner of profile boxes represent correlation values between compared profiles

Table 1. Number of intersections and average correlations values among compared conditions by IQR setting idea (top right diagonal in yellow) and default value (lower left diagonal in blue).

		Number of cluster profile intersections (correlation value)									
		Unperturbed time (hr)				*Perturbed time (hr)*				*Wake*	APS
Conditions (hr)		0	2	4	6	0	2	4	6	4	4
Unperturbed	0		2 (0.79)	2 (0.79)	3 (0.76)	2 (1.00)				1 (0.79)	2 (1.00)
	2	*2 (0.79)*		2 (1.00)	2 (1.00)		3 (0.97)				
	4	-	*2 (1.00)*		2 (1.00)			2 (0.79)		2 (1.00)	
	6	*4 (0.71)*	*2 (1.00)*	*3 (0.98)*					2 (0.79)		
Perturbed	0	*2 (1.00)*					3 (0.75)	2 (1.00)	2 (1.00)	1 (0.79)	3 (0.95)
	2		*3 (0.99)*			** 4 (0.71)*		3 (0.73)	3 (0.73)		
	4			*2 (0.79)*		*2 (1.00)*	*3 (0.73)*		2 (1.00)		2 (1.00)
	6				*3 (0.82)*	*2 (1.00)*	*3 (0.73)*	*2 (1.00)*			
Active	wake 4	-		*2 (1.00)*		1 (0.79)					2 (0.79)
	APS 4	*2 (1.00)*				*3 (0.95)*		*2 (1.00)*		*2 (0.79)*	

* exist 1 insignificant intersection

The increasing correlations as time duration of sleep prolong up to 6 hours implies effect and causality relations that the longer the sleep period of *Drosophila*, the more sufficient time for the species to express itself till it manages to gain perfect correlation between time points.

The decrease of correlation in unperturbed vs. perturbed, however, suggests that the perfect correlation of unperturbed and perturbed gradually weaken. This phenomenon concluded that the perturbed sleeping hours result in diverse brain cluster expressions comparing to the habitual sleeping characteristics of *Drosophila*.

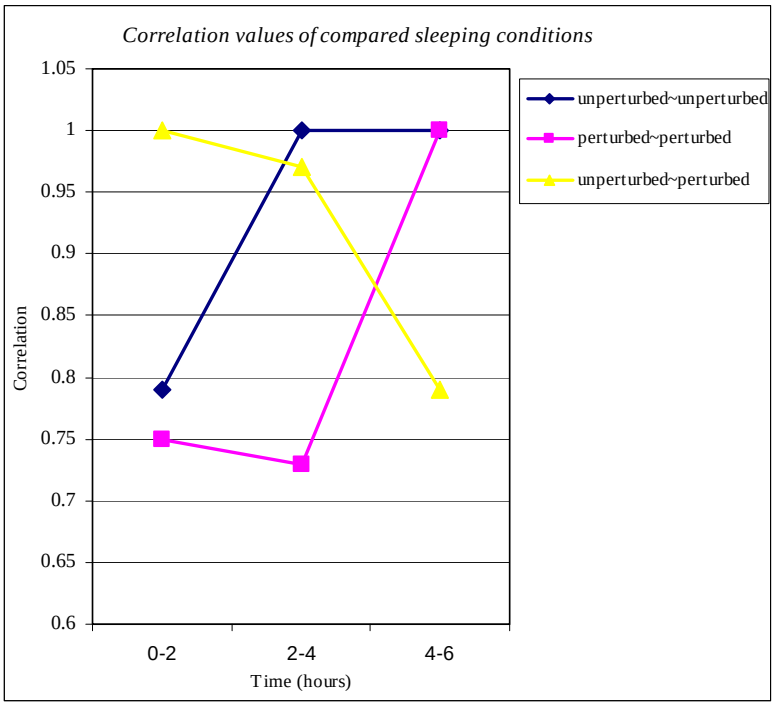

Figure 5. Graphs of three experimented condtitions of sleep duration changes in Drosophila

5. CONCLUSIONS

In pre-processing microarray temporal data for STEM clustering analysis, the technique of discarding unreliable data is a big issue. The existence of unclean data will have severe effect on the results of the clustering study. In this paper, we present the idea of cleaning microarray by tuning the value of minimum absolute expression change according to IQR of microarray. This effort is carried out separately according to experimental conditions and thus allows further STEM clustering evaluations and comparisons among clusters obtained. The technique of IQR filter boundary tuning is that it ensures cleaner data to pass filters and enables more significant cluster comparisons between time points. To the best of our knowledge, none of the literature studies has demonstrated the resetting of minimum absolute expression change value according to IQR to allow the implementation of STEM clustering method. This idea is an advantage as it is more sensitive in cases of small variations in gene expressions. The IQR setting idea is demonstrated using *Drosophila Melanogaster* species temporal expression data. We manage to show that clustering based on the IQR filtered data by respective experimental conditions is in good agreement with the default setting. But, in extended comparisons of subsequent time points, the IQR setting is seen to outperform the default setting. It is more reliable and capable to distinguish characteristics of sleep and deprive sleep durations. For the future study, it will be interesting to implement the process of IQR tuning filtration posterior to STEM clustering. This could direct microarray analysis to main goal of clustering prior to cleaning irrelevant results. The effort could also be extended to forecast expressions in the future time points. It shall be rewarding to incorporate diverse types of temporal microarray, for instance clinical tests microarrays of several medical treatment conditions and microarrays of longer time points, or higher expressions involving extreme values for deeper evaluations.

ACKNOWLEDGMENTS

We would like to acknowledge the Fundamental Research Grant Scheme for supporting this work.

REFERENCES

Bar-Joseph, Z., Gerber, G., Simon, I., Gifford, D.K. and Jaakkola, T.S. (2003), Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. *PNAS* 100(18), 10146–10151.

Ben-Dor, A., Shamir, R. and Yakhini, Z. (1999), Clustering gene expression patterns. *J. Comput. Biol.* 6(3/4), 281–297.

Ernst, J., Nau, G. J. and Bar-Joseph, Z. (2005), Clustering short time series gene expression data. *Bioinformatics.* 21, i159–i168.

Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000),Using BN to analyze expression data. *J. Comput. Biol.* **7**, 601–620.

Jiang, D., Pei, J., Ramanathan, M., Tang, C. and Zhang, A. (2004), Mining coherent gene clusters from gene sample-time microarray data. *In 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 430-439.

Krebs, O., Golebiewski, M., Kania, R., Mir, S. and Saric, J. (2007),SABIO-RK: A data warehouse for biochemical reactions and their kinetics. *Journal of Integrative Bioinformatics.* 4 (1), 49-58.

Schadt, E. E., Li, C., Su, C. and Wong, W. (2001a), Analyzing high-density oligonucleotide gene expression array data.*Cellular Biochem.* 80, 192–202.

Schadt, E. E., Li C., Ellis, B. and Wong, W. (2001b), Feature extraction and normalization algorithm for high-density oligonucleotide gene expression array data. *Cellular Biochem.* 84(37), 120–125.

Sharan, R. and Shamir, R. (2000), Click: A clustering algorithm for gene expression analysis. *Int. Conf. Intell. Sys. Mol. Biol.* 8, 307–316.

Strauch, M., Supper, J., Spieth, C., Wanke, D. and Kilian, J. (2007), A two-step clustering for 3-D gene expression data reveals the main features of the Arabidopsis stress response. *Journal of Integrative Bioinformatics.* 4 (1), 54.

Tamayo P., Slonim D., Mesirov J., Que, Z. and Golub T. (1999), Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci.* 96 (60), 2907-2912

Vingron, M. (2001), Bioinformatics needs to adopt statistical thinking. *Bioinformatics.* 17, 389–390.

Zhao, L. and Zaki M. J. (2005), Tricluster: An effective algorithm for mining coherent clusters in 3D microarray data. *Proc. of the 2005 ACM SIGKDD International Conference on Management of Data.* 694-705.