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

Using emergent clustering methods to analyse short time series gene expression data from childhood leukemia treated with glucocorticoids

Chaiboonchoe, A , S. Samarasinghe and D. Kulasiri

CfACS – Centre for Advanced Computational Solution, Lincoln University, Christchurch, New Zealand
Email:amphun.chaiboonchoe@lincolnuni.ac.nz

Abstract: Acute lymphoblastic leukemia (ALL) causes the highest number of deaths from cancer in children aged between one and fourteen. The most common treatment for children with ALL is chemotherapy, a cancer treatment that uses drugs to kill cancer cells or stop cell division. The drug and dosage combinations may vary for each child. Unfortunately, chemotherapy treatments may cause serious side effects. Glucocorticoids (GCs) have been used as therapeutic agents for children with ALL for more than 50 years. Common and widely drugs in this class include prednisolone and dexamethasone. Childhood leukemia now has a survival rate of 80% (Pui, Robison, & Look, 2008). The key clinical question is identifying those children who will not respond well to established therapy strategies.

GCs regulate diverse biological processes, for example, metabolism, development, differentiation, cell survival and immunity. GCs induce apoptosis and G1 cell cycle arrest in lymphoid cells. In fact, not much is known about the molecular mechanism of GCs sensitivity and resistance, and GCs-induced apoptotic signal transduction pathways and there are many controversial hypotheses about both genes regulated by GCs and potential molecular mechanism of GCs-induced apoptosis. Therefore, understanding the mechanism of this drug should lead to better prognostic factors (treatment response), more targeted therapies and prevention of side effects.

GCs induced apoptosis have been studied by using microarray technology in vivo and in vitro on samples consisting of GCs treated ALL cell lines, mouse thymocytes and/or ALL patients. However, time series GCs treated childhood ALL datasets are currently extremely limited. DNA microarrays are essential tools for analysis of expression of many genes simultaneously. Gene expression data show the level of activity of several genes under experimental conditions. Genes with similar expression patterns could belong to the same pathway or have similar function. DNA microarray data analysis has been carried out using statistical analysis as well as machine learning and data mining approaches.

There are many microarray analysis tools; this study aims to combine emergent clustering methods to get meaningful biological insights into mechanisms underlying GCs induced apoptosis. In this study, microarray data originated from prednisolone (glucocorticoids) treated childhood ALL samples (Schmidt et al., 2006) (B-linage and T-linage) and collected at 6 and 24 hours after treatment are analysed using four methods: Self-organizing maps (SOMs), Emergent self-organizing maps (ESOM) (Ultsch & Morchen, 2005), the Short Time series Expression Miner (STEM) (Ernst & Bar-Joseph, 2006) and Fuzzy clustering by Local Approximation of MEmbership (FLAME) (Fu & Medico, 2007).

The results revealed intrinsic biological patterns underlying the GCs time series data: there are at least five different gene activities happening during the three time points; GCs-induced apoptotic genes were identified; and genes active at both time points or only at 6 hours or 24 hours were determined. Also, interesting gene clusters with membership in already known pathways were found thereby providing promising candidate gens for further inferring GCs induced apoptotic gene regulatory networks.

Keywords: *Childhood Leukemia, Clustering, Emergent self-organizing maps, Glucocorticoids, Short time series gene expression.*

1. INTRODUCTION

Leukemia is a cancer in white blood cells and there are 2 types of childhood leukemia: acute (rapidly growing) or chronic (slowly growing) but most childhood leukemia is acute and can divided into 2 groups: Acute lymphoblastic leukemia (ALL) and Acute myelogenous leukemia (AML). In The United States, about 60% of leukemic children suffer from ALL and about 38% from AML (www.kidhealth.org). ALL can be further divided into 2 subgroups: T-lineage with homogeneous gene expression patterns and B-lineage with many subgroups. About 85% of ALL are B- ALL (www.cancer.org).

Glucocorticoids (GCs) are the most important drug for treating children with ALL. GCs induce apoptosis in malignant lymphoid. Apoptosis is a multi-step and multi-pathway cell death programme which varies with multicellular organisms. Several research groups have studied glucocorticoids-response genes in GCs-induced apoptosis pathway by using gene expression profiling. However, there is controversy among existing hypotheses, and GCs-induced apoptosis mechanism is still an active research area. To understand GCs-induced apoptosis, there is a need to understand it at a genetic level. Microarray technology makes it possible to investigate the expression of thousands of genes in one experiment. Microarray clustering is one approach to gene expression data analysis. Clustering or unsupervised classification approaches use the expression profiles and divide data into highly similar groups without predefined group labels.

Time series can generally be divided into two major groups: short and long time series. More than 80% of the microarray data are short time series containing only eight time points or fewer (Ernst et al., 2005). This leads to the key challenge of how to analyze the limited time series gene expression data in order to maximize the utilization of invaluable data. Although there are techniques for clustering gene expression, more methods are being developed. We select existing clustering methods according to their strength which relate to the nature of gene expression analysis: co-expression, time series and one gene belongs to more than one group (fuzzy).

In this paper, the Self-organizing maps (SOMs), Emergent self-organizing maps (ESOM), the Short Time series Expression Miner (STEM) and fuzzy clustering by Local Approximation of Membership (FLAME) were used to analyze short time series gene expression data extracted from prednisolone (glucocorticoids) treated childhood leukemia patients before treatment, and 6 hours and 24 hours after treatment. The original experiments and basic data analysis can be found in Schmidt et al., (2006). Data are for three T-ALL and ten B-ALL patients.

Objectives of this study are to:

(1) Validate the original authors' results for differentially expressed genes between 0 and 6 hours and extend the analysis to 24 hours.

(2) Identify differentially expressed genes 6 hours after treatment, 24 hours after treatment and those differentially expressed between 6 and 24 hours after treatment and genes that are turned on or off at these time points for both T-ALL and B-ALL and compare the two diseases subtypes. Further, identify GC-induced apoptosis genes that are active at both time points or only at 6 hours or 24 hours after treatment for the differentially expressed gene set.

(3) Extract intrinsic biological patterns underlying the time series data and find meaningful biological knowledge from the combination of four clustering methods: (i) SOMs, (ii) ESOM, (iii) STEM and (iv) FLAME.

2. BACKGROUND

Among emergent clustering methods, we focused on three categories: neural networks, time series and fuzzy clustering. Neural networks are a well-known and widely used clustering method and it has been used in cancer gene expression analysis (Golub et al., 1999). Self-organizing map (SOMs) based neural networks is a nonlinear clustering method with attractive features of data visualization and dimensionality reduction. Recently, evolving self-organizing maps have been developed to allow a map to grow from the predefined map structure which is a restriction of the traditional SOMs; for example, Emergent Self-organizing maps (ESOM) (Ultsch & Morchen, 2005).

Time series expression data can play a vital role in understanding the mechanism of disease progression, the role of genes in the process and modeling gene regulatory networks. Many clustering algorithms have been applied to time series data and more details can be found in Wang et al., (2008). Among available methods, the Short Time series Expression Miner (STEM) (Ernst & Bar-Joseph, 2006) was particularly created to

analyze short time series gene expression data. This method uses the change in direction and magnitude of the inputs with time.

Fuzzy clustering was introduced to overcome crisp clustering where one gene belongs to one cluster. Each gene is assigned a cluster membership which indicates the degree of belonging in each cluster. Well-known methods include Fuzzy C-Means (Bezdek & Ehrlich, 1984) and one recently developed method is Fuzzy clustering by Local Approximations of MEmberships (FLAME) (Fu & Medico, 2007).

This study combines four selected methods (SOMs, ESOM, STEM and FLAME) and the results from this study are used to identify gene clusters responsive to GCs and to further infer gene networks and pathways. Furthermore, The Database for Annotation, Visualization and Integrated Discovery (DAVID) was selected for incorporating known biological knowledge. DAVID is a web-based program for functional annotation and bioinformatics microarray analysis (Huang et al., 2007). DAVID has many functions, of which functional annotation and gene ID conversion are used in this paper.

3. METHODS

3.1 Dataset

This study uses the microarray dataset referred to in the article "Identification of glucocorticoid-response genes in children with acute lymphoblastic leukemia" by Schmidt et al., (2006). Raw data in the format of CEL files and normalized microarray data were obtained online from the Gene Expression Omnibus (GEO). Raw data comprise 13 patients (three T-ALL patients and ten B-ALL patients) and were collected at three time points: 0 hour, 6/8 hours, and 24 hours.

The raw data was then reprocessed starting from normalization and selection of differentially expressed genes, with a same log ratio threshold as in the original article. Specifically, all 39 files were processed and normalised by Robust Multi-array Average (RMA) as in the original study; however, our study not only used R as the original article, we also used RMAExpress and Matlab to calculate gene expression matrix. Furthermore, our study extends the original authors' work to the analysis of 24 hours expression and comparison of 6 hours and 24 hours gene expression patterns.

3.2 Computational Methods

Four methods have been used in this study: SOMs, ESOM, STEM, and FLAME. Each method was selected according to their strengths and they are all user friendly recently developed software that have the potential for further development. They have never been used with time series data (except for STEM). A summary of these methods is given below:

3.2.1 Self-organizing maps are unsupervised neural networks that preserve the exact topology of data space on a two-dimensional grid of neurons. The components of SOMs are nodes or neurons that are normally arranged in a 2-dimensional hexagonal or rectangular grid. Inputs are mapped from high dimensions to this two-dimensional map space. Training starts with assigning random neuron weight vectors with the same dimensions as the input vectors and then using Euclidean distance to calculate distance between an input vector and each of the weight vectors. The weight of the winner neuron (with the smallest distance) and its nearest neighbor neurons are updated after presentation of each input vector (or a batch of input vectors) until the map is converged when weight change is negligible (Samarasinghe, 2006). Peltarion Synapse, a commercial neural networks software, was used and trained map neurons were clustered using ward clustering.

3.2.2 Emergent Self Organizing Maps (ESOM) is an extension of SOMs which allows a map to grow from the initial map. ESOM is arranged in a toroid grid using a large number of neurons than typical SOMs. ESOM-map can be visualized in three forms: U-Matrix (distance-based), P-Matrix (density-based) and U*-Matrix (distance and density based). U-Matrix indicating the distance of each neuron to its nearest neighbours is the same as the U-Matrix in SOMs, so this study uses P-Matrix in order to analyze density of the gene expression data. P-Matrix is a matrix with entries denoting density of neurons in the neighbourhood of a neuron. A neuron with a large P is located in a dense data region while a small P indicates sparse data regions in the data space. Data can be analyzed by using publicly available ESOM software (Databionics ESOM Tool) developed by Ultsch and Morchen (2005). This software is available to download from http://databionic-esom.sourceforge.net/.

3.2.3 The operation of the Short Time series Expression Miner (STEM) algorithm can be divided into 3 major steps: (i) selecting reference model profiles (gene expression patterns) that are constructed before

analysis using all possible gene expression log-ratio increments (from ±1 to ±3, for example) for the 2 time steps of 6 h and 24 h with respect to time 0h, (ii) identifying significant model profiles that match actual gene expression patterns according to p-values by using a permutation method, and (iii) grouping significant profiles. STEM is a Java-based program; STEM version 1.3.4 was used and downloaded from (http://www.cs.cmu.edu/~jernst/stem/).

3.2.4 Fuzzy clustering by Local Approximation of MEmbership (FLAME) algorithm starts by calculating similarities of expression patterns using Pearson correlation, and then creating a connected graph of all K-Nearest Neighbors (KNN) of each expression vector. Each expression pattern is classified into one of 3 objects based on its density: (i) cluster supporting object (CSO) which has higher density than their neighbors, (ii) outlier, which has lower density than its neighbors, and (iii) the 'rest'. In the next step, the Local Approximation of fuzzy membership of the three types of objects is determined and each object is updated by a linear combination of the fuzzy memberships of its nearest neighbors. Finally, the clusters can be contracted based on fuzzy memberships into two categories: one gene to one cluster or one gene to multiple clusters. Clusters determined by FLAME contain gene expression patterns belonging to one of three types of clusters: (i) clusters which include one CSO and its nearest neighbours (ii) outliers cluster and (iii) all CSOs cluster. FLAME is integrated with Gene Expression Data Analysis Studio (GEDAS). (Software freely available to download from http://sourceforge.net/projects/gedas)

Each programme allows the user to adjust parameters. Synapse (SOMs) used 100,000 epochs (number of batch iterations) in training with initial and final learning rates being 0.5 and 0.001, respectively, and 14,000 epochs during clustering starting with 0.1 learning rate that reached 0.001 at completion. An ESOM was trained using online learning of a map of size 50× 82, with 20 training epochs, correlation distance function and linear function for cooling strategy for radius and learning rate. STEM was clustered with 0.7 minimum correlation coefficient and 0.05 significance level. FLAME was used with Pearson Correlation with 10 K-Nearest Neighbors with no threshold.

4. RESULTS AND DISCUSSION

4.1 Validation /extension results and identification of differentially expressed genes

- In the original paper, Schmidt et al., (2006) have combined the data for T-ALL and B-ALL in the analysis and selected differentially expressed genes under two conditions: (i) log ratio of ±0.7 or higher (ii) log ratio of ±1 or higher, for at least 6 out of 13 patients. The authors' results contained 62 probes (49 genes) for early response (6 hours) and 66 probes (55 genes) for late response (24 hours). (A gene is represented by a probeset, or probe for short, that needs to be converted to gene symbol and some probes are repeats). They focused on early response only in subsequent analysis; only 32 probes were used after deleting cell cycle genes. We re-analysed using the same method and criteria and found more probes which include all probes from the origin paper.
- In using their approach, we found that combining B-ALL and T-ALL data compromises the accuracy of selection of differentially expressed genes in both T-ALL and B-ALL. It is possible that the selected differentially expressed genes come entirely from B-ALL patients, as there were only three T-ALL patients. Therefore, we separated the two types of patients and a new set of differentially expressed genes were selected for T-ALL and B-ALL separately for each time point. Criteria used were log ratio of ± 1 or higher for at least 5 out of 10 B-ALL patients and 2 out of 3 T-ALL patients.
- We also analysed data for early response (6 hours) and late response (24 hours), but we added analysis for response between 6 and 24 hours because this can give more information about gene activity at different times. The results are shown in Table 1 which shows the number of differentially expressed genes 6 hours after treatment, between 6 hours and 24 hours and 24 hours after treatment (before and after deleting cell cycle genes). Log-ratio and criteria mentioned in the methods section were used to find differentially expressed genes.
- Our analyses found 237 probes (203 unique probes after removing repeats) for T-ALL for the combined time points and 257 (207 unique probes) for B-ALL and these were combined into one set. The final set contained 380 unique probe sets (30 probes are common to T-ALL and B-ALL, of which three probes were not found in the original paper). These were converted from probe sets ID to gene symbol by using DAVID. Then the cell cycle genes were deleted from this data set (cell cycle gene list was retrieved from KEGG, Cell cycle database and the original article). After deleting cell cycle genes, T-ALL contained 222 probes (172 unique probes) and B-ALL contains 190 probes (155 unique

probes) for the combined time points. The final set had 327 unique probes (304 genes) responsive to GCs (20 probes are common to T-ALL and B-ALL).

Table 1. Differentially expressed genes 6 hours after treatment, between 6 hours and 24 hours and 24 hours after treatment (before and after deleting cell cycle genes).

	0-6 hours				6-24 hours				0-24 hours			
	Before		After		Before		After		Before		After	
	Induced	Repressed	Induced	Repressed	Induced	Repressed	Induced	Repressed	Induced	Repressed	Induced	Repressed
T-ALL	19	10	19	9	59	51	56	49	58	40	56	33
B-ALL	24	23	24	9	16	13	16	9	73	108	71	61

Then the 327 probes (304 genes) were assessed for their time of activation (0h, 6h, and 24 h or in between). These patterns can be classified into five groups with example gene lists as follows:

- Genes differentially expressed (turned on) only at 6 hours or 24 hours: EGR1, SOCS1, IGHM, LYZ, and PIK3IP1.
- Genes turned on from 6 hours to 24 hours: S100A8, ZBTB16, and P2RY14.
- Genes turned on at 6 hours and stay at same expression level at 24 hours: GIMAP7, SLA, FKBP5, ZBTB16, SNF1LK, PFKFB2, EPPK1, WFS1, C6ORF85, TNFSF8, BTNL9, and TFP1.
- Genes turned on at 6 hours and not differentially expressed (turned off) at 24 hours: HES1, S100A12, GNG11, BCL2L11, ZNF24, and 1552230_AT.
- Genes turned off at 6 hours but turned on at 24 hours: STAB1, TNFSF8, KIAA0101, DTL, TYMS, RRM2, PPBP, FEN1, TMSL8, TUBB1, PF4, NRGN, and HBG2.

This information will be used for the study of GCs mediated gene regulatory networks. These five groups could be further verified by a review of the existing research articles.

4.2 Extract intrinsic biological patterns with four emergent clustering methods applied to gene clustering

After identifying and classifying GCs-induced apoptosis genes, further analysis focused on finding similar gene expression patterns which may have similar function. Gene expression data have high dimensionality (large number of genes) with few samples (patients). SOMs cluster genes with similar expression pattern vectors across the time points. We then explored emergent SOMs tools and selected ESOM to analyze data. We present here some results for clustering GCs-induced apoptosis genes from SOMs and ESOM for two typical patients: patient 2 (T-ALL) (shown on left hand side) and patient 13 (B-ALL) (shown on right hand side) in Figure 1 and 2.

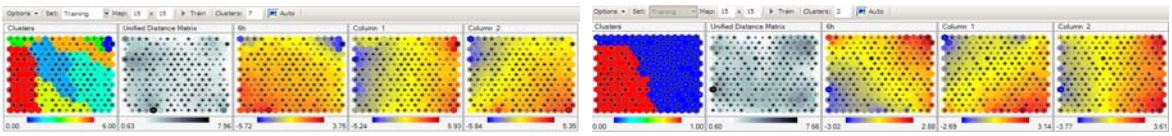

Figure 1. Self-organizing map (maplets from left to right: clusters, U-Matrix, log-ratio for time 0-6h, 6-24h and 0-24h for 2 selected patients).

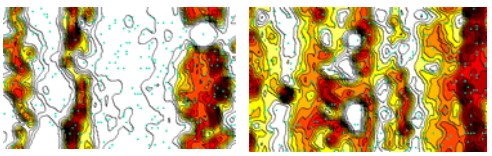

Figure 2. ESOM: P-Matrix (density-based) for the 2 selected patients, green dots represent best match (closet) vector to each data point.

In Figure 1, Lighter areas in the U-Matrix map indicate neurons that are clustered closer to each other and reveal some cluster patterns picked by the algorithm on the leftmost panels for the two patients. The cluster maplet reveals seven clusters for T-ALL and two clusters for B-ALL. Figure 1 indicates that clusters are located not far away from each other (no distinct gap or darker colour in U-Matrix). SOMs are good to use when data analysis starts because they give an overview of data but clustering from SOMs may not be consistent due to the instability of neural gas, the method used to cluster neurons on the trained map. Each training produces a different cluster. However, we can visualize the macro picture of the three scenarios (log-ratio for time 0-6h, 6-24h and 0-24h) for the two patients. The red areas show highly up-regulated genes,

whereas the blue areas show highly down-regulated genes. Maplets clearly indicate that genes are either turned on or off, up-regulated or down-regulated, or stay the same at the three time points. In Figure 2, ESOM provides better visualization of separated clusters when compared with Unified dimension matrix (U-Matrix), but it is still difficult to differentiate clusters in the case of B ALL (patient 13). With the P-Matrix (density-based), the darker colour represents high density, while the lighter colour represents low density. T-ALL patient has 3 distinct clusters while B patient has 3 or more clusters. ESOM has the same problem as SOMs that each training produces a different map. Also, the number of clusters are user-defined, so different users can define different numbers of clusters.

From these two methods we found that the patient 2 (T-ALL) genes are clustered close to each other and the number of genes in clusters highly varied (dense to sparse). The patient 13 (B-ALL) genes also clustered closed to each other but the number of genes in the clusters was quite similar and quite high (dense). We then further analysed the 327 probes using STEM and FLAME for each patient separately, and the results for the same two patients depicted in Figure 1 are shown in Figure 3(a) and 3(b) (patient 2 (T-ALL) (shown on left hand side) and patient 13 (B-ALL) (shown on right hand side)).

a) b)

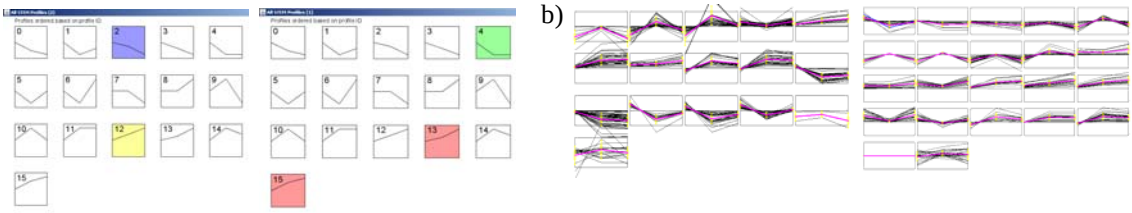

Figure 3. a) STEM: for T-ALL (patient 2) - there are two clusters of significant gene profiles: cluster one (profiles 2) and cluster two (profiles 12). STEM for B-ALL (patient 13)- there are two clusters: cluster one (profile 4) and cluster two (profiles 13, 15). Colours denote significant clusters and same colour belongs to the same cluster, b) FLAME: for T-ALL (patient 2) - line plots of gene expression for each cluster: first 14 clusters consist of a CSO and its nearest neighbours, 15th cluster contains outlier, the last (16th) cluster consists of all CSOs; for patient 13 (B-ALL) first 20 clusters are CSO and neighbour and the last cluster contains just CSOs.

A consistent cluster pattern can be retrieved from STEM and FLAME (Figure 3(a) and 3(b)). STEM (Figure 3(a)) process starts from converting raw expression data for 6 hours and 24 hours into log-ratio with respect to the first time point (0 hour); then using reference profiles as the standard, permutations are used to identify the significant model profiles-patterns that do not happen by chance. Finally, significant profiles are clustered. After clustering genes for each patient, group results for T-ALL and B-ALL were assessed with the result that B-ALL showed two main clusters: cluster one (profiles 2, 3, 4) and cluster two (profiles 8, 10, 11, 12, 13, 14, 15). T-ALL had three clusters: cluster one (profiles 0, 2, 3), cluster two (profiles 8, 12, 13) and cluster three (profiles 11). Results for two selected patients are shown in Figure 3(a): T-ALL (patient 2) on the left hand side and B-ALL (patient 13) on the right hand side. Results from DAVID confirmed these distinct clusters. After processing significant genes from each cluster by DAVID, results indicated that each cluster is involved in a unique KEGG pathway; for example, T-ALL cluster 1 has genes in Type I diabetes mellitus pathway, cluster 2 has genes in Leukocyte tranendothelial migration pathway and cluster 3 has genes in Notch signaling pathway. For B-ALL, cluster 1 has genes in One carbon pool by folate and cluster 2 has genes in Cytokine/cytokine receptor reaction.

FLAME (Figure 3(b)) identified clusters without the need for pre-defined number of clusters while providing fuzzy clustering characteristics. The number of clusters vary for each patient when using FLAME, for example, 14 clusters for patient 2 (T-ALL) and 20 clusters for patient 13 (B-ALL) as shown in Figure 3(b). The mean of number of cluster for T-ALL is 19 clusters and 18.4 clusters for B-ALL. Even though STEM and FLAME used different algorithms and gene expression input vector format (0, 6 and 24 h (STEM) and 0-6h, 6-24h and 0-24h (FLAME)), the patterns of each profile (STEM) and cluster (FLAME) have similar characteristic and similar/same number of genes in each profile/cluster.

The four emergent clustering methods have common limitations which is that they can be used to analyze one sample/patient at the time. Therefore, if combined results or comparisons from the four methods are desired, they have to be done manually. It is very difficult to compare/combine result for SOMs and ESOMs as the distance and density have different ranges, but it is more practical and easier to compare/combine profiles/clusters from STEM and FLAME.

5. CONCLUSIONS

We validated the original paper's results, considering the a priori knowledge that T-ALL and B-ALL are differentiated, and we purposed a new criteria to find GCs-induced apoptosis gene sets and found 327 probes (304 genes) differentially expressed. These genes can classified into five different gene activities happening at the three time points: for example some genes are active at a particular time point while some other genes are active at all times. Then we further analysed the data with four emergent clustering tools. Each computational method provided different insights into short time series data. SOMs and ESOM are visualization methods for high dimensional data projected onto a map and give an overview of how data are organized in term of distance and density; the drawback is clusters from SOMs are not consistent and ESOM requires user to define the number of cluster. The more consistent clusters were obtained from STEM and FLAME. STEM is used to find expression patterns that are statistically significant and have only a very little probability of happening by chance. FLAME can be used to find clusters without predefined groups and to verify clusters from STEM. But FLAME considers all clusters as possible, and does not require statistical analysis.

In conclusion, the four methods (SOMs, ESOM, STEM and FLAME) analyzed GCs treated childhood ALL short time series gene expression profiles and can reveal some underlying temporal gene activation characteristics of GCs-induced apoptosis. Because we need to analyse groups of patients, the challenge is to develop data analysis tools which can simultaneously analyse multiple samples of time series data.

A further interest is to use clusters from STEM and FLAME corresponding to genetic networks and then integrate with information from DAVID to analyze gene regulatory networks. Our future research will take this step forward in looking at the GCs responsive gene networks that control the gene expression patterns behind the scene, which in turn will help identify target genes for better treatment procedures.

ACKNOWLEDGMENTS

This research was funded by a Graduate research scholarship and Postgraduate research funding; attending this conference was funded by Postgraduate conference funding offered by Lincoln University, New Zealand.

REFERENCES

Bezdek, J. C., & Ehrlich, R. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers and geosciences, 10*(2), 191-203.

Ernst, J., & Bar-Joseph, Z. (2006). STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics, 7*(1), 191.

Ernst, J., Nau, G. J., & Bar-Joseph, Z. (2005). Clustering short time series gene expression data. *Bioinformatics, 21*(Suppl 1), S159-S168.

Fu, L., & Medico, E. (2007). FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. *BMC Bioinformatics, 8*, 3.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., et al. (1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science, 286*(5439), 531.

Huang, D. W., Sherman, B. T., Tan, Q., Kir, J., Liu, D., Bryant, D., et al. (2007). DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Research, 35*(Web Server issue), W169.

Kohonen, T. (1997). Exploration of very large databases by self-organizing maps. *Neural Networks, 1997., International Conference on, 1*, PL1-PL6.

Pui, C. H., Robison, L. L., & Look, A. T. (2008). Acute lymphoblastic leukaemia. *The Lancet, 371*(9617), 1030-1043.

Samarasinghe, S. (2006). *Neural Networks for Applied Sciences and Engineering: From Fundamentals to Complex Pattern Recognition*: Auerbach Publications. USA.

Schmidt, S., Rainer, J., Riml, S., Ploner, C., Jesacher, S., Achmuller, C., et al. (2006). Identification of glucocorticoid-response genes in children with acute lymphoblastic leukemia. *Blood, 107*(5), 2061.

Ultsch, A., & Morchen, F. (2005). ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM. *Data Bionics Research Group, University of Marburg, 17*

Wang, X., Wu, M., Li, Z., & Chan, C. (2008). Short time-series microarray analysis: methods and challenges. *BMC Systems Biology, 2*(1), 58.