Integrating Gene Expression Data from Microarrays Using the Self-Organising Map and the Gene Ontology

Ken McGarry*, Mohammad Sarfraz, and John MacIntyre

School of Computing and Technology, University of Sunderland,
St Peters Campus, St Peters Way, SR6 ODD, UK
ken.mcgarry@sunderland.ac.uk

Abstract. The self-organizing map (SOM) is useful within bioinformatics research because of its clustering and visualization capabilities. The SOM is a vector quantization method that reduces the dimensionality of original measurement and visualizes individual tumor sample in a SOM component plane. The data is taken from cDNA microarray experiments on Diffuse Large B-Cell Lymphoma (DLBCL) data set of Alizadeh. The objective is to get the SOM to discover biologically meaningful clusters of genes that are active in this particular form of cancer. Despite their powers of visualization, SOMs cannot provide a full explanation of their structure and composition without further detailed analysis. The only method to have gone someway towards filling this gap is the unified distance matrix or U-matrix technique. This method will be used to provide a better understanding of the nature of discovered gene clusters. We enhance the work of previous researchers by integrating the clustering results with the Gene Ontology for deeper analysis of biological meaning, identification of diversity in gene expression of the DLBCL tumors and reflecting the variations in tumor growth rate.

1 Introduction

Microarrays are an exciting and recent technological breakthrough that has enabled the detailed analysis of cellular activity and condition [1]. Recent work has highlighted how components of metabolic pathways can be identified and how the protein targets of drug treatment can be determined using expression profiles [2]. Microarray technology can deliver an extremely detailed analysis of cellular activity and condition [3]. Recent work has highlighted how components of metabolic pathways can be identified and how the protein targets of drug treatment can be determined using expression profiles for example Alizadeh et al [4] discovered a new sub-class of cancer with implications for clinical treatment. Microarray experiments are producing unprecedented quantities of genome data, the management and analysis of this data is starting to receive greater attention [5]. However, there is no one technique that appears to be superior, either for data management or data analysis.

* Corresponding author.

J.C. Rajapakse, B. Schmidt, and G. Volkert (Eds.): PRIB 2007, LNBI 4774, pp. 206–217, 2007.
© Springer-Verlag Berlin Heidelberg 2007

Microarrays have been used extensively for gene expression analysis and genotyping [6]. Expression analysis seeks to uncover the activity level of certain genes and groups of genes. This is of vital importance in drug discovery where not only are the anticipated effects on the target genes must be confirmed but also for any side-effects on non-target genes must to be monitored. Genotyping seeks to discover and identify many of the mutations within a single gene and can be used for the screening of individuals for particular diseases [7]. Obtaining such information at an early stage will lead to to improved clinical treatment [8].

Microarrays are small glass slides or chips that contain many thousands of genes (strands of DNA) formed as spots which are laid out in a regular grid-like structure. The genes are selected by scientists from gene libraries, and because of their microscopic size they must be located on the glass substrate by automated robotic equipment. The selected genes are usually chosen because they are deemed important for the particular biological process to be investigated. The microarrays are then introduced to the biological samples (DNA that have been labeled by fluorescent materials), which then bind to the original DNA placed on the glass substrate. The microarray image is then scanned and digitised by a laser system. Image processing software is used to reveal the intensity of the fluorescent labels and depending on the type of microarray, their colour. The intensity of the spot is proportional to the level and activity at which the genes are being expressed. Colour, where applicable, is used to identify sample and control populations.

The starting point for any microarray experiment is to define the biological question to answer [9]. For example, a scientist may wish to pursue the hypothesis that a certain number of specific genes are active (up-regulated) in a particular type of cancer and if treated with a particular drug should be inactive (down-regulated). The choice of microarray must also be made, often Affymetrix Gene chips are used in parallel with cDNA microarrays [10].

This paper is concerned with analyzing gene expression data generated from microarrays. We use the self-organizing map (SOM) because of its clustering and visualization capabilities. SOM is a vector quantization method that reduces that simplifies and reduces the dimensionality of original measurement and visualizes individual tumor sample in a SOM component plane. The data is taken from cDNA microarray experiments on Diffuse Large B-Cell Lymphoma (DLBCL) data set of Alizadeh [4]. Diffuse Large B-Cell Lymphoma is the most prevalent lymphoid cancer in adults and accounts for 30-40% of cancers, unfortunately, 50% patients cannot be cured.

The remainder of this is paper is structured as follows; section two discusses the details of the new microarray technology and the problems inherent in the data they generate for machine learning researchers; subsections deal with the characteristics of the SOM that make it suitable for bioinformatic work and the gene ontology system which enables the representation and processing of information about gene products and functions. Section three describes the data, experimental setup and preprocessing issues specific to the microarray data and the experimental results, finally section four presents the conclusions.

2 The Biological Basis of Microarray Technology

Figure 1 shows the internal structure of a typical microarray, the substrates can be glass slides, plastic slides or membranes where the cDNA can be deposited. They have a regular matrix structure, each spot corresponds to a gene sequence. The same gene sequences are usually repeated elsewhere on the chip for reasons of precision and accuracy. Several thousand genes may be placed on an individual chip, the cost of running microarray experiments is directly related to the number of genes per chip.

Although the process of creating microarrays and the analysis of the resultant data is fraught with difficulties their essential operation is relatively straightforward to understand. A set of DNA sequences stored in libraries that correspond to specific genes selected by scientists for their experiment are transferred or *spotted* onto a glass slide by robots. Cell cultures are taken from the patients (a sample and a control) and each is labelled by a fluorescent dye, usually red for the sample population and green for the control population. These cultures are then introduced to the microarray and allowed to bind or *hybridise* with their complementary target cDNA sequences on the chip. The more active a gene is, the more mRNA it should produce and so the intensity and colour of the spot corresponding to that gene ought to appear greater than non-active genes. If the control population is in greater quantity then it will appear green, if the sample population is in greater quantity then it will appear red, if the spot is yellow

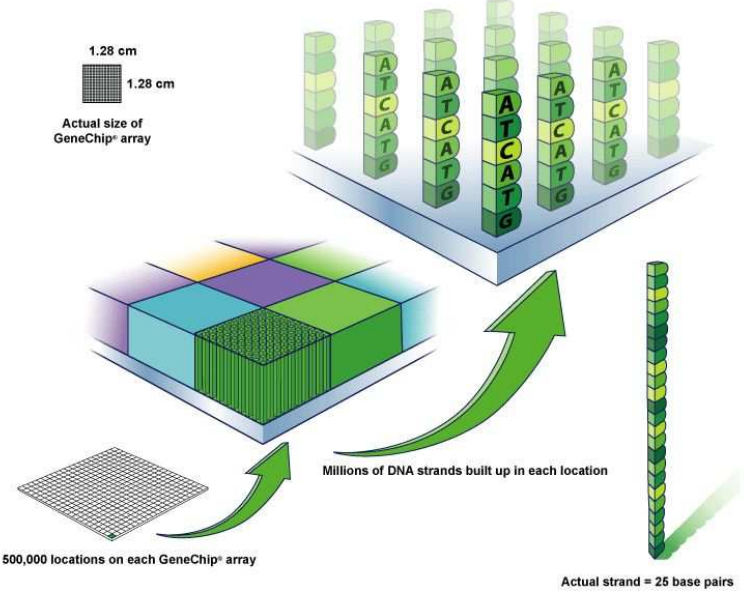

Fig. 1. Each spot is composed of millions of cDNA strands, diagram courtesy of Affymetrix Corporation

then both populations are expressed in equal quantities, if the spot is black then no hybridisation has occurred.

The basic idea behind microarray analysis is to examine the intensities of the spots which is an indirect indication of the level of expression of the genes. The expression levels are often compared against biologically related samples to see which genes are differentially expressed. This can be displayed as a ratio between the sample and control genes, there are disadvantages to using only expression ratios for data analysis. The ratios can help determine important relationships between genes but they also remove information relating to the absolute gene-expression levels. The information pertaining as to wether a gene is up- or -down regulated appears differently when using ratios; i.e. a up- factor of 2 have a value of 2 while those genes that are down-regulated by 2 have a value of -0.5 [11]. Transforming the data using a Log2 base produces a more intuitive range of values, see figure 2. This is a simple way to compare the two channels. Points that are above the diagonal in this plot correspond to genes that have higher expression levels in the sample than in the sample.

Typically, the first and most commonly used technique is to normalise the data, this manipulates the hybridisation intensities to balance them in order to make meaningful comparisons [12]. Normalisation usually needs to be applied because of various problems with experimental bias such as background intensities of the microarrays are not uniform, also differences can occur between pen-tips/print-tips, or blocks. These must be compensated for by normalizsation, hopefully the information will be available to normalise each block separately. Normalisation of data means that weaker signals are amplified, this could

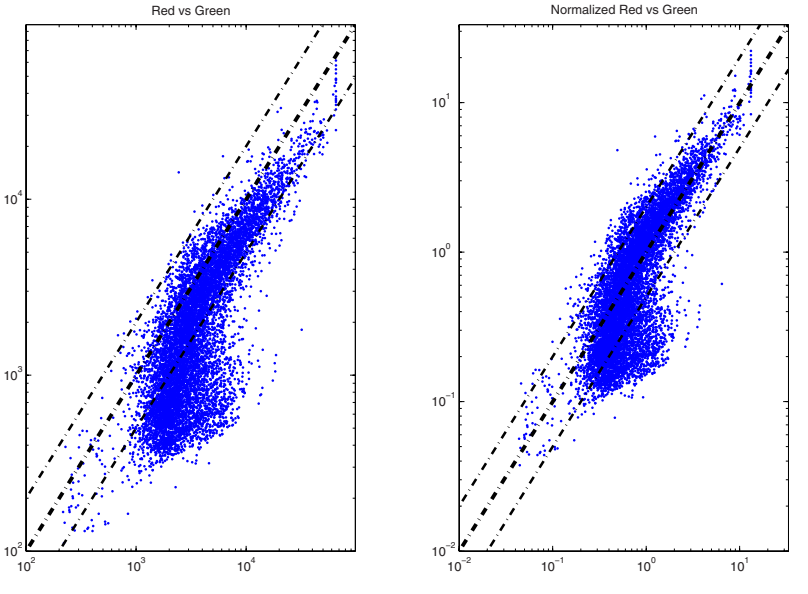

Fig. 2. Comparison of Normalisation of intensity data

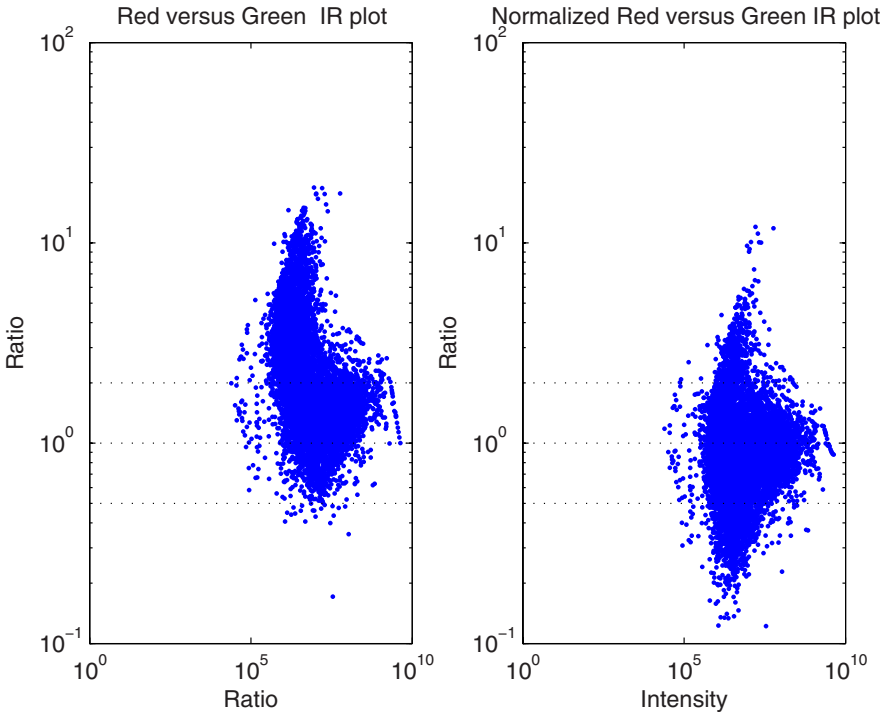

Fig. 3. Comparison of Normalisation of intensity/ratio data

mean they are related to important cellular activity that is expressed in small quantities of cDNA or perhaps could just be noise. Replicates, are one way of determining such effects.

It is also useful to plot the log_2 ratios against the intensity for each spot. Figure 3 shows how such a plot can highlight the difference.

Typically, the first and most commonly used technique is to normalise the data, this manipulates the hybridisation intensities to balance them in order to make meaningful comparisons. Normalisation usually needs to be applied because of various problems with experimental bias such as background intensities of the microarrays are not uniform. Normalisation of data means that weaker signals are amplified, this could mean they are related to important cellular activity that is expressed in small quantities of cDNA or perhaps could just be noise. Replicates, are one way of determining such effects.

2.1 Kohonen Self-Organising Feature Map (SOM)

The Kohonen SOM consists of a simple architecture. Since its initial introduction by Kohonen several improvements and variations have been made to the training algorithm. The SOFM consists of two layers of neurons, the input and output layers. The input layer presents the input data patterns to the output layer and

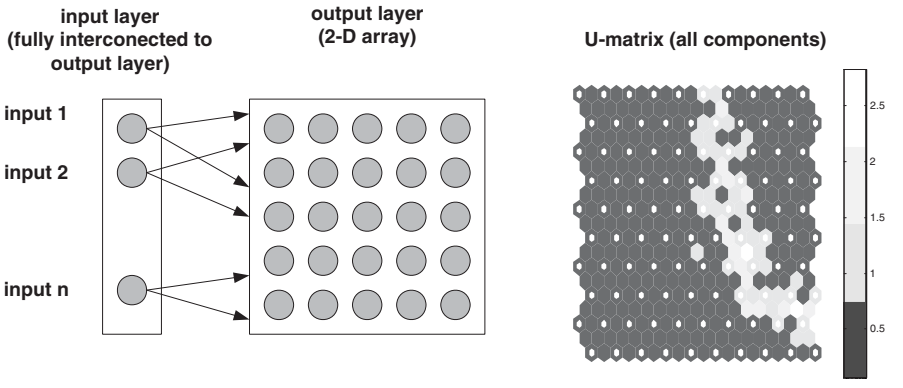

Fig. 4. Architecture of SOM, showing a regular grid of neurons. The U-matrix technique calculates the weighted sum of all Euclidean distances between the weight vectors for all output neurons. The resulting values can be used to interpret the clusters learned by the SOM. Each white dot represents a neuron and the colours represents different values of the weights, a distinct boundary is formed forming two large clusters.

is fully interconnected. The output layer is usually organised as a 2-dimensional array of units which have lateral connections to several neighbouring neurons. The architecture is shown in Figure 4.

Each output neuron by means of these lateral connections is effected by the activity of its neighbours. The activation of the output units according to Kohonens original work is by equation 1. The modification of the weights is given by equation 2 :

$$O_j = F_{min}(d_j) = F_{min}(\sum_i (X_i - W_{ji})^2) \tag{1}$$

$$\Delta W_i j = O_j \eta (X_i - W_{ji}) \tag{2}$$

where:
O_j= activation of output unit, X_i = activation value from input unit, W_{ji} = lateral weights connecting to output unit, d_j = neurons in neighbourhood, F_{min}= unity function returning 1 or 0, η= gain term decreasing over time.

The lateral connections enable the SOM to learn "competitively", this means that the output neurons compete for the classification of the input patterns. During training the input patterns are presented to the SOM and the output unit with the nearest weight vector will be classed as the winner.

The Kohonen self-organising feature map (SOM) is a neural network which is unsupervised technique that represents multi-dimensional patterns into 2-dimensional form for visualisation [13]. It also has the important feature of topological preservation i.e. clusters that are close to each other represent patterns that are very similar. The SOM is often used to group microarray gene expression data into related clusters, for example Kaski selected a subset of 1551 yeast genes of known functional classes [14,15]. Since neural networks are not

amenable to internal scrutiny (they are known as black boxes), Kaski was interested in determining the internal representation by using U-matrix analysis to show *how* the SOM partitioned the boundaries between the clusters.

2.2 The Gene Ontology

The use of ontologies is increasingly perceived as a way forward to overcome the complexity of biological information, for comprehensive introductions see [16]. A substantial amount of biological information is hierarchial in nature and the inter-relationships between the various pieces of knowledge can be meaningfully formalized, structured and represented by an ontology. One should not confuse Gene Ontology with a database of gene sequence or with a catalogue of gene product, rather than it gives us an idea of how gene product behaves at cellular level. It is not a way to bring together all the available biological datasets. The authors of GO have tried to provide a practically useful framework for keeping track of biological annotations which are applied to gene products.

GO is divided in to three disjoint term hierarchies, which are cellular component, biological process and molecular function. A cellular component is just a component of a cell with a condition that it is a part of large object, which might be a gene product or anatomical structure. A biological process is defined in GO as:"A phenomenon marked by changes that lead to a particular result, mediated by one or more gene product" [17]. Biological process terms can be quite specific

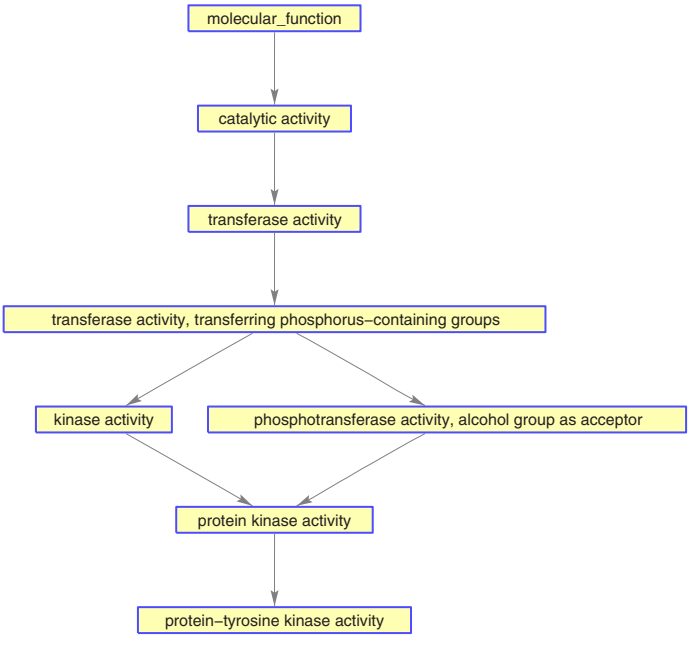

Fig. 5. Gene Ontology identifies gene JNK3 as active in protein-tyrosine kinase activity

(glycolysis) or very general (apotosis). Molecular function and biological process terms are clearly closely interrelated. Molecular Function describes activities at molecular level, like that of binding activities or catalytic activities, In GO it represent activities rather than molecules or complexes that perform the action, and do not specify the context in which action take place.

3 Experimental Results

The work of Alizadeh is often cited as a clustering success, whereby the authors were able to identify a new sub-class of cancer [4]. The novel variety was revealed through hierarchial clustering of tumors DLBCL (diffuse large B-cell lymphoma) data. The authors identified two distinct groups that were highly correlated with patient survival rates (40% of patients respond well to conventional treatment), these patients showed *germinal centre B-like DLBCL* stages of expression. This implied a major breakthrough for the treatment for this variety of cancer as the 60% of patients who succumbed to the disease showed *activated B-like cells* stages of expression. Sources of experimental data: All the data used in this study including survival data of lymphoma patients was obtained from the web supplement of the publication of Alizadeh available at http:/llmpp.nih.gov/lymphoma/data.shtml.

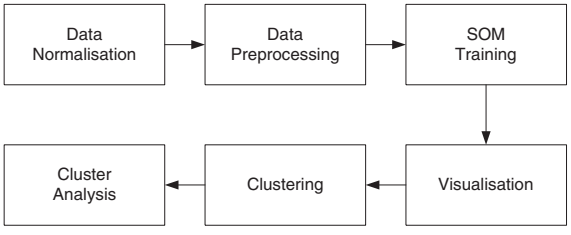

Fig. 6. Experimental setup and process

3.1 Data, Experimental Setup and Preprocessing

The various stages involved are highlighted in figure 6. The fluorescent intensity of each gene was tested and if greater than 1.4 times the local background were considered well measured. The ratio values were log-transformed (base 2) and stored in a table (rows, individual cDNA clones; columns, single mRNA samples). The Alizadeh data was preprocessed by the Lowess function with zero-norm with linear models and kernel methods. Each feature was given mean zero value and standard deviation was reduced to one. After cleaning the data that is removing all those which were under expressed and any bad measurement in the data, the original data set of 4026 genes was reduced to 3535 genes from 96 samples.

Figure 7 shows the U-Matrix of DLBCL entire data set, the individual clusters are quite well differentiated. The name of the genes superimposed over the

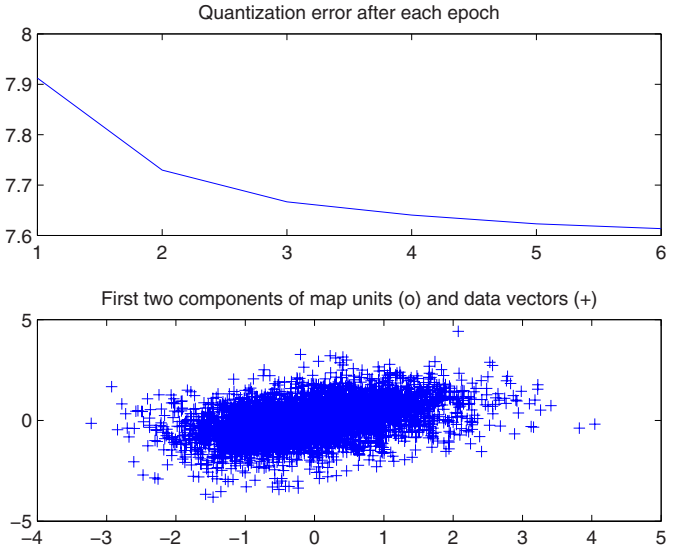

Fig. 7. Training run on DLBCL data

map unit so it is very easy to observe and analysis which genes are part of a particular cluster. The expression data can be judged by the colours, predominately reddish colour implies that a particular gene is highly expressed. The bluer the colour implies that a particular gene is less expressed. Despite their powers of visualization, SOMs cannot provide a full explanation of their structure and composition without further detailed analysis. The only method to have gone someway towards filling this gap is the *unified distance matrix* or U-matrix technique of Ultsch [18]. Further U-matrix research involving the analysis of individual component features was undertaken by Kaski [19]. Recent work by Malone makes explicit the contribution of each variable in the cluster to be assessed for characterising the cluster and can be expressed in rule format [20].

A deeper analysis of the SOM component plane (figure 8) reveals 42 DLBCL samples and three DLBCL lines (OCILy3, OCILy10 and OCILy1), the topology of the SOM is 20x15 and the colour scale of component plane represent the mean ratio in each map node.

Through the proposed approach applied above one can directly observe gene expression patterns of different lymphomas sub types i.e. DLBCL, CLL and FL, as it can be seen by the figures above that there are four prominent clusters identified in DLBCL 4, 2, 9 and the large group of clusters of 1 and 12 a short summary of the genes included in these cluster are listed table 1. After selecting the genes in the second subset file, the annotations have to be extracted from the ontology website. The particular information of interest for humans is gene-association-goa-human. It contains up to date annotations of Homo Sapiens, the more interesting genes were tested to get their ancestor list and also their root graph.

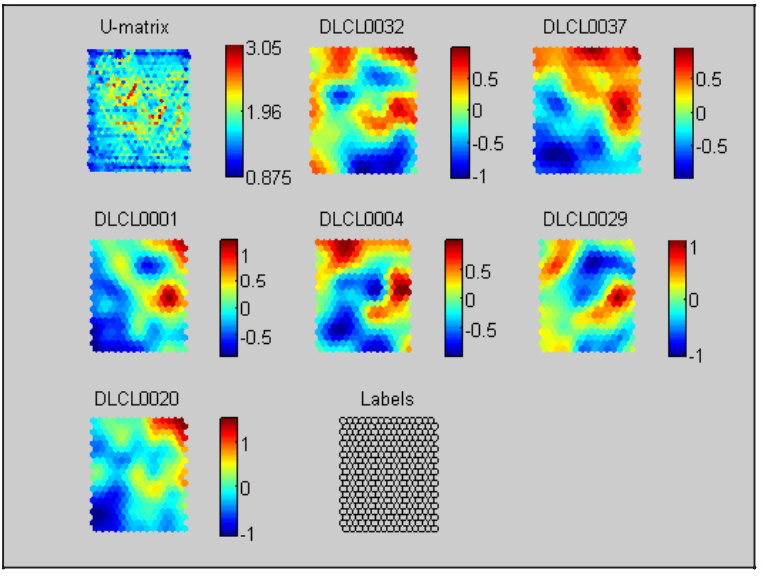

Fig. 8. Umatrix and SOM component planes

Table 1. Important DLBCL Genes clustered by the SOM

GeneName	GO ID	Description
TP73L	GO:0045892	Tumor protein p73-like
JNK3	GO:0004713	Catalysis of ATP and a protein tyrosine.
LYSp100	GO:0006952	Defense/immunity protein activity
RAD50	GO:0030674	physically linking the bound proteins or complexes to each other
CD44	GO:0016337	The attachment of one cell to another cell via adhesion molecules
GADD34	GO:0030968	Results in changes in the regulation of transcription and translation
CD5	GO:0025383	Involvement in DLBCL tumor progression

The DLBCL data was applied to the Gene Ontology to look at the significance of interesting genes and Gene Ontology terms that are used in the micro array. For the ontology study the data used all 3535 genes, first we applied K-means clustering was done to select only interesting genes during this all under expressed genes were removed, the total number of gene were reduced to 1157, than clustering was done into 4 sets. The difficulty of course is accurately identifying "interesting" genes.

4 Conclusions

We have demonstrated the use of Self Organising Map as a tool for analysis of gene expression data. The approach taken in our paper for the analysis of gene expression data were consistent with results originally published. However, the

aim of this study was to demonstrate the visualization capabilities of SOM with the original data. We also integrated the Gene Ontology with the discovered clusters of genes, which provides additional domain knowledge regarding gene function and common biological pathways. Finally in this study, the theoretical and practical approach of analysis of gene expression data of human Diffuse Large B cell Lymphoma have been discussed using SOM. We conclude that the SOM provides an excellent perfect platform for visualization and analysis of microarray data, and it will be very useful in extracting biologically meaningful information, when combined with domain knowledge such as the Gene Ontology.

Acknowledgements

This work was part supported by a Research Development Fellowship funded by HEFCE and the Biosystems Informatics Institute (Bii).

References

1. Berkum, N., Holstege, F.: Dna microarrays raising the profile. Current Opinions in Biotechnology 12(1), 48–52 (2001)
2. Soinov, L., Krestyaninova, M., Brazma, A.: Towards reconstruction of gene networks from expression data by supervised learning. Genome Biology 4(1), 1–10 (2003)
3. Sherlock, G.: Analysis of large-scale gene expression data. Current Opinion in Immunology 12, 201–205 (2000)
4. Alizadeh, A., Eisen, M., Davis, R., Ma, C.: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403, 503–511 (2000)
5. Kuo, P., Kim, E., Trimarchi, J., Jenssen, T., Vinterbo, S., Ohno-Machado, L.: A primer on gene expression and microarrays for machine learning researchers. Journal of Biomedical Bioinformatics 37, 293–303 (2004)
6. Huges, T., et al.: Functional discovery via a compendium of expression profiles. Cell 102, 109–126 (2000)
7. Lu, Y., Han, J.: Cancer classification using gene expression data. Information Systems 28, 242–268 (2003)
8. Peterson, C., Ringer, M.: Analyzing tumor gene expression profile. Artificial Intelligence in Medicine 28(1), 59–74 (2003)
9. Moreau, Y., Aerts, S., Moor, B.D., DeStrooper, B., Dabrowski, M.: Comparison and meta-analysis of microarray data: from the bench to the computer desk. Trends in Genetics 19(10), 570–577 (2004)
10. Kuo, P., Jenssen, T., Butte, A., Ohno-Machado, L., Kohane, I.: Analysis of matched mRNA measurements from two different microarray technologies. Bioinformatics 18(3), 405–412 (2003)
11. Quackenbush, J.: Computational analysis of microarray data. Nature Reviews Genetics 2, 418–427 (2001)
12. Quackenbush, J.: Microarray data normalisation and transformation. Nature Genetics Supplement 32, 496–501 (2002)
13. Kohonen, T., Oja, E., Simula, O., Visa, A., Kangas, J.: Engineering applications of the self-organizing map. Proceedings of the IEEE 84(10), 1358–1383 (1996)

14. Kaski, S., Nikkilä, J., Törönen, P., Castrén, E., Wong, G.: Analysis and visualization of gene expression data using self-organizing maps. In: Proceedings of NSIP-01, IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing 2001, Baltimore, USA (2001)
15. Nikkila, J., Kaski, S., Toronen, P., Castren, E., Wong, G.: Analysis and visualization of gene expression data using self-organizing maps. Neural Networks 8(9), 953–966 (2002)
16. Bard, J., Rhee, S.: Ontologies in biology: design applications and future challenges. Nature Reviews Genetics 5, 213–222 (2004)
17. Ashburner, M.: Gene ontology: tool for the unification of biology. Nature Genetics 25, 25–29 (2000)
18. Ultsch, A., Siemon, H.P.: Kohonens self organizing feature maps for exploratory data analysis. In: Proceedings of the International Neural Network Conference, pp. 305–308 (1990)
19. Kaski, S.: Dimensionality reduction by random mapping: Fast similarity computation for clustering. In: Proceedings of IJCNN'98, International Joint Conference on Neural Networks, Piscataway, NJ, vol. 1, pp. 413–418 (1998)
20. Malone, J., McGarry, K., Bowerman, C., Wermter, S.: Rule extraction from kohonen neural networks. Neural Computing Applications Journal 15(1), 9–17 (2006)