Visualisation of Distributions and Clusters Using ViSOMs on Gene Expression Data

Swapna Sarvesvaran and Hujun Yin

University of Manchester Institute of Science and Technology (UMIST)
Department of Electrical Engineering and Electronics
Manchester, M60 1QD, UK

Abstract. Microarray datasets are often too large to visualise due to the high dimensionality. The self-organising map has been found useful to analyse massive complex datasets. It can be used for clustering, visualisation, and dimensionality reduction. However for visualisation purposes the SOM uses colouring schemes as a means of marking cluster boundaries on the map. The distribution of the data and the cluster structures are not faithfully portrayed. In this paper we applied the recently proposed visualisation induced Self-Organising Map (ViSOM), which directly preserves the inter-point distances of the input data on the map as well as the topology. The ViSOM algorithm regularizes the neurons so that the distances between them are proportional in both the data space and the map space. The results are similar to the Sammon mappings but with improved details on gene distributions and the flexibility to nonlinearity. The method is more suitable for larger datasets.

1 Introduction

Microarray technologies make it straightforward to monitor simultaneously the expression patterns of thousands of genes during cellular differentiation and response [1, 2]. Tens of thousands of data points are generated from every experiment. DNA arrays provide a snapshot of all the genes expressed in a cell at a certain time. One of the ultimate goals of biological research is to determine the proteins involved in specific physiological pathways. Hence DNA arrays play a major role in understanding biological processes and systems ranging from gene regulation, to development and to disease from simple to complex. The information obtained can be studied and analysed, to identify the underlying genetic causes of many human diseases, drug discovery and clinical research. One way of discovering pathways and families of similarly acting proteins is to monitor the expression levels of messenger RNA (mRNA), which encodes for the corresponding proteins. The state of a particular cell and its functions is reflected in the levels of mRNA. So subjecting a cell to environmental stimuli and measuring the mRNA levels of genes of interest over time provides expression patterns for the genes.

The concept of microarrays is as follows: mRNAs are extracted from genes that are under study, converted into corresponding complimentary DNAs (cDNA), and tagged with a florescent dye. This is then washed over a glass slide (DNA chip) bearing a grid spotted with DNA sequences of known genes. Tagged cDNAs hybridise (bind)

Z.R. Yang et al. (Eds.): IDEAL 2004, LNCS 3177, pp. 78–84, 2004.

with corresponding DNA sequences on the microarrayer. Analysing the location and intensity of the florescent signals we can determine the levels of activity for each gene. The DNA chip allows scientists to study the entire genome of an organism. This presents a problem, in a statistical aspect, as the data produced from microarray experiments are enormous and trying to visualise datasets of high dimensionality proves very difficult. Since there is no single "best method" available to analysis and visualise microarray data, various methods have been proposed. Numerous dimensionality reduction methods exist that have been used on expression level datasets [7, 9, 12].

Section 2 briefly describes various projection methods. Section 3 describes the related work, ViSOM and potential applications. Section 4 gives a brief explanation about the datasets used and details about the proposed work and the results, together with discussions, are presented in Section 5. Finally Section 6 concludes.

2 Projection Methods

Dimensionality reduction methods map the original data typically into two dimensions, in order to display them onto a screen. The mapping, in order to be useful, needs to serve a human observer by preserving important structures of the original data. The best projection method is not self-evident, but depends on the distribution and nature of the original data and the usage of the resulting configuration. Two popular methods are the principal component analysis (PCA) and multidimensional scaling (MDS). SOMs have also been used as a dimensionality reducing technique, and in conjunction with other clustering methods such as the *k*-means and hierarchical clustering [14].

2.1 Principal Component Analysis

PCA allows data to be displayed in two dimensions with as much of the variation in the data as possible. It helps to filter noise and reduce the dimensionality of the data without a significant loss of information, making the data more accessible for visualisation and analysis. For a more in depth view on PCA, its application to microarray data and its extensions to nonlinear forms see [4, 6, 7, 10]. One of the disadvantages of PCA is its inability to capture nonlinear relationships in a dataset and if the input dimensionality is much higher than two, the projection onto a linear plane will provide limited visualisation power [16].

2.2 Multidimensional Scaling

MDS, well described in [3], searches for a low dimensional space, which is usually Euclidean, where each point in the mapping space represents one object/variable (genes in the microarray aspect) and such that the distances between the points in the map space, match as well as possible the distances of these points in the input space. That is, it tries to preserve the pairwise distances between data points, so that they are proportional in both the mapping space and the input data space. MDS is generally nonlinear and can reveal the overall structure of the data, but cannot provide the un-

derlying mapping function [16]. Sammon [8, 9] mapping is a popular MDS method; its algorithm is based upon the Newton optimisation techniques. Since Sammon mapping is a point-to-point mapping, like other MDS methods, every time a new point is introduced, the projection has to be recalculated from scratch based on all data points, making it computationally intensive especially when dealing with large datasets (like microarray data). Therefore it requires large amounts of computer memory. Torkkola, et al. [12] suggest, combining Sammon with the SOM algorithm to over come these problems. The Sammon mapping is applied to the results of the SOM algorithm, which has already achieved a substantial data reduction by replacing the original data with fewer representative prototypes.

2.3 Self-organising Maps

Kohonen's SOM is one of the most popular artificial neural networks [5]. The SOM is both a projection method, which maps high-dimensional data into low-dimensional space, and a clustering method so that similar data samples tend to be mapped to nearby neurons (topology preservation). The SOM has been used in data mining and visualization for complex datasets. The network consists of a number of neurons or nodes, usually arranged on a rectangular or hexagonal grid. The SOM is used to reduce the amount of data by clustering and constructing a nonlinear projection of the data onto a lower-dimensional display. For visualisation purposes the SOM uses a colouring scheme such as U-matrix [14], to visualise the relative distances between data points in the input space on the map. But this does not faithfully portray the distribution of the data and its structure.

3 Related Work

The ViSOM, proposed in [17, 18], is a nonlinear projection method for data visualisation but of simple computational structure compared to Sammon mapping that requires the first and second order derivative for every data point in every iteration. ViSOM projects high dimensional data in an unsupervised manner, similar to the SOM, but constrains the lateral contraction force between the neurons and hence regularises the inter-neuron distances with respect to a scalable parameter that define and controls the resolution of the map. The ViSOM preserves the inter-point distances as well as the topology as faithfully as possible therefore providing direct visualisation of the structure and distribution of the data. This paper used a smoothed version of the ViSOM. The algorithm is described as follows [17]:

1. Initialise the weights with principal components or to small random values.
2. At time step t, an input $\boldsymbol{x}(t)$ is drawn randomly from the dataset or data space. A winning neuron, say $\boldsymbol{v}$, can be found according to its distance to the input, $v = \arg\min_{c \in \Omega} \left\| x(t) - w_c \right\|$
3. Update the winner according to equation $\Delta w_v(t+1) = \alpha(t)\left[x(t) - w_k(t)\right]$

4. Update the weights of the neighbouring neurons using

$$\Delta w_k(t+1) = \alpha(t)\eta(v,k,t)\left([x(t)-w_v(t)]+[\xi+(1-\xi)(\frac{d_{vk}}{\Delta_{vk}\lambda}-1)][w_v(t)-w_k(t)]\right)$$

   Here d_{vk} and Δ_{vk} are the distances between nodes v and k in the data space on the map respectively, ξ is the smooth variable varying from 1 to 0 gradually with time during the training period, λ the resolution parameter depending on the size of the map and the variance or breadth of the data. The smaller the value of λ, the higher resolution the map can provide.
5. Refresh the map by randomly choosing a neuron and using its weight vector as the input for a small percentage of updating times (say for 20% of the iterations). Then the process is repeated until the map converges.

The constraint is introduced gradually for a smooth convergence. More details on this aspect and also the relation to principal curves/surfaces can be found in [17, 18]. This algorithm has already been applied to visualise high dimensional datasets [17, 18], but not microarray datasets.

4 Experiments

Several experiments have been conducted and their results are presented in section 5. The experiments are to demonstrate the usefulness of the ViSOM in visualising multivariate data and its advantages over other methods. The ViSOM has not been previously applied to gene expression datasets.

In the first example, the publicly available dataset of Saccharomyces cerevisiae bakers yeast is used[1]. A sample of size 232x17 (232 rows and 17 columns) was chosen as these 232 genes have been fully identified [1]. Four methods: PCA, Sammon mapping, SOM and ViSOM, were applied to this dataset. The second example uses the rat dataset[2], 112x9 was used same as [16].

5 Results and Discussion

The results shown in Fig. 1 are the results of various projection methods applied to the first dataset. In [1], 6220 (Saccharomyces cerevisiae, bakers yeast) trancripts were monitored. To obtain synchronous yeast culture, cdc28-13 cells were arrested in late G1 at START by raising the temperature to 37˚C, and the cell cycle was reinitiated by shifting cells to 25˚C. Cells were collected at 17 time points taken at 10 min intervals, covering nearly two full cell cycles. Out of which 416 showed cell cycle-dependent periodicity. 232 biologically characterized genes that showed transcriptional periodicity is listed in [1]. These are the genes used in this paper and referred to as the mitotic dataset.

In the plots seen in Fig. 1, the crosses (x) indicate all the 232 genes listed in [1]. The different shapes, i.e., triangles, circles, squares, diamonds and plus signs indicate genes that were previously identified to be cell cycle regulated by traditional

[1] Dataset available at http://genomics.stanford.edu

[2] Dataset available at http://rsb.info.nih.gov/mol-physiol/PNAS/GEMtable.html

methods[3]. Triangles indicate functionally characterized genes in the G2/M phase, diamonds the S/G2 phase, squares are from the M/G1 Boundary, circles represent Late G1, and plus signs are known genes in the S phase, the onces marked are the histones (proteins that are required for normal transcription at several loci) [11]. Not all the genes from the "known regulated genes" list have been marked, not all of them are in the list given in [1] showed transcriptional periodicity. Various shapes and colours are used to specifiy the five phases of the cell cycle.

In Fig. 2 the rat dataset is used [16]. The dataset consists of 112 genes over 9 conditions. This study was conducted so that relationships between members of important gene families during different phases of rat cervical spinal cord development, assayed over nine time points before (E=embryonic) and after birth (P=postnatal) could be discovered.

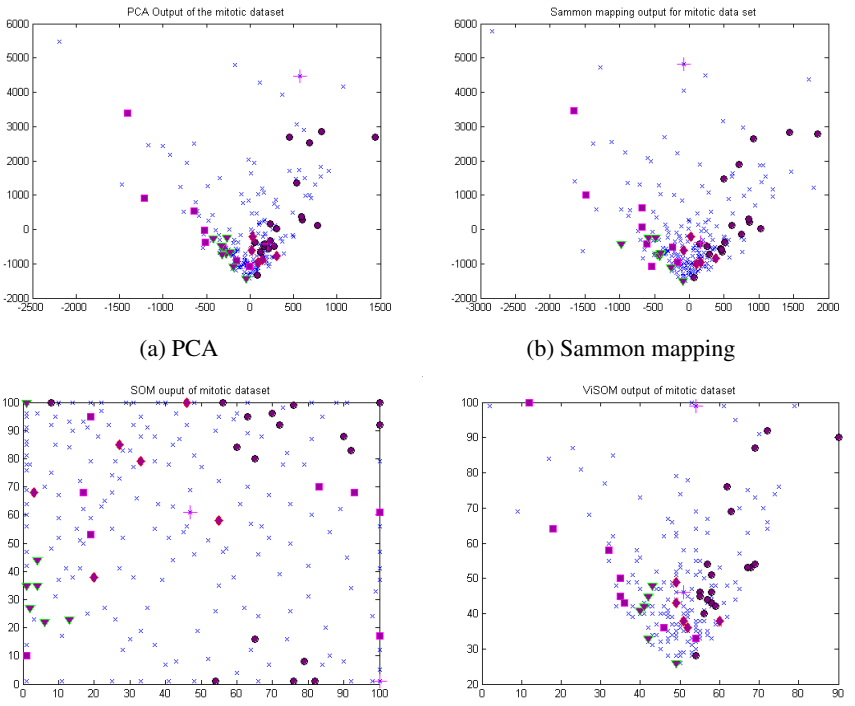
A rectangular ViSOM was applied to both the datasets and the projected data on the map is shown in Fig.1(d) and Fig.2(b). For comparison, a SOM of the same size and structure has been applied to map the mitotic data and the result is shown in Fig.1(a). The Sammon output for the rat dataset is shown in Fig.2 (a). The intial states of the Sammon mapping, SOM and ViSOM were all placed on a plane spanned by the first two principal components of the data. As can be seen, the ViSOM result closely resembles that of the Sammon mapping except that the data point are more seperated in the ViSOM (i.e. it has captured more details of intra-cluster and inter-point distribution) so each individial cross can be seen more clearley , instead of a lot of overlapping as seen in the Sammon output. The Sammon method is better than the linear PCA in revealling nonlinear structural details, and in the SOM it is impossible to see the inter-cluster and intra-cluster distribution. It can be asssumed that, points plotted near genes with known functions have similar functions to the genes surrounding it or are involved in similar biological pathways.

The advantage of applying the ViSOM to biological data is that the algorithm can be generalised so that no matter how big the data size , the ViSOM algorithm can be adapted accordingly. It is not computationally intensive like Sammon mapping, which requires storing all interpoint distances and second order optimisation processes. Both Sammon mapping and ViSOM can preserve the inter-cluster and intra-cluster details as well as the inter-point distribution of the data, this enable biologists to view each point or gene in the projected space more clearly compared to the other three methods mentioned.

6 Conclusion

In this paper, the ViSOM has been applied on gene expression datasets. The use of ViSOM intends to uncover the structure and patterns from the datasets, and to provide graphical representations that can support understanding and knowledge construction. The ViSOM has been compared to the SOM and Sammon mapping. It is similar in structure to that of the SOM and has similar capabilities as the Sammon mapping; preserving the inter-point distribution of the data. It allows for new points to be added to be projected on to the lower dimensional map without the need for re-calculation from scratch based on all data points. The ViSOM constrains the lateral contraction force within the updating neighborhood; without this the ViSOM is the same as the SOM.

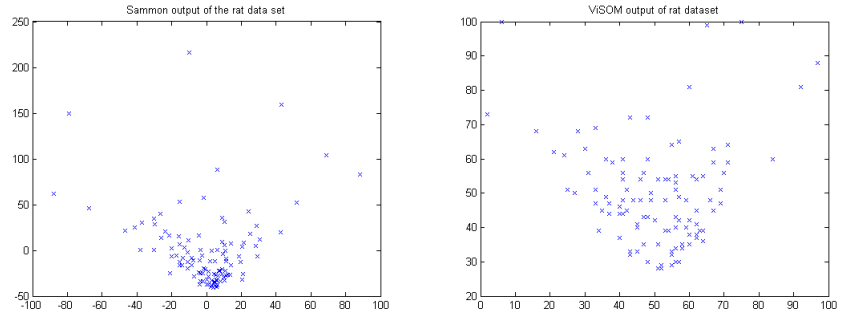
[3] List available at http://genome-www.stanford.edu/cellcycle/data/rawdata/



(a) PCA (b) Sammon mapping

(c) 100x100 SOM (d) 100x100 ViSOM

Fig. 1. Projections of mitotic dataset. Each projection shows genes whose functions have been identified within the mitotic cell cycle. The different shapes show characterized genes in different phases of the cell cycle: circles – Late G1, squares – M/G1 boundary, diamonds – S/G2, triangles – G2/M, and crosses – S. The application of the ViSOM algorithm to this dataset resulted in a better visualisation of the genes compared to the PCA, Sammon mapping and SOM. The inter-point distances as well as the neighbouring genes from the original data space are preserved in the lower 2-D space.



(a) Sammon mapping (b) A 100x100 ViSOM.

Fig. 2. Projections of rat dataset. It can be clearly seen that the projections of the rat dataset onto a ViSOM map are more discernable compared to the Sammon mapping of the same dataset.

The ultimate goal for researchers in the area of microarray data visualization, is to design tools for visual representations that will allow biologists to view appropriate underlying distributions, patterns, and therefore contribute to enhance their understanding of microarray analysis results. So they can then predict various genomic pathways and protein functions.

References

1. Cho, R. et al.: A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle, *Molecular Cell*, Vol.2, 65-73, July1998.
2. Chu, S. et al.: The transcriptional program of sporulation in budding yeast, *Science* 282, 699-705, 1998.
3. Cox, T. F., Cox, M. A. A.: *Multidimensional scaling*, London: Chapman and Hall, 1994.
4. Karhunen, J., Joutsensalo, J.: Generalisation of principal component analysis, optimisation problems, and neural networks, *Neural Networks* 8, 549-562, 1995.
5. Kohonen, T.: *Self-Organizing Maps*, 2nd edition, Springer. 1995.
6. Kramer, M.A.: Nonlinear principal component analysis using autoassociative neural networks, *AICHE Journal*, 37, 233-243, 1991.
7. Raychaudhuri, S.et al.: Principal Components Analysis to Summarize Microarray Experiments- Application to Sporulation Time Series, *Pac Symp Biocomput*: 455-66, 2000.
8. Ripley B. D.: *Pattern recognition and neural network*, Cambridge, UK: Cambridge University Press, 1996.
9. Sammon, J.W.: A nonlinear mapping for data structure, *IEEE Transactions on Computer*, 18, 401-409, 1969.
10. Schölkopf,B., Smola, A., Müller, K.R.: Nonlinear component analysis as a kernal eigenvalue problem, *Neural Computation*, 10, 1299-1319, 1998.
11. Spellman P.T. et al.: Comprehensive Identification of Cell Cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridisation, *Molecular Biology of the Cell* 9, 3273-3297, 1998.
12. Torkkola, K. et al.: Self-organizing maps in mining gene expression data, Information Sciences 139: 79-96, 2001.
13. Törönen, P. et al.: Analysis of gene expression data using self-organising maps, *FEBS Letters*, 451, 142-146, 1999.
14. Ultsch, A.: *Self-organizing neural networks for visualization and classification*. In O.Opitz, B. Lausen, & R. Klar (Eds.), Information and classification (pp. 864-867), 1993.
15. Wang et al.: Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study, *Bioinformatics* 3:36, 24 Nov 2002.
16. Wen, X. et al.: Large-Scale Temporal Gene Expression Mapping of CNS Development, *Proc Natl Acad Sci USA*, 95:334-339, 1998.
17. Yin, H.: *Visualisation induced SOM (ViSOM)*, In N. Allinson, H. Yin. L. Allinson, &J. Slack (Eds.), Advances is self-organising maps, (pp. 81-88), Proceedings WSOM'01, London: Springer, 2001.
18. Yin, H.: Data visualisation and manifold mapping using the ViSOM, *Neural Networks* 15: 1005-1016, 2002.