Proc. NSIP'01, IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing, to appear

ANALYSIS AND VISUALIZATION OF GENE EXPRESSION DATA USING SELF-ORGANIZING MAPS

Samuel Kaski, Janne Nikkilä

Helsinki University of Technology
Neural Networks Research Centre
P.O.Box 5400, 02015 HUT, Finland
{samuel.kaski,janne.nikkila}@hut.fi

Petri Törönen, Eero Castrén, Garry Wong

University of Kuopio, A.I.Virtanen-institute
P.O.Box 1627, 70211 Kuopio, Finland
toronen@hytti.uku.fi
{eero.castren,garry.wong}@uku.fi

ABSTRACT

Cluster structure of gene expression data obtained from DNA microarrays is analyzed and visualized with the Self-Organizing Map (SOM) algorithm. The SOM forms a non-linear mapping of the data to a two-dimensional map grid that can be used as an exploratory data analysis tool for generating hypotheses on the relationships, and ultimately of the function of the genes. Similarity relationships within the data and cluster structures can be visualized and interpreted. The methods are demonstrated by computing a SOM of a set of yeast genes of known functional classes. The cluster structure is visualized and the clusters are characterized in terms of the properties of the expression profiles of the genes. Additionally, relationships of the functional classes and division of the classes into subclasses are visualized.

1. INTRODUCTION

DNA microarray technologies have provided the opportunity to obtain gene expression data on a massive scale. DNA microarrays are ordered samples of DNA placed in high density on a solid support such that each sample represents a particular gene. This array can then be assayed for changes in the expression patterns of the representative genes after different treatments, different conditions, or tissue sources. Examples of different conditions are 1) cells before and after drug treatments 2) tissues from young vs. old age 3) healthy vs. cancerous tissues 4) yeast used in fermentation for beer vs. yeast used in fermentation for wine. Thus many important biological, physiological, medical, and industrial phenomena can be studied using arrays. Since the number of different genes on an array can number up to hundreds of thousands and the number of possible treatment conditions is virtually limitless, the amount of data being generated is substantial. Existing paradigms for the analysis of gene expression data have focused on three important aspects: preprocessing of and feature extraction from the data, clustering the data, and visualization. The goal of preprocessing and feature extraction is to transform the data in a suitable form for subsequent analysis. The goal of clustering is principally to group the data into meaningful sets. This step provides both data reduction and an analysis of the original data. The goal of visualization is to present the data in a way that both similarities and differences of the expression profiles and clusters of them can be seen.

This work was supported by the Academy of Finland, in part by the grant 50061.

Our laboratories have been using Self-Organizing Maps (SOM) as a rational method for clustering gene expression data sets. SOM is an unsupervised neural network algorithm that after preprocessing of the data, can cluster the data into biologically meaningful groups. While many other algorithms have been used for clustering and classifying gene expression data (e.g. hierarchical clustering, k-means clustering), less attention has been paid toward visualizing the data.

The SOM visualizes the high-dimensional data set, in this paper the gene expression patterns, on a graphical map display. Close-by locations on the map represent data that are similar in the original high-dimensional space, in our case genes having similar profiles of expression in the experimental treatments. The density structures of the data, i.e. clusters and substructures within them, can be visualized on the same display using for example the so-called U-matrix method. In this paper we explore the use of a U-matrix and newer methods of detecting cluster borders as visualization tools in gene expression analysis. In the cluster visualizations there are dark stripes indicating separation of clusters with abrupt changes. These isolated clusters may ultimately contain gene expression data with the underlying genes having distinct functions from genes found in neighboring clusters. We will apply recently developed interpretation methods for characterizing such clusters. Since the complexity and the amount of gene expression data continues to increase, visualization methods may be increasingly valuable to analyze and provide functionally meaningful results.

2. EXPLORATORY CLUSTER ANALYSIS WITH THE SOM

2.1. The SOM

The Self-Organizing Map (SOM [5, 6]) is a discrete grid of map units (the hexagons in Fig. 2). Each map unit represents certain kinds of data, in the present paper genes expressed in similar ways in the chosen set of treatments. The input data is represented in an ordered fashion on the map: Map units close-by on the grid represent more similar expression profiles and units farther away progressively more different profiles. The map is a similarity diagram that presents an overview of the mutual similarity of the large amount of high-dimensional expression profiles.

The mapping is defined by associating a n-dimensional *model vector* $\mathbf{m}_i$ with each map unit i, and by mapping each expression profile $\mathbf{x}$ to the map unit having the closest model vector. In this paper the length of all vectors is normalized to unity and we use the inner product metric, for which the closest model vector c must fulfill the condition

$$\mathbf{x}^T \mathbf{m}_c \geq \mathbf{x}^T \mathbf{m}_i \tag{1}$$

for all i. The mapping becomes ordered and represents the data when the values for the model vectors are computed in an iterative process. In "on-line" type of computation at step t one expression profile $\mathbf{x}(t)$ is selected at random, the closest model vector c is found by (1), and the model vectors are adapted according to

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{ci}(t)\mathbf{x}(t) \ , \tag{2}$$

and normalized to unit length. Here $h_{ci}(t)$ is the *neighborhood function*, a decreasing function of the distance of the units c and i on the map grid. We have used a Gaussian function that decreases in width and height during the iterative computation. For more details, variants, and different methods of computing SOMs see [6, 7].

2.2. Data Analysis with SOM

The SOM bears similarity to two kinds of traditional methods of data analysis: Dimensionality reduction methods and clustering methods. Projection and multidimensional scaling methods can be used to reduce the dimensionality of the data that can then be visualized in the low-dimensional space. According to recent evidence [8] the similarity diagrams formed by the SOM are more trustworthy in the sense that if two data points are close-by on the display they are more likely to be close-by in the input space as well. (Note that it is impossible to construct perfect mappings that reduce dimensionality; different methods make different kinds of compromises.)

Clustering methods reduce the amount of data by grouping similar data samples together. Most methods make (implicitly) some assumptions about the desired cluster shape and then extract a set of clusters that best fit the data. The SOM has often been used in the same fashion, by treating each map unit as a separate cluster. However, several neighboring map units may actually represent one cluster. Suitable visualizations can aid in gaining an overview of the cluster structure and extracting clusters having arbitrary shapes.

The hierarchical clustering methods commonly used in biology, and already applied to gene expression data (e.g. [4]), visualize a tree of subclusters of increasing detail. The problem with the trees is that they are sensitive to small variations in the data, and for large data sets they become huge. It is then virtually impossible to extract the essential cluster structures from them. We argue that the SOM, together with suitable visualization methods, can provide an overview of the data collection, and interactive visualization and analysis methods can be used for extracting and interpreting clusters in the data.

3. RESULTS

We clustered the expression profiles of genes of the budding yeast *Saccharomyces cerevisiae.* The expression data was collected by microarrays during diauxic shift [1], the mitotic cell division cycle [2], sporulation [3], and shock treatments [4], forming 79-dimensional data vector for each gene. The original data is available at `http://genome-www.stanford.edu/clustering/`.

Since our primary goal in this paper is to demonstrate that the SOM-based exploratory tools are useful for analysis of gene expression data, we chose a subset of 1551 genes of known functional classes. It is known (cf. [9]) that these classes are expressed in mostly different ways in this collection of treatments. Hence we can use the data to verify our methods. We may additionally gain some new insight into the relations of the classes and subdivisions of the classes.

Genes had been classified to functional categories according to the MIPS yeast functional classification catalogue[1]. Based on this classification and [9] the following set of functional classes were chosen: "Amino-acid biosynthesis," "Glycolysis and Gluconeogenesis," "Pentose-phosphate pathway," "Tricarboxylic-acid pathway," "Respiration," "Fermentation," "Ribosomal proteins," "Proteolysis," "Cytoplasmic degradation," "Sugar and carbohydrate transporters," "Pheromone response generation," "Organization of the cytoplasm," "Nuclear organization," "Organization of chromosome structure," and "Mitochondrial organization".

3.1. Visualizing Clusters

The density or cluster structure of the data can be visualized by the so-called U-matrix [10] computed of a SOM. The U-matrix (Fig. **1a**) visualizes the distances between model vectors of neighboring map units with gray levels. It is known that the density of the model vectors reflects the density of the data items, and therefore the distances between model vectors are smaller in clustered areas and larger in sparser areas.

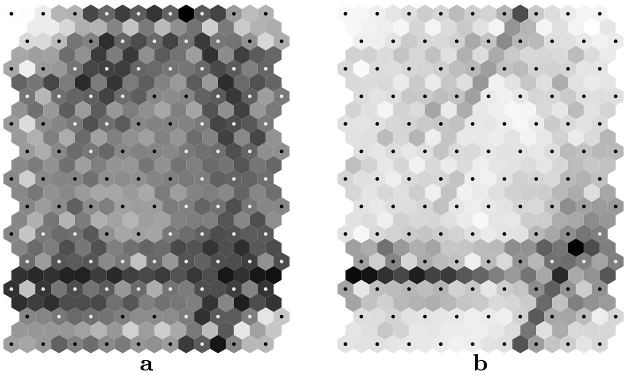
The problem with the U-matrix is that the visualization may be very noisy and it may be difficult to distinguish clustered areas. We have developed a heuristic method [11] that complements the U-matrix by detecting and emphasizing salient cluster borders. In the method, the difference between each model vector and the centroid of the closest data points is first computed (intuitively speaking, it points towards the closest cluster or increasing density), and then the differences at neighboring model vectors are compared. The result of the comparison is plotted in gray levels (Fig. **1b**). If there is a dark stripe then there either is a sharp minimum of data density in between the neighboring model vectors, or the model vectors are in between two clusters.

3.2. Interpreting the Clusters

3.2.1. Relation to the Functional Classes

We begin the analysis by plotting the distribution of the functional classes on the SOM map computed using expres-

[1]http://www.mips.biochem.mpg.de/



a b

Figure 1: Visualizations of the cluster structure in gene expression data. **a** U-matrix. **b** A new method for detecting cluster borders. Light shades: clusters; dark shades: sparser areas or gaps in between clusters. The dots denote map units. Note that in these displays there is a hexagon in between each pair of map units, whereas on the subsequent figures only the map units themselves are been shown. The same SOM is shown in all figures.

sion data only; the class information was not utilized in any way.

Three representative classes are visualized in Figure 2. "Cytoplasmic degradation" (Fig 2**a**) is an example of classes where the majority of genes are focused to a single map unit. The great majority of genes clustered to the map unit at the lower right corner contain proteasome subunits involved in protein degradation. Consequently, another functional class, "proteolysis", which also contains proteasome subunits, displayed a distribution similar to the class "cytoplasmic degradation" (not shown).

"Sugar and carbohydrate transporters" (Fig. 2**b**) represent classes where genes are divided to several loci on the map. This is not unexpected, since the expression patterns in this class vary according to which sugar pathway they belong to, which sugar is available, and the rate limiting enzyme. Genes involved in amino-acid biosynthesis (not shown) behave in the same way. The expression depends on which amino acid is being synthesized, what is the current availability or need for that amino acid, and what is the rate-limiting enzyme for that pathway.

In the third example, genes classified to "ribosomal proteins" (Fig. 2**c**) are divided into three clusters at the opposite corners of the map. Examination of the gene content in each of these clusters revealed that the major cluster in upper left corner contains almost exclusively different ribosomal protein subunits. In contrast, the majority of genes located in the upper right corner are mitochondrial ribosomal proteins. Another group of mitochondrial proteins is clustered to the lower left corner. This analysis suggests that ribosomal proteins and mitochondiral ribosomal proteins should be classified into separate functional classes (and the latter perhaps into two distinct classes) because they show clearly different gene expression patterns.

3.2.2. Characterization of the Clusters

The clusters found by the U-matrix and our enhanced method (Fig. 1) have traditionally been analyzed with three methods: (1) By plotting class distributions as we have done above, (2) By plotting the model vectors, in this case the expression profiles represented by the map units [12, 13], and (3) by plotting the distribution of the original data variables, in the present paper the expression levels at a certain treatment and a time point, on the map. The problem with the approaches (2) and (3), for large maps and high-dimensional data, is that they present a huge amount of data. We aim at summarizing the essential data by concentrating on characterizing the most salient clusters.

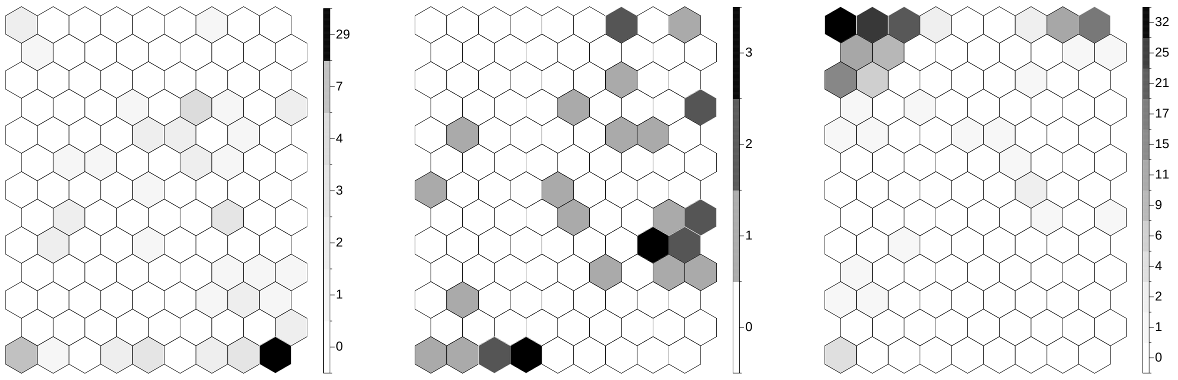
The problem we are addressing is what distinguishes one SOM area from its surrounding areas, i.e., what changes as one moves over a black stripe on the cluster displays of Figure 1. In terms of the original variables this can be seen as a question of what variables contribute most to the change. For example in linear factor analysis such contributions are standard interpretation tools.

In this paper we interpret the contributions in a straightforward and intuitive manner: We compute the difference between the representations, the (means) of the model vectors, on each side of the cluster border.

Three cluster areas are interpreted in Figure 3. The cluster 1 was selected according to the U-matrix in Figure 1**a**. Figure 3 (top left) shows that, compared to the most similar other genes around the cluster, the genes coding ribosomal proteins are more down-regulated in the heat and cold shock treatments and at some time points of the sporulation curve. This kind of behavior is to be expected, and the SOM shows that it is indeed characteristic of the genes encoding ribosomal proteins. The other findings, that the genes are relatively more up-regulated in the ELU and CDC treatments, and at some time points of sporulation, may be interesting as well.

The most obvious cluster, cluster 3 in Figure 3, contains genes from the groups of cytoplasmic degradation (Fig. 2**a**) and proteolysis. Compared to the genes in the area towards the top of the map they are clearly down-regulated in the heat and cold treatments and in the end of the diauxic shift. On the other hand, compared to the genes on the map area to the left, the genes in the cluster are up-regulated in the end of the diauxic shift. This group is also strongly upregulated when compared to the others in DTT shock. This is natural as the DTT shock affects the protein folding and forces the yeast to recycle proteins more and degradate the misfolded proteins.

Cluster 2 is an example of unexpected findings in the data, revealed by the exploratory analysis tools. Some such findings may form hypotheses for further inspection. Cluster 2 could not have been anticipated based on the known functional classes; it contains genes from several classes including amino-acid biosynthesis, sugar and carbohydrate transporters, and nuclear organization. The main feature (Fig. 3 bottom left) that distinguishes it from the upper area on the map is that its genes are clearly more up-regulated in sporulation. When the actual model vectors (not shown) are inspected it seems that the region is characterized by up-regulation in the sporulation treatment, and an almost flat profile elsewhere. This indicates that all the genes in



a: cytoplasmic degradation (87)

b: sugar and carbohydrate transporters (32)

c: ribosomal proteins (173)

Figure 2: Distribution of genes in three functional classes on the map. The number in parentheses denotes the total number of genes in the class. The scale shows the number of genes mapped to the map units.

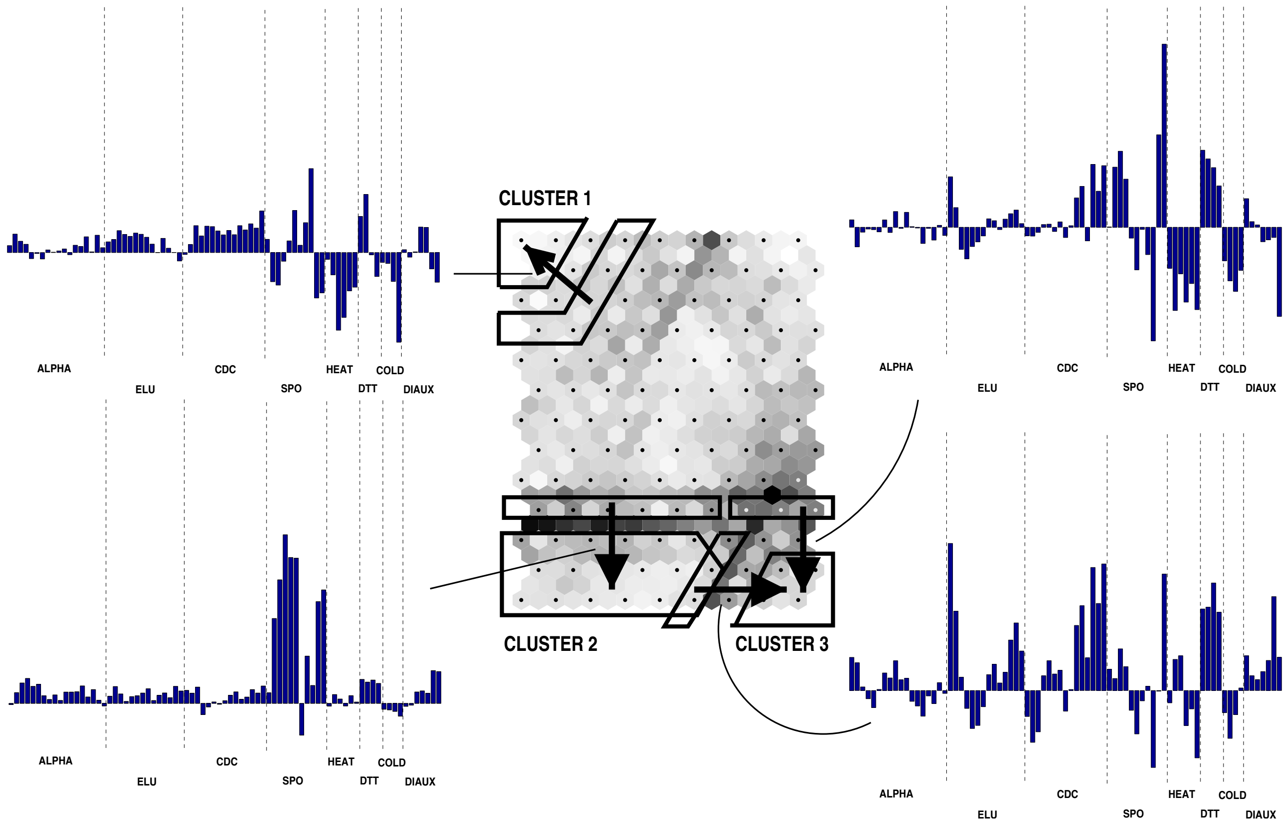


Figure 3: Characterization of three clusters in terms of the original variables. The profiles show how much the expression of the genes changes (on the average) when moving on the SOM according to the arrows. The changes can be used to characterize what is special in the expression of the genes. Abbreviations: ALPHA=cell-cycle: alpha factor arrest and release, ELU=cell-cycle: elutriation, CDC=cell-cycle: cdc15 arrest and release, SPO=sporulation, HEAT=heat shock, DTT=dtt shock 1mm, COLD=cold shock (see `http://genome-www.stanford.edu/clustering/`.

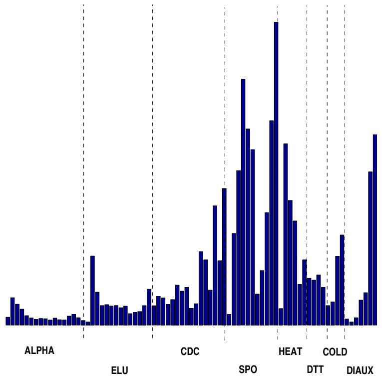


Figure 4: The (mean) contribution of the time points within treatments to the SOM-based analysis.

this cluster are connected to spore formation, regardless of their functional classification.

There are additional interesting areas on the map, distinguishable either by the U-matrix display in Figure 1 or by the distribution of the functional classes on the map. We will investigate such areas in more detail in subsequent papers, with maps computed from more comprehensive sets of genes.

3.2.3. Contributions of the variables

To provide an overview of how much each variable contributes to the SOM we computed the average of the local contributions over the map. In linear factor analysis it is straightforward to judge the contribution or communality of each input variable to the analysis. Since the SOM is non-linear, however, the contributions are different in different locations on the map.

The non-linear two-dimensional map can be approximated locally, at a certain map unit, by two *local linear factors* [11], and the contribution on the local factors can be computed. We approximate the total contributions of the variables by the average local contributions over the whole map.

It is evident (Fig. 4) that the contributions of different treatments and of different time points vary greatly. The SOM takes predominantly into account the CDC, sporulation, heat shock, and diauxic shift. This is mainly the result of the different variance in the treatments. If it is desired to emphasize the other treatments more, the contributions can be changed by scaling the treatments differently in the preprocessing phase.

4. CONCLUSIONS

We demonstrated the use of the Self-Organizing Map as an exploratory data analysis tool in the analysis of the gene expression data collected with DNA microarrays. It was shown that for a set of genes with known functional classes the SOM-based visualizations were compatible with prior knowledge of the properties of the classes.

The SOM-based visualizations offer an intuitive way to form hypotheses about the data. Some hypotheses were generated even for the relatively well-known set of yeast genes. For example the class “Ribosomal proteins” seems to consist of two or more subgroups suggesting that the functional classification concerning this subclass could be revised to correspond the behavior of the genes. Another example is a cluster (number 2 in Fig. 3) connected to spore formation. We will apply the methods in subsequent papers to less well known data, and analyze the hypotheses in more detail.

5. REFERENCES

[1] J.L. DeRisi, V.R. Iyer, and P.O. Brown, ”Exploring the metabolic and genetic control of gene expression on a genomic scale,” *Science*, vol. 278, pp. 680-6, 1997.

[2] P.T. Spellman et al., ”Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization,” *Mol Biol Cell*, vol. 9, pp. 3273-97, 1998.

[3] S. Chu et al., ”The transcriptional program of sporulation in budding yeast [published erratum: Science vol. 282 p. 1421, 1998],” *Science*, vol. 282, pp. 699-705, 1998.

[4] M. B. Eisen et al., “Cluster analysis and display of genome-wide expression patterns,” *PNAS*, vol. 95, pp. 14863–8, 1998.

[5] T. Kohonen, “Self-organized formation of topologically correct feature maps,” *Biol. Cybern.*, vol. 43, pp. 59–69, 1982.

[6] T. Kohonen, *Self-Organizing Maps*, Springer, Berlin, 1995, (Third, extended edition 2001).

[7] S. Kaski, J. Kangas, and T. Kohonen, “Bibliography of self-organizing map (SOM) papers: 1981–1997,” *Neural Computing Surveys*, vol. 1, no. 3&4, pp. 1–176, 1998.

[8] J. Venna and S. Kaski, “Neighborhood preservation in nonlinear projection methods: An experimental study,” in *Proc. Int. Conf. on Artificial Neural Networks, 2001*, submitted.

[9] P. Pavlidis et al., “Gene functional classification from heterogeneous data,” in *Proc. Fifth Int. Conf. on Computational Molecular Biology*, 2001, to appear.

[10] A. Ultsch, “Self-organizing neural networks for visualization and classification,” in *Information and Classification*, O. Opitz, B. Lausen, and R. Klar, Eds., pp. 307–13. Springer-Verlag, Berlin, 1993.

[11] S. Kaski, J. Nikkilä, and T. Kohonen, “Methods for exploratory cluster analysis,” in *Proc. of the Int. Conf. on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet*. Scuola Superiore G. Reiss Romoli, 2000.

[12] P. Törönen, M. Kolehmainen, G. Wong, and E. Castrén, “Analysis of gene expression data using self-organizing maps,” *FEBS Letters*, vol. 451, pp. 142–6, 1999.

[13] P. Tamayo et al.,“Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation,” *PNAS*, vol. 96, pp. 2907–12, 1999.