In Advances in Self-Organizing Maps, pp. 124-131, Springer, 2001

SOM-Based Exploratory Analysis of Gene Expression Data

Samuel Kaski

Helsinki University of Technology
Neural Networks Research Centre
P.O. Box 5400, FIN-02015 HUT, Finland
samuel.kaski@hut.fi

Abstract. Applications of new SOM-based exploratory data analysis methods to bioinformatics are described. Cluster structures are revealed in data describing the expression of a set of yeast genes in several experimental treatments. The structures are visualized in an intuitive manner with colors: The similarity of hue corresponds to the similarity of the multivariate data. The clusters can be interpreted by visualizing changes of the data variables (expression in different treatments) at the cluster borders. The relationship between the organization of the SOM and the functional classes of the proteins encoded by the genes may additionally reveal interesting relationships between the functional classes, and substructures within them.

1 Introduction

Recent methodological breakthroughs are transforming the cellular-level research of biological organisms. Several large-scale projects to decipher the genetic sequences have been carried out, and new methods for genome-wide measurement of gene expression have been introduced. The three-dimensional structures of the proteins are being modeled. All these approaches generate massive amounts of data, into which important knowledge about the function of the genes and proteins is hidden. It has been said that biology is changing from a hypotheses-driven science to a data-driven science. The currently very active field of bioinformatics provides the necessary new data-analysis methods. It is the glue between biology, computer science, and various modeling approaches.

Currently the challenges in genetics are shifting from analyzing the genetic sequences to analyzing the function of the genes. The Self-Organizing Map (SOM) has potential as a tool in the required large-scale analyses. SOMs have already been applied to clustering of yeast gene expression data [8,9].

In this paper SOM-based methods for exploratory data analysis will be described and applied to yeast gene expression analysis. We have developed methods for uncovering and visualizing cluster structures in an easily understandable manner, and for interpreting them in terms of the original data variables.

2 Gene Expression Data

The expression of even hundreds of thousands of genes can be measured simultaneously with DNA microarrays or "DNA chips." They are essentially arrays of DNA from each gene, placed as a matrix on a "chip." The array can be used to measure the amount of messenger RNA of each gene, an indicator of the activity of the gene, at a certain time. The effects of a drug treatment, for example, can be monitored by comparing the expression with a control condition. When this monitoring is done in a genome-wide scale and in many treatments, a huge amount of data about gene function is generated. One aim of our bioinformatics project is to analyze this data with SOMs.

The data set used in the experiments consists of measurements of the expression level of 1551 genes of the budding yeast *Saccharomyces cerevisiae*. The measurements containing 79 time points in total were made in 8 treatments, resulting in an "expression profile" for each gene. The data is available at `http://genome-www.stanford.edu/clustering`. For more details on the data and its analyses see [3].

3 SOM-based Data Exploration

The Self-Organizing Map [6,7] has been a very useful tool in several kinds of applications. Over 4000 scientific articles and several books have already been written on the SOM (`http://www.cis.hut.fi/research/som-bibl/`; cf. [1]).

One of the main applications of the SOM is data analysis. Similarity relationships within a data set can be visualized on a graphical SOM display. Also other aspects of the data, clusteredness or density structures, and distribution of the values of the data variables, can be visualized on the same display. Combined use of these visualizations helps in gaining an overview of the data set, and in interactive exploration of its properties. The SOM-based methods are perhaps at their best in the exploratory or data-driven phase of data analysis and research, in which hypotheses concerning properties or contents of the data set are formed. The exploratory phase is often followed by a hypothesis-driven phase in which the formed hypotheses are then confirmed.

There are many subtle differences between the SOM and alternative methods. For example, experimental results indicate that visualizations formed by the SOM are more trustworthy in the sense that if two data samples are close-by on the SOM display then they are close-by in the original space as well, at least more often than for alternative methods. This result was obtained empirically by comparing the results of the SOM and traditional multidimensional scaling-based non-linear projection methods [11]. Such trustworthiness is of course important in data analysis.

3.1 Visualization of Cluster Structures

Each data sample, here a gene expression profile, is mapped onto a certain point on the SOM grid. As a result of the SOM algorithm the data becomes organized on the grid so that close-by points contain genes having similar expression profiles.

The locations of genes coding proteins of three sample functional classes have been shown in Figure 1. Based on the displays it is evident that one of the classes (**a**) contains mostly genes that are expressed similarly in the set of treatments. The other classes are more diverse, containing subgroups that can be identified based on the SOM visualizations. We are currently investigating the biological significance of these results.

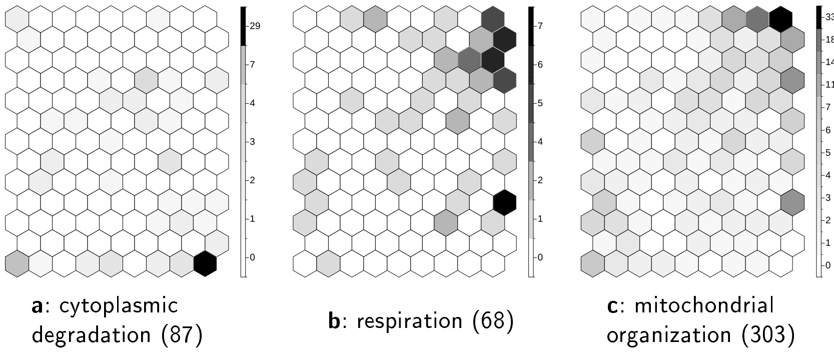

Fig. 1. Distribution of genes of three functional classes on the map. The number in parentheses denotes the total number of genes in the class. The scale shows the number of genes mapped onto the map units.

U-matrix. The SOM is defined by an ordered set of model vectors, one vector attached to each map unit or grid point. The model vectors represent the data in an ordered fashion, and the point density of the model vectors in the data space reflects the probability density of the data. Hence density structures of the data, i.e., cluster structures, can be visualized by visualizing the density of the model vectors. The so-called U-matrix display [10] visualizes the relative distances between neighboring model vectors by shades of gray.

The U-matrix of the gene expression data is shown in Figure **2a**. Some clustered areas are visible, but the noisiness of the visualization is problematic. In fact, for this map all pairwise distances between neighboring model vectors are relatively similar as can be seen in Figure **3a**.

A new method for detecting cluster borders. We have developed [2] a method that uses both the model vectors and the local distribution of the

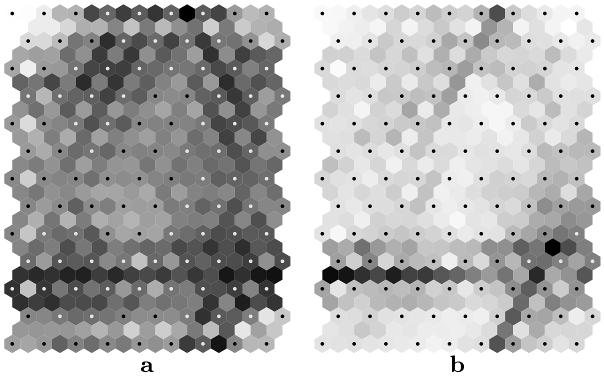

Fig. 2. Visualizations of the cluster structure in the gene expression data. **a** U-matrix. **b** Our new method for detecting cluster borders. Light shades: clusters; dark shades: sparser areas or gaps in between clusters. The dots denote the map units.

data to detect and visualize salient cluster borders not necessarily visible in the U-matrix type displays. It first computes, for each location of the SOM grid, a (heuristic) approximation of the local gradient of the data density. It is given by the difference between the model vector and the centroid of a set of closest data points. Then the gradient approximations at neighboring map units are compared by computing the norm of their difference, and the result of the comparison is visualized as a gray level on the map. If the difference in the gradients is large (dark shade in the visualization), the two map units are close to different clusters and there is a cluster border between them.

The cluster structures of the gene data have been visualized in Figure 2**b**. Two cluster borders are clearly visible as dark stripes.

Illustrative Visualization with Colors. The cluster structures are visible in the displays of Figure 2. However, the interpretation of the gray shades still requires background knowledge on the SOM methodology. One of the main goals in data exploration is to automatically form overviews that are immediately understandable. An analyst could use such visualizations as reports for executives or laymen, and the visualizations could even form the basis of a decision support system.

Color is a good candidate for grouping data according to its properties; color-based visualizations of data groups have already been used so pervasively that their interpretation is common knowledge. We have introduced a method for assigning colors to SOM displays such that the perceptual differences of the colors reflect the distances in the original data space as faithfully as possible. Similar data are colored similarly.

This seemingly simple task is actually far from trivial if the data space is high-dimensional, like the gene expression space is. The coloring corresponds to a (non-linear) projection of the model vectors of the SOM to the color space. We have used a constant-lightness "slice" of the CIELab color space, and discouraged very non-saturated colors. Then the resulting colors differ predominantly only in one perceptual dimension, the hue, which makes the interpretation easier. The available color space is further bounded by the limits of the display device, and hence the projection must be able to flexibly conform to a very irregular-shaped two-dimensional area while still remaining ordered and retaining color resolution. A suitable yet computationally relatively demanding projection method has been described in more detail in [5].

When the SOM of Figure 2**a** is projected to the color space it can be seen that all distances between neighboring model vectors are relatively uniform (Fig. 3**a**). The color will then change in a uniform fashion on the map. The new cluster display in Figure 2**b** can be projected similarly, resulting in a display where the found cluster borders are reflected as rapid changes in colors (Fig. 3**b**). Different clusters will then get different hues. The hue will change more slowly within the clusters, while still reflecting their inner substructures. The resulting colored maps can be viewed at `http://www.cis.hut.fi/sami/wsom01/`.

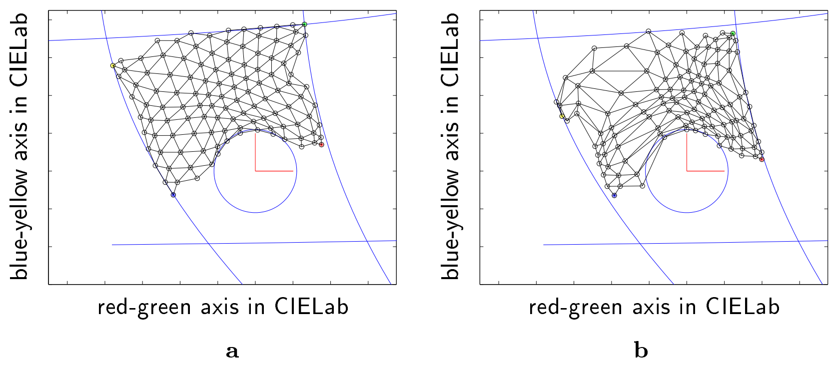

Fig. 3. Projection of the SOM grid onto the constant-lightness cross-section of the color space by preserving **a** the pairwise distances between neighboring model vectors, and **b** the indicators of cluster borders (shades of gray in Fig. 2**b**). The outer borders denoted by the long curves limit the colors displayable by a typical CRT tube, and a set of non-saturated colors in the middle has been encircled.

3.2 Interpretation of the Structures

The SOM displays are commonly interpreted utilizing three kinds of visualizations. Well-known samples, or distributions of classes of samples as in Figure 1, can be used as "landmarks" on the map. Another option is to plot the model vectors as curves or histograms at each SOM grid point. In the present case the model vectors are typical gene expression profiles that could be depicted within the map units. The third possibility is to plot the distribution of the values of the original data variables, or even some new variables, with gray shades on the map. Visualizations of the values of a variable in the model vectors are often called "component planes."

The problem with all these interpretation methods is that for large maps and large, high-dimensional data sets it may be difficult to uncover the essentials. Summaries of relevant information are therefore needed. Our approach to interpretation is to interactively choose an area of the map, usually a cluster, and characterize it by showing what distinguishes it from its surroundings.

In earlier works we have computed contributions of the original variables to the "local factors" that describe the local change in the SOM [2]. A cluster is then characterized by the set of variables having the largest contributions on the cluster borders. In this paper a computationally simpler and perhaps more easily interpretable method is demonstrated.

A sample cluster is characterized in Figure 4 in terms of how the data variables change on its borders. The average difference between neighboring model vectors on different sides of the border is computed and shown as a bar graph. From the visualization it can be seen that, for instance, the genes in the cluster are consistently more up-regulated in the DTT shock treatment than the genes in the surroundings. This is biologically understandable since the shock affects protein folding, and misfolded proteins must be degraded. It is known from Figure 1 that the cluster contains proteins related to cytoplasmic degradation.

Note that to characterize *any* area of the map, not only a clear-cut cluster, it can be interpreted in the same fashion. The result then tells in what sense the data within an area is similar, and in what sense it is different from the data of the surrounding areas.

4 Conclusions

The main conclusion of this paper is that SOM-based exploratory data analysis methods are promising tools for bioinformatics applications, in particular for the analysis of gene expression and function. The investigation of the biological meaning of the discoveries is still underway, but already now it is safe to say that the methods are applicable.

The cluster structures in gene expression data were visualized in new ways which emphasize cluster borders and make the clusters easily interpretable by

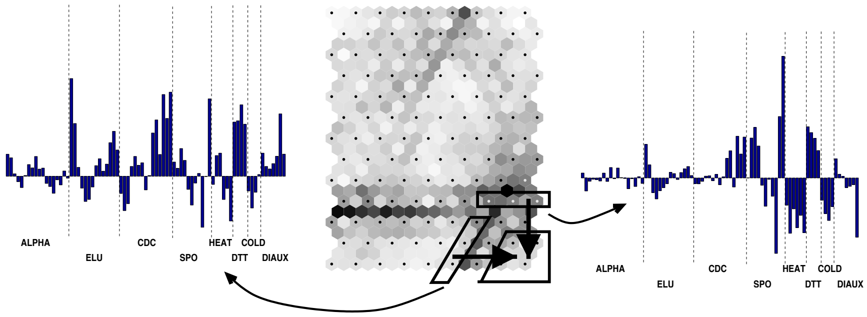

Fig. 4. A characterization of a sample cluster in terms of the original variables. The profiles show how the expression of the genes changes (on the average) when moving on the SOM according to the arrows. The changes can be used to characterize what is special in the expression of the genes. Abbreviations: 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, DIAUX=diauxic shift.

coloring the data according to their similarity. The clusters were interpreted by computing characteristic changes occurring in the data at the cluster borders.

The present study was based on the usual inner-product similarity measure between gene expression profiles. It is evident, however, that not all time points and treatments are equally important and they should be weighted according to their importance. More generally, the metric of the data space should concentrate on measuring the important dissimilarities. In subsequent studies we will apply new methods [4] that *learn* a suitable distance measure. The metric is learned from data in which the expression profiles have been paired with auxiliary data that implicitly defines what is important. For gene expression data the auxiliary data could denote for instance the functional class of the proteins the genes encode.

Acknowledgement

The research was supported by the Academy of Finland, in part by the grant 50061. I would like to thank our bioinformatics group, the bioinformatics group of the A.I. Virtanen Institute of the University of Kuopio, and Mr Jarkko Venna for the contributions.

References

1. S. Kaski, J. Kangas, and T. Kohonen. Bibliography of self-organizing map (SOM) papers: 1981–1997. *Neural Computing Surveys*, 1(3&4):1–176, 1998.

Available in electronic form at http://www.icsi.berkeley.edu/~jagota/NCS/: Vol 1, pp. 102–350.
2. S. Kaski, J. Nikkilä, and T. Kohonen. Methods for exploratory cluster analysis. In *Proceedings of SSGRR 2000, International Conference on Advances in Infrastructure for Electronic Business, Science, and Education on the Internet*. Scuola Superiore G. Reiss Romoli, 2000. (Proceedings on CD-ROM, ISBN 88-85280-52-8).
3. S. Kaski, J. Nikkilä, P. Törönen, E. Castren, and G. Wong. 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. Accepted for publication.
4. S. Kaski, J. Sinkkonen, and J. Peltonen. Bankruptcy analysis with self-organizing maps in learning metrics. *IEEE Transactions on Neural Networks*, 2001. Accepted for publication.
5. S. Kaski, J. Venna, and T. Kohonen. Coloring that reveals cluster structures in multivariate data. *Australian Journal of Intelligent Information Processing Systems*, 6:82–88, 2000.
6. T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.
7. T. Kohonen. *Self-Organizing Maps*. Springer, Berlin, 1995. (Third, extended edition 2001).
8. P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrowsky, E. S. Lander, and T. R. Golub. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences, USA*, 96:2907–2912, 1999.
9. P. Törönen, M. Kolehmainen, G. Wong, and E. Castrén. Analysis of gene expression data using self-organizing maps. *FEBS Letters*, 451:142–146, 1999.
10. A. Ultsch. Self-organizing neural networks for visualization and classification. In O. Opitz, B. Lausen, and R. Klar, editors, *Information and Classification*, pages 307–313. Springer-Verlag, Berlin, 1993.
11. J. Venna and S. Kaski. Neighborhood preservation in nonlinear projection methods: An experimental study. In *Proceedings of ICANN'01, International Conference on Artificial Neural Networks*. 2001 Submitted.