Analysis and visualization of gene expression data using Self-Organizing Maps [1]

Janne Nikkilä [a,*,2], Petri Törönen [b,2], Samuel Kaski [a], Jarkko Venna [a], Eero Castrén [b], Garry Wong [b]

[a] *Helsinki University of Technology*

Neural Networks Research Centre

P.O.Box 9800, 02015 HUT, Finland

fax +358-9-755 4892

[b] *University of Kuopio, A.I.Virtanen Institute*

P.O.Box 1627, 70211 Kuopio, Finland

---

* Corresponding author (reprint requests)

Email addresses: janne.nikkila@hut.fi (Janne Nikkilä), toronen@hytti.uku.fi (Petri Törönen), samuel.kaski@hut.fi (Samuel Kaski), jarkko.venna@hut.fi (Jarkko Venna), eero.castren@uku.fi (Eero Castrén), garry.wong@uku.fi (Garry Wong).

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

[2] These authors contributed equally to this work.

Analysis and visualization of gene expression data using Self-Organizing Maps

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 yeast genes. The relationships of known functional classes of genes are investigated by analyzing their distribution on the SOM, the cluster structure is visualized by the U-matrix method, and the clusters are characterized in terms of the properties of the expression profiles of the genes. Finally it is shown that the SOM visualizes the similarity of genes in a more trustworthy way than alternative methods, multidimensional scaling and hierarchical clustering.

Key words: Clustering, exploratory data analysis, gene expression, information visualization, Self-Organizing Map (SOM)

The list of symbols

$\mathbf{x}_k$ sample vector in input space $\mathbb{R}^n$

$\mathbf{m}_i$	model vector of the Self-Organizing Map
$c(\mathbf{x})$	the index of the model vector closest to $\mathbf{x}$
$h_{c(\mathbf{x}),i}$	neighborhood function
$C_k(\mathbf{x}_i)$	the set of those k data vectors that are closest to $\mathbf{x}_i$ in the original data space
$\hat{C}_k(\mathbf{x}_i)$	the set of those k data vectors that are closest to $\mathbf{x}_i$ after the projection
$U_k(\mathbf{x}_i)$	the set of data vectors $\mathbf{x}_j$ for which $\mathbf{x}_j \in \hat{C}_k(\mathbf{x}_i) \wedge \mathbf{x}_j \notin C_k(\mathbf{x}_i)$ holds
$V_k(\mathbf{x}_i)$	the set of data vectors $\mathbf{x}_j$ for which $\mathbf{x}_j \notin \hat{C}_k(\mathbf{x}_i) \wedge \mathbf{x}_j \in C_k(\mathbf{x}_i)$ holds
$r(\mathbf{x}_i, \mathbf{x}_j), i \neq j$	the rank of $\mathbf{x}_j$ when the data vectors are ordered based on their Euclidean distance from the data vector $\mathbf{x}_i$ in the original data space
$\hat{r}(\mathbf{x}_i, \mathbf{x}_j), i \neq j$	the rank of $\mathbf{x}_j$ when the data vectors are ordered based on their distance from the data vector $\mathbf{x}_i$ after the projection

1 Introduction

DNA microarray technologies (cf. Schena et al., 1995; DeRisi et al., 1997; Chu et al., 1998) provide 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 or conditions, or after sampling from different tissue sources. Examples of some possible comparisons 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 the 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 unsupervised analysis of gene expression data have focused on three important aspects: preprocessing and feature extraction from the data, clustering, and visualization. The goal of preprocessing and feature extraction is to transform the data into a suitable form for subsequent analysis. The goal of clustering is to group the data into meaningful sets. The goal of visualization is to present the data in a way that both similarities and differences can be seen.

While the initial intent was to profile the expression patterns of individual genes with microarrays, the ability to cluster these patterns on a genome-wide scale and to access the pertinent genes in these clusters, has expanded the utility of microarrays to inferring the function of specific genes. Although the biological validation of hypotheses derived from microarray data remains necessary, the reliance on microarray generated data for individual gene information has risen to the forefront. On a larger scale, the analysis of many combined microarray data sets has taken this a step further to characterize more sophisticated biological phenomena such as cancer, development, and psychosocial effects.

Early microarray generated data sets contained perhaps 3000-6000 genes and up to 10 conditions or time points. While clustering and visualization of this data is not a trivial task, a variety of algorithms have been used and have proved to be extremely useful. Nonetheless, currently available technology can produce microarrays with the number of genes to be analyzed nearly an order of magnitude higher. Moreover, public access to data sets has allowed for the analysis of hundreds of different treatments or time points for a given set of genes. Such rapidly escalating complexity in gene expression data sets requires improved methods for both their analysis and visualization, if the data generated are going to be useful. The ability to visualize, as a means of understanding the relationships between genes and treatments in an ever increasingly complex data set environment, is currently an immense and ongoing challenge.

The potential for clustering methods including SOM to study gene function was recognized early on in yeast experiments. Initial publications showed that genes within particular metabolic pathways fell into the same clusters (DeRisi et al., 1997; Chu et al., 1998) and this led to the proposal that clustering could be used to predict gene function. These ideas were more formally discussed and methods illustrated soon thereafter (Eisen et al., 1998). The yeast system seems particularly amenable to gene function prediction based on clustering since essentially all genes from the organism are available on microarrays and many of the genes (>35%) have been assigned formal functions. Internal validation for gene function prediction has come from currently available data sets that include results derived from hundreds of experiments. Clustering has identified a large set of genes ($\sim$900) that show a similar regulation to different types of drastic environmental changes (Gasch et al., 2000), and this demonstrates internal validation for a set of stress responsive genes. In contrast, there has been a paucity of external validation for new gene function assignments that are based on independent methods such as gene knock-out and complementation techniques. The rate at which new gene function assignments have been proposed based on clustering has been much more rapid than actual new assignments based on other more traditional laboratory methods.

In more complex biological systems such as mice, clustering expression profiles from different developmental stages and body regions has revealed clusters that have ubiquitous or tissue specific expression. The data itself reveals an initial characterization of gene functions (Miki et al., 2001). In that study, clustering also revealed a remarkable coordination of gene expression within 78 metabolic pathways among different tissues, thus providing internal validation for the clustering and assignment of functions.

One of the most useful applications of clustering to predict gene function is in medical diagnostics. This application for microarray analysis and clustering methodologies is currently an especially active area. Clustering methods have been used on microarray data to distinguish tumor types, to predict clinical outcomes, or to predict responsiveness to therapy. Clustering strategies to detect biomarkers as well as to predict outcome have been applied to breast tumors (Perou et al., 2000), leukemia (Golub et al., 1999), prostate cancers

(Dhanasekaran et al., 2001), esophageal cancers (Selaru et al., 2002), and nervous system tumors (Pomeroy et al., 2002) to name a few. Hierarchical clustering, K-means, and SOMs have been used. Finally, it should be noted that in order for clustering methods to become more trusted and useful as diagnostic tools, the level of annotation and a common ontology similar to the standard found in yeast needs to be brought forth. While the human and mouse data is moving toward this goal, the functions of single genes may be pleotropic and thus assignment of a single gene to a single function may cover some but not all cases, adding further complexity to this problem.

Our laboratories have been using the Self-Organizing Map (SOM) as a rational method for analyzing gene expression data sets. The SOM is an unsupervised neural network algorithm that, after preprocessing of the data, can cluster the data into biologically meaningful groups. The first studies applying SOM to gene expression data were published by Tamayo et al. (1999) and Törönen et al. (1999). 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.

There seems to be a common misunderstanding in the bioinformatics community that in SOM each map unit should be regarded as a separate cluster. Actually several neighboring units may model a cluster together. In this work we will use the SOM for *visualizing* a high-dimensional data set of 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

In this section we will briefly introduce the basic SOM algorithm and discuss its relationship with alternative methods. In data analysis applications the SOM can be complemented with a set of tools that help in interpreting the data. The tools will be introduced in Section 3, in connection with analysis of gene expression data.

2.1 The SOM

The Self-Organizing Map (SOM; Kohonen 1982, 1995) is a discrete grid of map units (the hexagons in Fig. 2). Each map unit can represent certain kinds of data, and in the present paper the units represent genes expressed in similar ways in a 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 represent progressively more different profiles. The map is a similarity diagram that presents an overview of the mutual similarity of the large number 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[3], for which the closest model vector $c(\mathbf{x})$ must fulfill the condition

$$\mathbf{x}^T \mathbf{m}_{c(\mathbf{x})} \geq \mathbf{x}^T \mathbf{m}_i \tag{1}$$

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

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

and normalized to unit length. Here $h_{c(\mathbf{x}),i}(t)$ is the *neighborhood function*, a decreasing function of the distance of the units $c(\mathbf{x})$ 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

[3] Note that normalization may enhance noise if the norm of a vector is very small, i.e. if a gene is expressed in a similar way in the treatments(s) and in the control condition. It would probably be advisable to discard such genes from the computation; in this paper we did not investigate this issue further.

of computing SOMs see (Kohonen, 1995; Kaski et al., 1998).

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.

Clustering methods reduce the number of data samples by grouping similar 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. We will use the term cluster below to denote the group of data (here genes) occurring in a set of neighboring SOM units having close-by model vectors. Suitable visualizations such as the U-matrix used in Section 3 can aid in gaining an overview of the cluster structure and extracting clusters having arbitrary shapes. Note that we acknowledge that it is impossible to give an exact definition of a cluster. There simply are more and less dense domains in the data space, and fundamentally it is variations in this density structure we wish to study.

The hierarchical clustering methods commonly used in biology, and already applied to gene expression data (e.g. Eisen et al., 1998), visualize a tree of sub-clusters 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, complemented with suitable visualization methods, can provide an overview of the data collection. Interactive visualization and analysis methods can be used for extracting and interpreting clusters in the data. In (Oja et al., 2002) it has been demonstrated that similar inferences can be made with the SOM and with hierarchical clustering.

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 (Venna and Kaski, 2001) the similarity diagrams formed by the SOM are more trustworthy than alternative methods, in the sense that if two data points are close-by on the display they are 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. The same kind of a study will be repeated for gene expression data in Section 3.3. Hierarchical clustering is included in the comparison.

3 Gene expression analysis

In an earlier study we have demonstrated with a small gene data set that the SOM-based exploratory tools are useful for analyzing gene expression data and functional classes (Kaski et al., 2001). We now aim at producing new biological hypotheses about the density (i.e. cluster) structure of the yeast gene expression space and the relationships of that data space to the functional classification of the genes.

We clustered the expression profiles of genes of the budding yeast *Saccharomyces cerevisiae*. The expression data was collected by microarrays during diauxic shift (DeRisi et al., 1997), the mitotic cell division cycle (Spellman et al., 1998), sporulation (Chu et al., 1998), and shock treatments (Eisen et al., 1998), forming a 79-dimensional data vector for each gene. The original data is available at `http://genome-www.stanford.edu/clustering/`. The data included 2460 genes. These were the genes for which the functional classification was known at the time of the creation of the data. The formed expression vectors were normalized to unit-length, but no other preprocessing was used. The data also included a short verbal description of the function of each gene. The descriptions were used in the analysis by doing keyword searches for certain specific functions.

The genes were classified to functional categories according to the MIPS yeast functional classification catalog[4] (Mewes et al., 2002). Some of these functional classes were known to be regulated in these treatments (cf. Pavlidis et al., 2001; Eisen et al., 1998).

3.1 Methods

The computation and analysis of the SOM was made with the (modified) SOM_PAK and SOM_Toolbox (for MATLAB) program packages available at `http://www.cis.hut.fi/research/software.shtml`. When computing a SOM a few of its properties and parameters need be selected, and finally it must be ascertained that the final SOM is of good quality. Advice will be given below; more advice is available in the documentation of the program packages, in the book (Kohonen, 1995), and in the SOM literature (see Kaski et al., 1998). The SOM is not particularly sensitive to choices of its size and

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

other parameters, although they do affect the results.

The size of the SOM determines the resolution of the visualization it produces. On a small SOM lots of data samples will be projected to each SOM unit, whereas on a large SOM the similarity relationships of the samples are more readily visible. We chose a good resolution, a SOM of 30x50 units, resulting in less than 2 data points in a unit on the average.

The topology of the SOM grid was chosen to be the plain two-dimensional rectangle, the rationale being that it is easy to visualize and to interpret. The topology of the grid was hexagonal, which is more homogeneous with respect to the directions on the SOM plane than the other frequently used alternative, rectangular topology. The results are usually very similar with both choices, though.

The second parameter needing a choice is the "stiffness" of the SOM, which is governed by the ratio of the final neighborhood width to the size of the SOM (its side length). Based on preliminary experiments, the final standard deviation of the Gaussian neighborhood function was chosen to be 2.0 times the distance of neighboring units on the SOM grid. One method for choosing the stiffness has been presented by Kaski and Lagus (1996).

The number of iterations in the computation process should be large enough to guarantee proper organization and convergence. Here we took a conservative approach. The SOM was computed in two phases. In the first phase (500,000 iterations) the neighborhood radius (standard deviation of the Gaussian neighborhood function) was large, starting from 14 units and linearly decreasing to 4 units. The height of the neighborhood function decreased linearly from 0.5 to 0. This phase was responsible for the rough ordering of the map.

The fine tuning of the map took place in the second phase. In this phase, during 3,000,000 iterations, the neighborhood radius decreased from 4 to 2 and the height from 0.02 to 0. As a result of the conservative large number of iterations, the computation time for a modern workstation with an Alpha processor took about 19 hours. The computation time scales quadratically with respect to the size of the SOM. There exist shortcut methods for speeding up the computation (see Kohonen et al., 2000); in this study we did not apply them.

Finally, the quality of the resulting map needs be ascertained. Several measures of goodness have been proposed. We used the (local) objective function of the SOM, which for the inner product metric reads

$$E = \sum_{\mathbf{x}} \sum_{i} h_{c(\mathbf{x}),i} \mathbf{x}^T \mathbf{m}_i \ . \tag{3}$$

The c is defined by (1). The SOM having the largest value of the objective function was chosen from five randomly initialized SOMs.

Note that a stiff SOM is not particularly sensitive to noise: each model vector attains a value that is an average over a number of data points. The averaging cancels out noise.

Note finally that it is possible, in principle at least, that the SOM "folds" when it tries to represent a data set with a high (local) dimensionality (Kohonen, 1995). Such effects can be minimized by proper choice of the stiffness, and if necessary the SOM can still be monitored after it has been computed: The distance of each model vector from the set of other model vectors can be plotted as gray shades on the SOM display. Strong oscillations may indicate folding. Alternatively, the distance of each data sample from all model vectors can be visualized to check whether there are several local maxima of the inner product.

3.2 Visualizing gene expression data

The density or cluster structure of the data can be visualized using the so-called U-matrix (Ultsch, 1993). The SOM provides a map display, on which the map units are located as a lattice. The U-matrix complements the SOM. With it, the density of the data (i.e. the clusters) can be visualized as gray shades on top of the SOM display.

The U-matrix is simply a collection of pairwise distances between the model vectors of neighboring SOM units. In Figure 1 every second hexagon represents an actual map unit, and the distances have been visualized as gray levels of additional hexagons inserted between the SOM units. Long distances correspond to dark shades and short distances to light shades. The gray level of the SOM units is chosen to be the average of the three hexagons on top of it.

The intuitive justification for using the U-matrix distances to visualize data density is the following: it is known that the density of the model vectors reflects the density of the data items. For special cases the functional relationship between the densities has even been solved in closed form. Distances between close-by data points (and model vectors) are smaller in dense and larger in sparser areas. Note that since the SOM is an ordered description of the data, the model vectors of neighboring map units are the close-by model vectors in the input space as well. Distances between neighbor units thus approximate (a function of) the density of the data. The exact functional relationships between the distances, the density, and the gray shades are not important for exploratory purposes.

There are some clearly distinguishable cluster areas visible as light regions in Figure 1, for example in the upper and lower right corners. Genes clustered together are regulated similarly in the set of treatments chosen for the analysis. The natural question to ask next is whether the biological functions underlying the similarity in regulation are new or already known. We will analyze this issue in detail for some of the clusters in Section 3.4. In this section we demonstrate how to proceed with the exploratory analysis: of acquiring an overview of what the data and its cluster structures are like.

Information of the biological function of the genes is available in at least two forms: in the MIPS hierarchical functional classification (`http://mips.gsf.de/proj/yeast/CYGD/db/index.html`) and in functional textual annotations of the SGD database (`http://genome-www.stanford.edu/Saccharomyces/`). In the first stage we plotted the distribution of the genes in the MIPS functional classes on top of the SOM. Samples are shown in Figure 2. We spotted some interesting areas in which a certain functional class has a (relatively) high concentration. Since the MIPS functional classes do not include the most detailed information on gene functions, we checked further details from the SGD database, for example by looking for genes with keyword patterns such as 'RIBOSOMAL PROTEIN' or 'RIBOSOMAL PROTEIN, MITOCHONDRIAL'. Landmarks were added on the SOM of Figure 1 to indicate resulting areas; they are already known to correspond to a certain biological function. Note that not all areas have a one-to-one correspondence to the cluster structure; an example will be analyzed further in Section 3.4.

These kinds of SOM-based visualizations provide an overview to the very large gene expression database, by making some relationships between genes, and between functional classes explicit. They may reveal unexpected relationships; in this study the SOM visualizations inspired a considerable amount of hypotheses for further investigation. We will inspect some of them in more detail below; we have tried to provide representatives of the main types of findings. Note that it would have required a considerable effort to arrive at these hypotheses without the SOM-based visualizations.

3.3 Comparison of methods for visualizing the similarity of expression profiles

One of the main uses of nonlinear projection methods such as the SOM is to visualize multivariate data, and in such visualizations it is crucial that the visualized proximities can be trusted upon: if two data samples are close to each other on the display they should be close-by in the original space as well. We will define and use a local measure of trustworthiness to compare the SOM with alternative visualization methods.

Projection methods differ in the type of data set properties they try to preserve. A family of traditional methods, which are based on multidimensional scaling (MDS; Torgerson, 1952), try to preserve the pairwise distances of the data samples as well as possible. That is, the pairwise distances after the projection approximate the original distances. In a variant of nonlinear MDS, non-metric MDS (Kruskal, 1964), only the rank order of the distances is to be preserved. Another variant, Sammon mapping (Sammon, 1969), emphasizes the preservation of local (short) distances relative to the larger ones. In this work we will transform the inner product similarity measure to the Euclidean distance the MDS methods assume by $\|\mathbf{x}-\mathbf{y}\|^2 = (1-\mathbf{x}^T\mathbf{y})/2$, which holds for normalized vectors.

Hierarchical clustering can be considered as a partitional clustering method, when the dendrogram is cut at some level, or as a method for visualizing the similarities of the data points (and ultimately clusters). Here we will measure the visualization capacity. Hierarchical clustering can be used to order the data into a linear table according to the similarities revealed by the dendrogram. We will compare this ordering to orderings the other methods produce. Note that the linear ordering is not unique, and we have fixed the order with the method recommended by Eisen (see the documentation of the program package at `http://rana.lbl.gov/`): ordering according to a one-dimensional SOM.

It may be argued that the dendrogram reveals a more comprehensive picture of the similarities than the linear ordering. The ultrametric distance (see Jain and Dubes, 1998) is a proper measure for the similarity relationships the dendrogram induces. We will additionally measure the trustworthiness of the ultrametric distance.

The hierarchical clustering was computed with both the correlation and the inner product metric, and the better result (correlation) was chosen.

For the SOM the distances were measured along the map grid, weighted by the U-matrix values: dark areas correspond to larger distances than light areas. More specifically, the distances were measured along the shortest paths of the grid induced by the model vectors into the data space. This corresponds to the visual distances on the U-matrix display.

We will measure preservation of neighborhoods of the data points in the projection (Venna and Kaski, 2001). For finite data the *neighborhoods* of data vectors are defined to be sets consisting of the k closest data vectors, for relatively small k. Topological concepts have been defined in terms of "arbitrarily small" neighborhoods but for discrete data we have to resort to finite neighborhoods. When the data are projected, the neighborhood is preserved if the set of the k nearest neighbors does not change.

Two kinds of errors are possible. Either new data may enter the neighborhood of a data vector in the projection, or some of the data vectors originally within the neighborhood may be projected further away on the 2D graphical display, or the ordering induced by the hierarchical clustering. The latter kinds of errors result from *discontinuities* in the mapping, and they have been measured extensively when quantifying the neighborhood preservation of SOMs. As a result of discontinuities not all of the proximities in the original data are visible after the projections.

We argue that the former kinds of errors are even more harmful since they reduce the *trustworthiness* of the proximities or neighborhood relationships that are visible on the display after the projection: some data points that seem to be close to each other may actually be quite dissimilar.

If the data manifold is higher-dimensional than the display, then both kinds of errors cannot be avoided and all projection methods must make a trade-off. Here we will focus on the new property of trustworthiness. The discontinuities in the mapping will be measured by the preservation of neighborhoods to show the trade-offs.

In principle, the errors could be measured simply as the average number of data items that enter or leave the neighborhoods in the projection. We have used slightly more informative measures: trustworthiness of the neighborhoods is quantified by measuring how far from the original neighborhood the new data points entering a neighborhood come. The distances are measured as rank orders; similar results have been obtained with Euclidean distances as well (unpublished). Using the notation in the list of symbols the measure for trustworthiness of the projected result is defined as

$$M_1(k) = 1 - \frac{2}{Nk(2N - 3k - 1)} \sum_{i=1}^{N} \sum_{\mathbf{x}_j \in U_k(\mathbf{x}_i)} (r(\mathbf{x}_i, \mathbf{x}_j) - k) \,, \tag{4}$$

where the term before the summation scales the values of the measure between zero and one[5]. In case of ties all orderings for samples with equal distance are considered equally likely, and the trustworthiness will be the average. In other words, the trustworthiness measure $M_1(k)$ gets larger values the less genes from outside the original neighborhood get projected to the new neighborhood. Furthermore, the farther the outside genes are, the bigger the resulting value.

Preservation of the original neighborhoods is measured by

$$M_2(k) = 1 - \frac{2}{Nk(2N - 3k - 1)} \sum_{i=1}^{N} \sum_{\mathbf{x}_j \in V_k(\mathbf{x}_i)} (\hat{r}(\mathbf{x}_i, \mathbf{x}_j) - k) \,. \tag{5}$$

As can be seen from Figure 3**a**, the SOM is more trustworthy than the alternatives, except for very small neighborhoods (1 or 2 neighbors), for which hierarchical clustering is better. For neighborhoods larger than about 500 genes the Sammon projection and NMDS are better. We argue, however, that the smaller neighborhoods are more important in visual inspection of the results.

The other side of the coin is continuity of the projection. Here the SOM is

[5] For clarity we have only included the scaling for neighborhoods of size $k < N/2$.

clearly better than hierarchical clustering, although worse than the MDS-based methods (Fig. 3**b**).

3.4 Generating hypotheses based on the visualizations

3.4.1 Ribosomal proteins break into three sub-clusters.

SOM-based visualization provides a simple and rapid means of evaluating the data to generate new biological hypotheses. The SOM allowed us to discern 3 distinct groups of ribosomal proteins (encircled manually in Figure 4 **A**). The expression profiles of all genes which clustered within each of these respective groups can be seen in Figures 4 **B**, **D**, and **F**. When only genes with a "ribosomal protein" descriptor were displayed (Figures 4 **C**, **E**, and **G**) the similarities in expression patterns of genes within each group were even more evident.

Of the 174 genes in the bottom right group, 121 encoded ribosomal proteins, a major piece of machinery for translation, and 20 others were involved in translation initiation, elongation, or other protein synthesis activities. Moreover, 14 were involved in small molecule synthesis including those for the amino acids lysine, glutamine, serine, and histidine. Only 2 genes were described as mitochondrial and encoded a translocase component and a protein folding function. Of the 100 genes in the bottom left group, 32 encoded ribosomal mitochondrial proteins, 7 were involved in mitochondrial translation and protein synthesis, and 10 in mitochondrial respiration. Interestingly, 11 genes encoded proteins involved in tRNA, rRNA, or mRNA splicing. Of the 25 genes from the topmost box, 6 encoded ribosomal proteins of which all are mitochondrial. An additional 3 genes encoded mitochondrial proteins involved in respiration or oxidative phosphorylation.

While it has previously been shown that genes within certain functional classes (e.g. TCA cycle) are tightly co-regulated transcriptionally during certain cellular events, the SOM in this work has provided several interesting insights regarding protein synthesis. First, the SOM has divided the ribosomal proteins into 3 different groups, suggesting that protein synthesis may occur in at least 3 different modalities. Also, the visualization of the expression profiles in Figure 4 confirms that the three groups all have clear differences in their expression profiles. Taking data from the bottom right SOM units, the results would suggest that translation involves not only production of a huge number of proteins for the translational machinery, but also de novo synthesis of the constituent amino acids needed for the generation of proteins. In contrast, mitochondrial protein synthesis involves co-regulation with genes necessary for energy generation and splicing factors that produce parts (tRNAs, rRNAs, mRNAs) for the machinery of protein synthesis. Finally, the yeast gene YIL098C has been described as "putatively" in the respiration category. This gene was grouped with 32 mitochondrial ribosomal genes and 10 other mitochondrial respiration genes in the bottom left box, strongly suggesting that it is a gene involved in respiration. Taken together, visualization of gene expression using the SOM has provided a rapid and accessible means of not only discerning modes of function, but also of assigning genes to functional categories.

An additional finding also for protein synthesis from the visualizations in Figures 1 and 2 is that ribosomal RNA transcription genes and ribosomal protein genes are grouped to separate regions. This is surprising as one would expect these to be co-regulated.

3.4.2 Interpretation of clusters based on functional classes

The U-matrix (Fig. 1) reveals several clusters as light areas on the display. Some of the clusters have a clear interpretation: they consist mainly of genes from a certain functional class. For instance, in the cluster in the bottom right corner 90% of the genes belonged to a class of ribosomal proteins. This was revealed by isolating the dense cluster visible in the U-matrix (Fig. 1) and analyzing the class distribution of that cluster. The inspection of the distribution of the ribosomal protein class confirmed the results of the analysis (Figure 4 **A**). The proportion of the next largest, relevant functional class (protein destination) in that cluster was only 6%. Note that the functional classification is hierarchical and the genes may belong to multiple functional classes.

3.4.3 Clusters having no clear connection to functional classes

The U-matrix reveals additionally clusters containing genes from several classes. Such clusters may reveal new biological connections but their interpretation is harder; they need to be interpreted with more data-driven methods. Since the genes are clustered together, we know that they behave in a similar way in the set of treatments. The similarities are, however, hard to discover from among the high-dimensional data. We will next introduce methods that aid in mining the expression profiles.

Clusters found by the U-matrix 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 (cf. Törönen et al., 1999; Tamayo et al., 1999), and (3) by plotting the distribution of the original data variables on the map. The problem with the approaches (2) and (3) is that, for large maps and high-dimensional data, they present a huge amount of data. In this Section we introduce methods that enable focusing on chosen interesting properties, and apply them to 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. We measure which treatments and time points contribute most to the change. For example, in linear factor analysis such contributions are standard interpretation tools.

In this paper we characterize a map area in a straightforward and intuitive manner: (1) the expression profiles of the genes within the area are visualized, and (2) the average expression profiles within the area and its surroundings are computed, and the differences are visualized. The change in the profiles reveals what changes when moving out from the area on the map.

Two cluster areas are interpreted in Figures 5 and 6. The cluster in the upper left corner of the map was selected according to the U-matrix in Figure 1. It could not have been anticipated based on the known functional classes; it contains genes from several classes including cellular organization, cell growth, and metabolism.

The main feature (Fig. 5) distinguishing it from the area below it on the map is that its genes are clearly more up-regulated in sporulation. Further analysis of the sporulation time series and the actual model vectors (not shown) revealed that this cluster is characterized by up-regulation in the sporulation treatment, and an almost flat profile elsewhere. This indicates that all the genes in this cluster are connected to spore formation, regardless of their functional classification.

We also analyzed this cluster in more detail (not shown) by dividing it into 8 smaller blocks (4 columns and 2 rows; each block containing about 4*4 SOM units) and visualizing the expression profiles. This analysis revealed that the time-lag of the upregulation in the sporulation time series is smallest in the right brink of the cluster, and increases progressively towards left. This result agrees well with the earlier results obtained with other methods (Chu et al., 1998), and the discovery would have been hard with typical clustering methods.

The cluster in Figure 6 was chosen based on the U-matrix visualization as well. It turned out to contain mainly genes classified into the protein folding and stabilization classes and the stress response class; both may be involved in stress response. Comparison with the neighborhood (shown in Figure 6) reveals that the genes in the cluster are upregulated in the sporulation, heat-, and dtt-shock treatments. Further analysis of the data and the SOM confirmed the association with the stress response.

3.4.4 A small unexpected cluster

The problem with the U-matrix is that the visualization may be noisy and it may be difficult to distinguish clustered areas from it. We have developed a method (Kaski et al., 2002) that complements the U-matrix by detecting and emphasizing salient cluster borders. In the present data set, the resolution of the U-matrix seems sufficient and we have, for the main part, used it for visualizing the clusters. With the new method we found, however, some additional potentially interesting areas, including a small cluster that is not clearly visible in the U-matrix display. We analyze the finding in this section.

The new method aims at finding locations in the data space in which the gradient of the data density has a clear minimum and changes its direction. Such locations are at cluster borders. The search is carried out at the locations represented by the SOM units. The principle is demonstrated in Figure 7. The difference of the gradients at neighboring model vectors is compared in the same way as the model vectors are compared in the U-matrix. Finally, the difference of the gradients is visualized as gray shades in between the map units. If there is a dark stripe in the visualization, 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.

The gradient is estimated by mean shift analysis (Cheng, 1995): the direction of the gradient at $\mathbf{m}_i$ can be estimated by the direction from $\mathbf{m}_i$ towards the centroid of data within a suitable local kernel (for example a Gaussian) centered at $\mathbf{m}_i$. Instead of a kernel having a fixed width we use an adaptive resolution by computing the centroid of the k data points closest to $\mathbf{m}_i$.

The mean shift visualization is illustrated in Figure 8. The borders of the "histone protein cluster" are more clear in the mean shift visualization than in the U-matrix. When analysing the genes in the surrounding area we found a group of genes with "protein glycosylation" in their description (boxed area in Fig. 9**A**) that overlapped the histone protein cluster (Fig. 9**B**). The expression profile of all the genes within the box of Figure 9**A** is shown in Figure 9**C**. There were 46 genes in total, 10 of which had "protein glycosylation" in their description and 9 genes had "histone". Protein glycosylation is mainly associated with processed, secreted, or membrane proteins, but the histone proteins are associated with the nucleus. The result that they are both co-regulated is unexpected. A view of their expression profiles (Figure 9**D** and **E**, respectively) shows the similarity of their expression profiles during the cell cycle (the first 3 parts of the profile in panel **D** and **E**) and sporulation (fourth part of the profile). These results suggests that protein glycosylation may be involved in cell cycle and sporulation processes that are co-regulated with histone proteins. The similarity in expression during sporulation could perhaps be explained by the association in the spore formation or may indicate a novel link between these two protein classes.

3.4.5 Contributions of the variables

The preprocessing stage is one of the most important stages of data analysis. One of the important decisions for gene expression data is whether all treatments and all time points should be stressed equally. In SOM-based analysis, prior knowledge of importance can be incorporated by scaling the variances of the treatments and time points accordingly.

In this study we wanted to acquire an overview of the whole data instead of emphasizing any special property, so we used the original scales of the treatments. Since the treatments and time points have different variances, this choice will cause the different treatments to affect the analyses inequally.

It is evident from the earlier analysis (Kaski et al., 2001) and from the analysis made for this work (not shown) 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. We have taken this into account when interpreting the results.

4 Conclusions

We have demonstrated the use of the Self-Organizing Map as a tool for exploratory analysis of gene expression data measured with DNA microarrays. New biological hypotheses were created for a set of genes with known functional classes. The results were consistent with the existing knowledge of functional classes of the genes.

The SOM-based visualizations offer an intuitive aid for forming hypotheses about data. We would also like to point out that SOM can be used in conjunction with other clustering methods to visualize clustering results (Vesanto and Alhoniemi, 2000), and that hierarchical SOMs could possibly be used to extend our present results.

Hierarchical clustering is an alternative, widely used visualization method. The functional classes of the genes could be visualized on the resulting dendrogram as easily as on a SOM. Nonetheless, the resulting display can be somewhat unwieldy, if there are thousands of nodes in the dendrogram, and hundreds of treatments. The principal advantage of the SOM-based display in comparison to such a very large dendrogram is the ability to visualize the entire data set. In this work we have additionally shown that SOM-based visualizations are more trustworthy.

Biological hypotheses were generated even for relatively well-known yeast genes. For example the class "ribosomal proteins" was shown to consist of two or even more subgroups, suggesting that the functional classification of this larger group could be further refined to take into account their different expression patterns and proposed functions during the varying treatments. Our analysis has also shown that clusters with the tightest co-regulation, such as the cluster of histone proteins, ribosome proteins, or proteasome complexes appear to be clusters of proteins belonging to a particular interacting protein complex.

In future work, we will apply the methods to less well characterized data, and explore in greater detail the hypotheses that will be generated from such an investigation.

References

Cheng, Y., 1995. Mean shift, mode seeking, and clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence 17, 790–799.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., Herskowitz, L., 1998. The transcriptional program of sporulation in budding yeast. Science 282, 699–705.

DeRisi, J. L., Iyer, V. R., Brown, P. O., 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278, 680–686.

Dhanasekaran, S. M., Barrette, T. R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K. J., Rubin, M. A., Chinnaiyan, A. M., 2001. Delineation of prognostic biomarkers in prostate cancer. Nature 412, 822–6.

Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. Proceedings of National Academy of Science of the USA 95, 14863–14868.

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., Brown, P. O., 2000. Genomic expression programs in the response of yeast cells to environmental changes. Molecular Biology of the Cell 11, 4241–57.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., Lander, E. S., 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286, 531–537.

Jain, A., Dubes, R., 1988. Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, NJ.

Kaski, S., Kangas, J., Kohonen, T., 1998. Bibliography of self-organizing map (SOM) papers: 1981–1997. Neural Computing Surveys 1 (3&4), 1–176, available in electronic form at http://www.icsi.berkeley.edu/∼jagota/NCS/: Vol 1, pp. 102–350.

Kaski, S., Lagus, K., 1996. Comparing self-organizing maps. In: Proceedings of International Conference on Artificial Neural Networks (ICANN) 1996. pp. 809 – 814.

Kaski, S., Nikkilä, J., Törönen, P., Castrén, E., Wong, G., 2001. Analysis and visualization of gene expression data using self-organizing maps. In: Proc. of NSIP-01, IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing.

Kaski, S., Nikkilä, J., Kohonen, T., 2002. Methods for exploratory cluster analysis. In: Szczepaniak, P. S., Segovia, J., Kacprzyk, J., Zadeh, L. A. (Eds.), Intelligent Exploration of the Web. Springer, Berlin. Accepted for publication.

Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. Biological Cybernetics 43, 59–69.

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

Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., Saarela, A., 2000. Self organization of a massive document collection. IEEE Transactions on Neural Networks 11, 574–585.

Kruskal, J. B., 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrica 29, 1–26.

Mewes H.W., Frishman D., Guldener U., Mannhaupt G., Mayer K., Mokrejs M., Morgenstern B., Munsterkotter M., Rudd S., Weil B., 2002. MIPS: a database for genomes and protein sequences. Nucleic Acids Research 30(1),31–4.

Miki, R., Kadota, K., Bono, H., Mizuno, Y., Tomaru, Y., Carninci, P., Itoh, M., Shibata, K., Kawai, J., Konno, H., Watanabe, S., Sato, K., Tokusumi, Y., Kikuchi, N., Ishii, Y., Hamaguchi, Y., Nishizuka, I., Goto, H., Nitanda, H., Satomi, S., Yoshiki, A., Kusakabe, M., DeRisi, J. L., Eisen, M. B., Iyer, V. R., Brown, P. O., Muramatsu, M., Shimada, H., Okazaki, Y., Hayashizaki, Y., 2001. Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays. Proceedings of the National Academy of Sciences of the USA 98, 2199–204.

Oja, M., Nikkilä, J., Törönen, P., Wong, G., Castrén, E., Kaski, S., 2002. Exploratory clustering of gene expression profiles of mutated yeast strains. In: Zhang, W., Shmulevich, I. (Eds.), Computational And Statistical Approaches To Genomics. Kluwer. In press.

Pavlidis, P., Weston, J., Cai, J., Grundy, W. N., 2001. Gene functional classification from heterogeneous data. In: Proceedings of the Fifth International Conference on Computational Molecular Biology (RECOMB), 249–255.

Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lonning, P. E., Borresen-Dale, A. L., Brown, P. O., Botstein, D., 2000. Molecular portraits of human breast tumours. Nature 406, 747–52.

Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., Kim, J. Y., Goumnerova, L. C., Black, P. M., Lau, C., Allen, J. C., Zagzag, D., Olson, J. M., Curran, T., Wetmore, C., Biegel, J. A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D. N., Mesirov, J. P., Lander, E. S., Golub, T. R., 2002. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 415, 436–42.

Sammon, Jr., J. W., 1969. A nonlinear mapping for data structure analysis. IEEE Transactions on Computers C-18, 401–409.

Schena, M., Shalon, D., Davis, R., Brown, P., 1995. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA microarray. Science 70, 467–470.

Selaru, F. M., Zou, T., Xu, Y., Shustova, V., Yin, J., Mori, Y., Sato, F., Wang, S., Olaru, A., Shibata, D., Greenwald, B. D., Krasna, M. J., Abraham, J. M., Meltzer, S. J., 2002. Global gene expression profiling in Barrett's esophagus and esophageal cancer: a comparative analysis using cDNA microarrays. Oncogene 21, 475–8.

Spellman, P., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., Futcher, B., 1998. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Molecular Biology of the Cell 9, 3273–97.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrowsky, E., Lander, E. S., Golub, T. R., 1999. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proceedings of the National Academy of Sciences of the USA 96, 2907–2912.

Torgerson, W. S., 1952. Multidimensional scaling: I. theory and method. Psychometrika 17, 401–419.

Törönen, P., Kolehmainen, M., Wong, G., Castren, E., 1999. Analysis of gene expression data using self-organizing maps. Federation of European Biochemical Societies Letters 451, 142–14.

Ultsch, A., 1993. Self-organizing neural networks for visualization and classification. In: Opitz, O., Lausen, B., Klar, R. (Eds.), Information and Classification. Springer-Verlag, Berlin, pp. 307–313.

Venna, J., Kaski, S., 2001. Neighborhood preservation in nonlinear projection methods: An experimental study. In: Dorffner, G., Bischof, H., Hornik, K. (Eds.), Artificial Neural Networks—ICANN 2001. Springer, Berlin, pp. 485–491.

Vesanto, J., Alhoniemi, E., May 2000. Clustering of self-organizing map. IEEE Transactions on Neural Networks 11, 586–600.

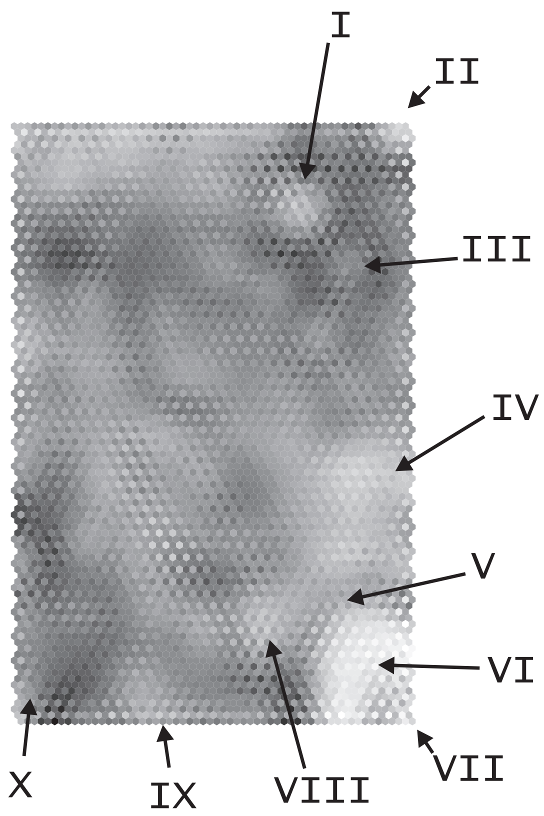
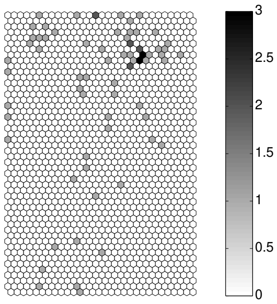
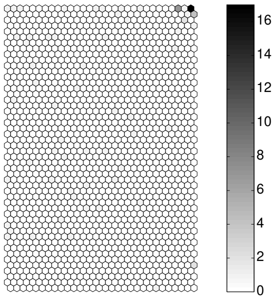


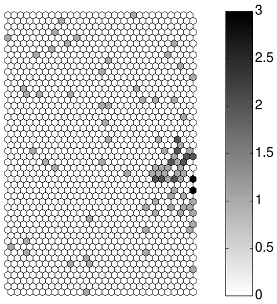
Fig. 1.



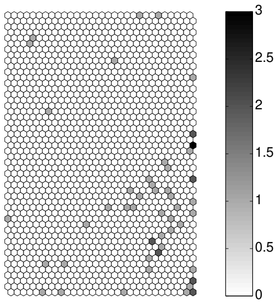
I: DNA synthesis



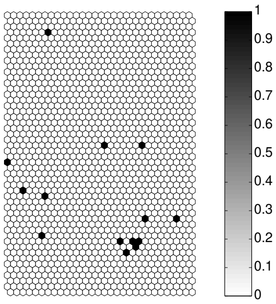
II: Cytoplasmic degradation



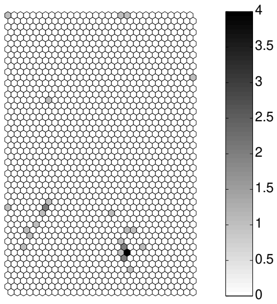
III: Ribosomal RNA transcription



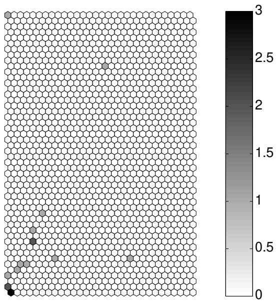
V and VII: Protein translation



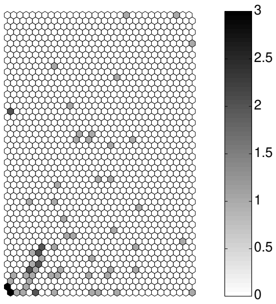
VIII: Fermentation



VIII and X: Glycolysis and glyconeogenesis



X: TCA cycle



X: Respiration

Fig. 2.

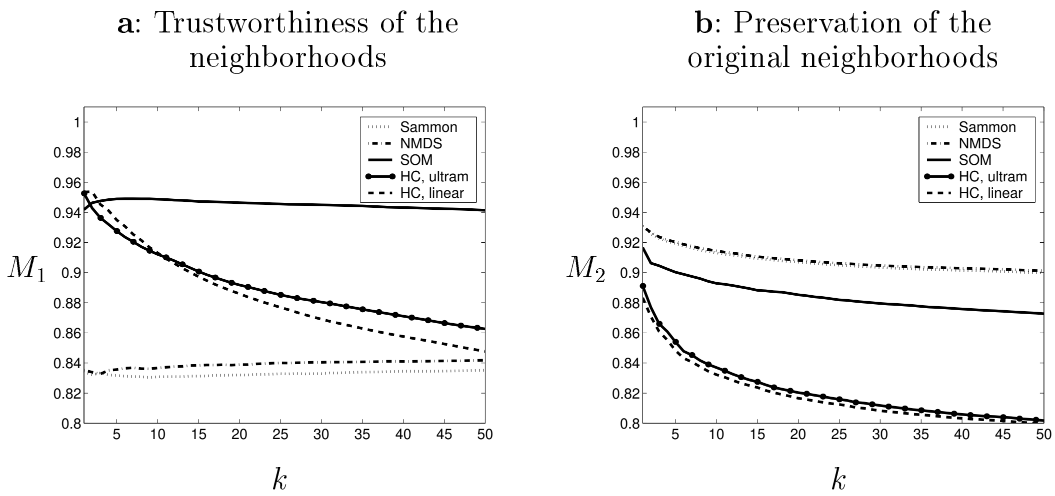

a: Trustworthiness of the neighborhoods
Sammon
NMDS
SOM
HC, ultram
HC, linear
M_1
k
b: Preservation of the original neighborhoods
Sammon
NMDS
SOM
HC, ultram
HC, linear
M_2
k

Fig. 3.

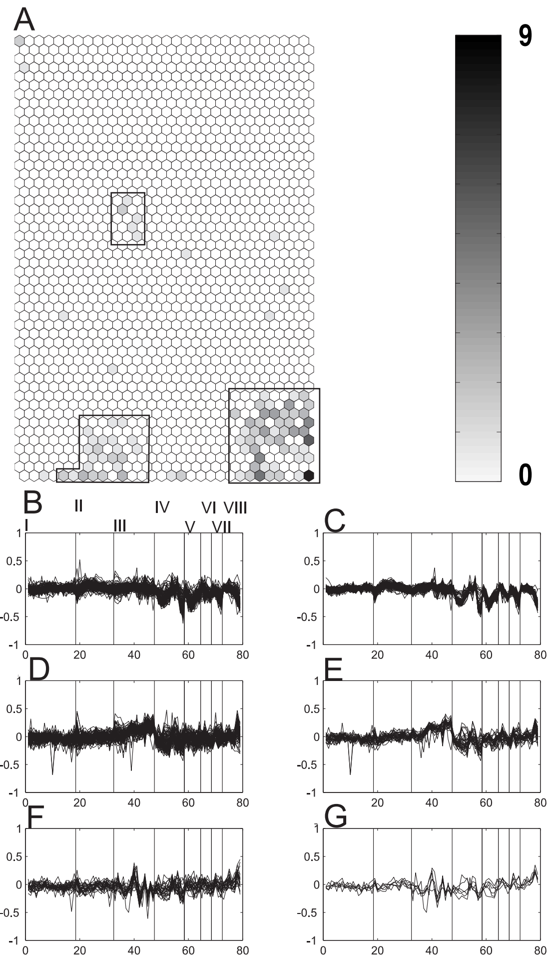


Fig. 4.

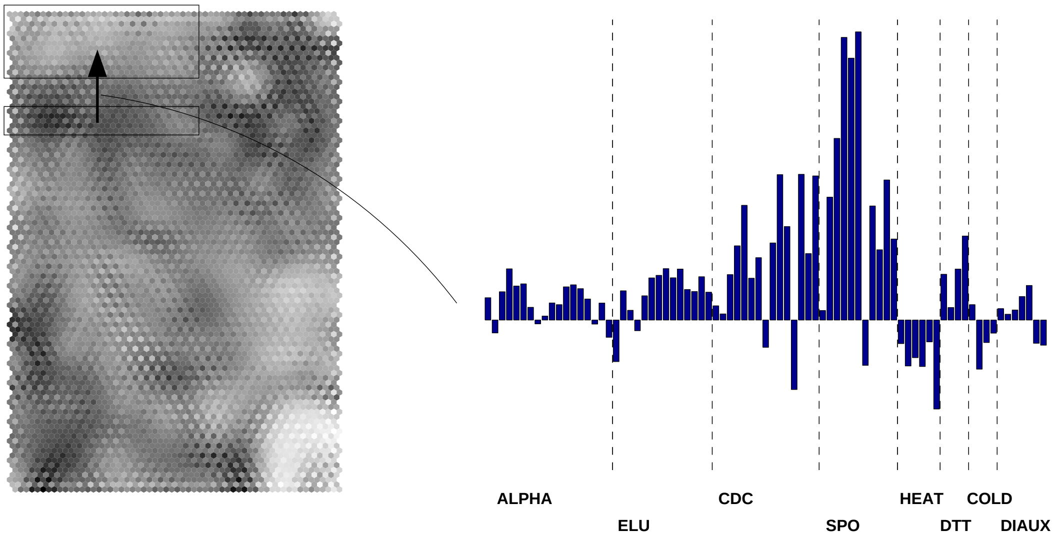


Fig. 5.

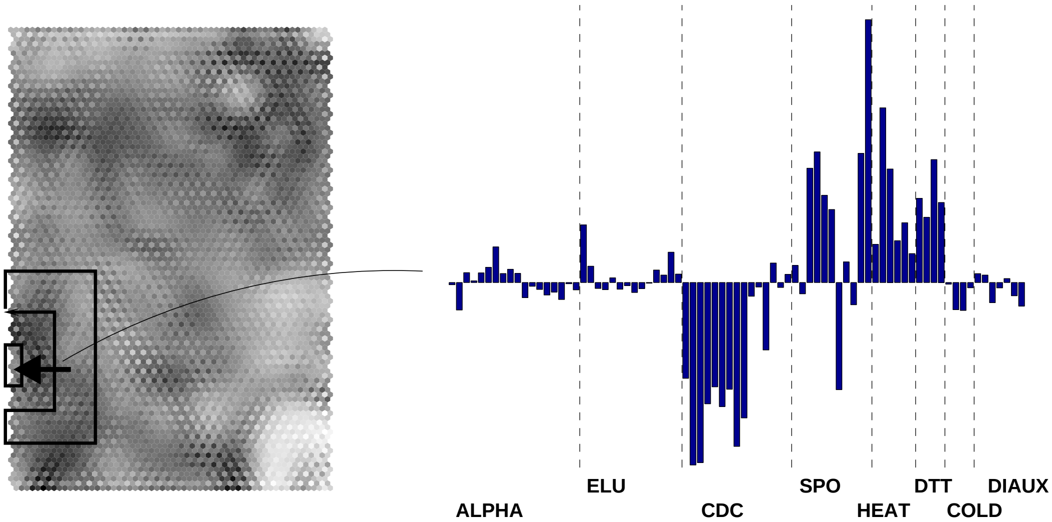


Fig. 6.

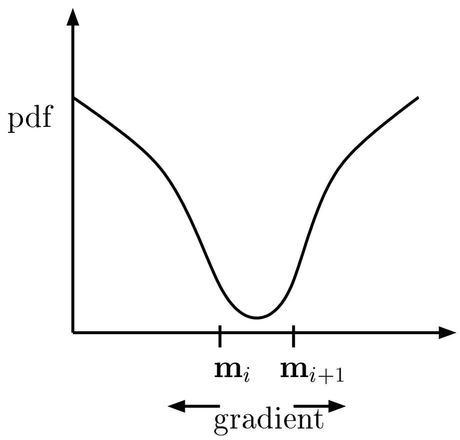


Fig. 7.

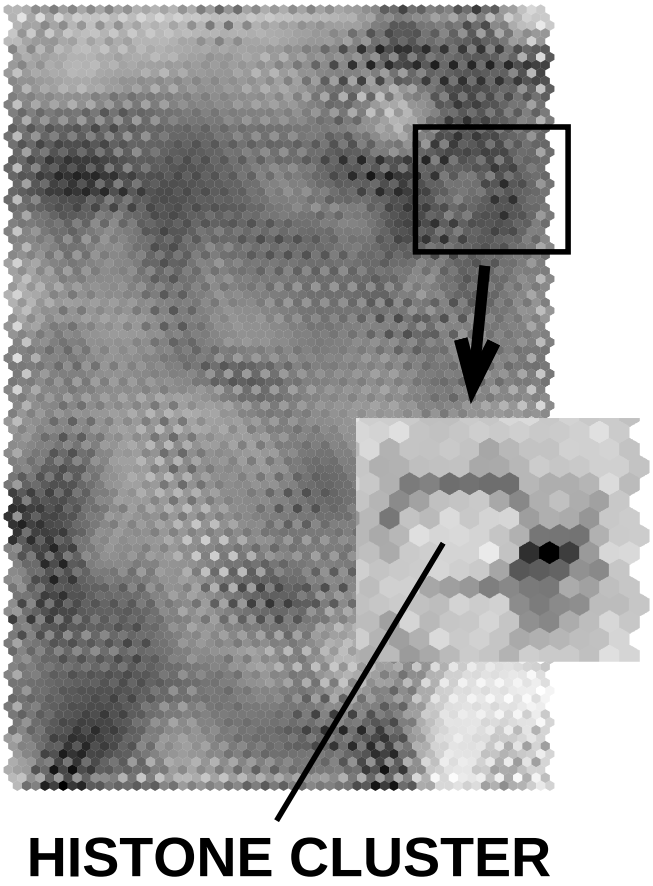


Fig. 8.

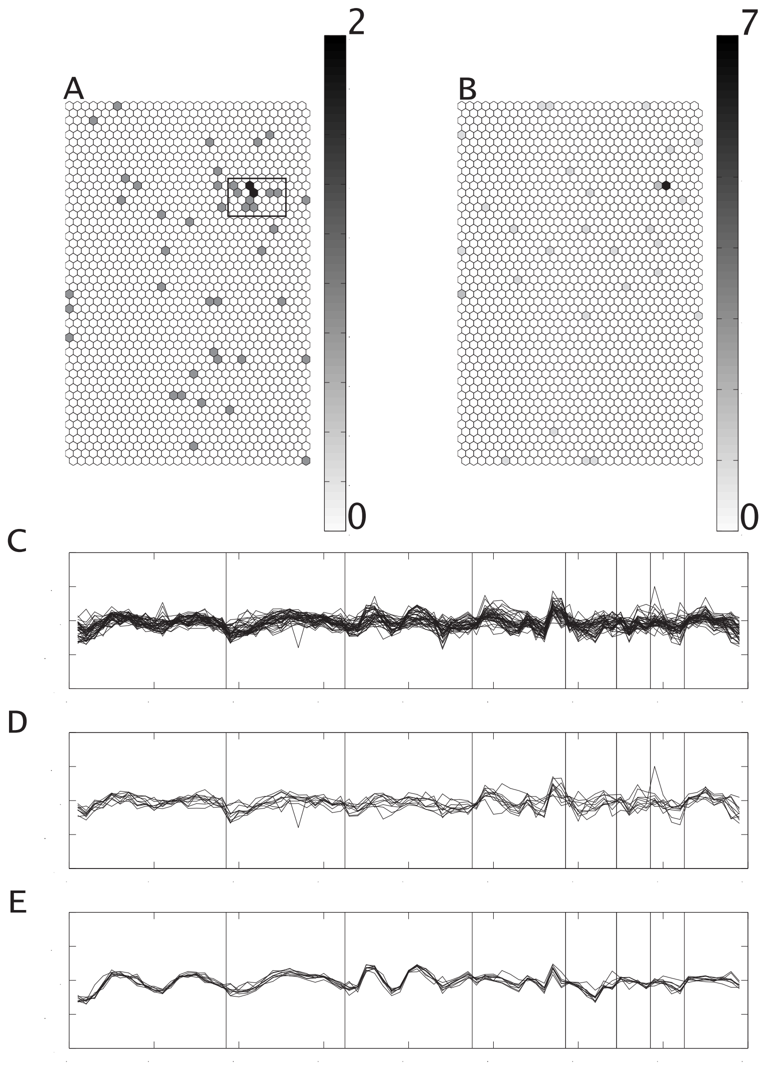


Fig. 9.

Fig. 1. The SOM of yeast gene expression data, with gray shades showing the density variations in the data. The dark regions represent sparse areas and white areas dense areas (i.e. clusters). In the U-matrix display every second hexagon is a SOM unit and every second unit represents the distance between the model vectors of the units. A set of landmarks, pointing to areas with relatively large content of a particular functional class of genes, have been added to the map. Distributions of the corresponding functional classes are investigated in more detail in Figure 2. I: DNA synthesis; II: cytoplasmic degradation; III: histone proteins; IV: ribosomal RNA transcription; V: protein translation; VI: protein synthesis; VII: ribosomal proteins and protein translation; VIII: glycolysis and fermentation; IX: mitochondrial ribosomal proteins; X: TCA cycle, respiration and glycolysis.

Fig. 2. Detailed distribution of genes of the functional classes marked onto the SOM in Figure 1 (the Roman numbers refer to Fig. 1). The areas III, VI, and IX will be analyzed in more detail in subsequent figures. Each hexagon denotes one SOM unit, and the scale to the right of each sub-image gives the number of genes of the class occurring in the SOM unit. Note that although the SOM is the same, in Figure 1 there are more hexagons since it is a U-matrix display.

Fig. 3. Trustworthiness of the neighborhoods after projection (**a**) and preservation of the original neighborhoods (**b**) as a function of the neighborhood size k.

Fig. 4. Groups of ribosomal proteins revealed by the SOM, and combined expression profiles of ribosomal genes in different groups. **A.** The SOM was constructed as described in the methods. The SOM units that contained genes with the descriptor "ribosomal protein" and their abundance were identified and are indicated by gray shades. The bar on right indicates the total number of ribosomal genes in each unit. The number of genes used for the SOM was 2460 and the number of treatments and time points were 79 (horizontal axis in **B-G**). The boxes indicate the boundary where genes were identified within the SOM units. Their expression profiles are shown in panels **B-G**. The number 0 represents no change in expression and values are normalized to a range of 1 to -1 (vertical axis). **B.** Expression profiles from all genes (174 total) grouped in the bottom left box across all 79 treatments or time points. **C.** Expression profiles from genes grouped in the bottom left box with "ribosomal protein" in their description (121 profiles). **D.** Expression profiles from all genes grouped in the bottom right box (103 profiles). **E.** Expression profiles from all genes grouped in the bottom right box (32 profiles) with "mitochondrial ribosomal protein" in their description. **F.** Expression profiles from genes grouped in the center box (25 profiles). **G.** Expression profiles with genes grouped in the center box with "mitochodrial ribosomal protein" in their description (6 profiles). The treatments were: **I**, cell-cycle, alpha factor arrest and release; **II**, cell-cycle, elutriation; **III**, cdc15 arrest and release; **IV**, sporulation; **V**, heat shock; **VI**, dithiothreitol (DTT) shock; **VII**, cold shock; **VIII**, diauxic shift.

Fig. 5. Characterization of the cluster in the upper left corner in terms of the data variables (time points in treatments). The profiles show how much the expression of the genes changes (on the average) when moving on the SOM according to the arrow. The changes can be used to characterize what is special in the expression of the genes of the clusters. 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(see `http://genome-www.stanford.edu/clustering/`).

Fig. 6. Characterization of the small left edge cluster in terms of the data variables (time points in treatments). The profiles show how much the expression of the genes changes (on the average) when moving on the SOM according to the arrow. The changes can be used to characterize what is special in the expression of the genes. Abbreviations are the same as in the Figure 5.

Fig. 7. A one-dimensional demonstration of the principle of the new method for locating cluster borders. The gradient (shown with arrows in the bottom) of the probability density function (pdf) of the data is evaluated at the locations of the model vectors, and salient turning points are detected by comparing the gradients at neighbor units.

Fig. 8. Mean shift visualization of the SOM. The large background image is the U-matrix, and the smaller image is a partial enlargement of the mean shift visualization showing more saliently the abrupt changes in the density structure. It would have been hard to detect this histone cluster based on the U-matrix alone. In both images dark shades denote sparser areas between potential clusters in the gene expression space.

Fig. 9. Analysis of the glycosylation genes and histone genes based on their distributions on the SOM. **A**. Shows the distribution and number of genes with the descriptor "glycosylation protein" in each SOM unit. The boxed area indicates units chosen for subsequent analysis. **B**. Shows the distribution and number of genes with the descriptor "histone protein" in each SOM unit. The expression profiles of all the genes found in the boxed area of the panel **A** are shown in the panel **C**. In panel **D** are the profiles of the genes with "protein glycosylation" in the descriptor, and in the panel **E** the profiles of the genes with "histone protein" in the descriptor. The 8 treatments and time points are the same as in Figure 4.