Wei Zhang and Ilya Shmulevich, editors,
Computational And tatistical Approaches To Genomics.
Kluwer Acadmic Publishers, 2002.

Chapter 1

EXPLORATORY CLUSTERING OF GENE EXPRESSION PROFILES OF MUTATED YEAST STRAINS*

Merja Oja[1], Janne Nikkilä[1], Petri Törönen[2], Garry Wong[2], Eero Castrén[2], and Samuel Kaski[1]

[1] *Helsinki University of Technology*
Neural Networks Research Centre
P.O. Box 9800, FIN-02015 HUT, Finland
{merja.oja,janne.nikkila,samuel.kaski}@hut.fi

[2] *University of Kuopio*
A. I. Virtanen Institute
P.O. Box 1627, FIN-70211 Kuopio, Finland
toronen@hytti.uku.fi,{garry.wong,eero.castren}@uku.fi

Abstract Yeast gene expression profiles are analyzed with self-organizing map (SOM)-based exploratory clustering and information visualization methods. Each profile describes the expression of the other yeast genes after one gene has been removed or its function has been blocked. Information about the relationships of the function of the genes is hidden in the relationships between the profiles. We demonstrate that the clusters found by the SOM are closely related to the clusters found earlier by hierarchical clustering. The advantage of the SOM is the intuitive visualization of the similarity relationships and cluster structures within the data set. We additionally compare different metrics for the functional classification of the genes.

Keywords: Bioinformatics, clustering, exploratory data analysis, gene expression, information visualization, self-organizing map (SOM), yeast

*This work was supported by the Academy of Finland, in part by the grants 50061 and 1164349.

Introduction

Genome projects are continuously revealing novel genes and the function of many of the genes discovered previously is unknown. The clarification of the function of these genes will be a great challenge and will require coordinated work by many different disciplines. Because many genes share a relatively high homology in different organisms, simple model organisms have been exploited in the analysis of gene function. For example, projects are underway to mutate every single gene in baker's yeast, *Saccharomyces cerevisiae* (Winzeler et al., 1999). One of the problems that has hampered functional analysis is that simple organisms display only a limited number of observable phenotypes and many mutations do not produce any phenotype at all.

Gene expression analysis has been recently proposed as an alternative means to analyse phenotypes of mutated yeast strains. Hughes et al. (2000) used the large mutation data set in an exploratory fashion to suggest inferences about gene function, for example about which pathway an uncharacterized gene belongs to. The global transcriptional response, i.e. the expression level of all the genes, of a cell after a mutation can be regarded as a detailed molecular phenotype. Tentative inferences on the similarity of function can be made based on similarity of the gene expression profiles, and the hypotheses can then be tested experimentally.

Hughes et al. (2000) clustered the mutations according to similarities in their expression profiles, to make similarities hidden within the large data set explicit. Hierarchical agglomerative clustering, a popular method for grouping gene expression data, was used. The hierarchical clustering methods visualize a tree of subclusters of increasing detail. The problem with the trees is that they are sensitive to small variations in the data that may affect the agglomeration process, and for large data sets the trees become huge. It is then hard to extract the essential cluster structures from them.

We have been using Self-Organizing Map (SOM)-based clustering and information visualization for the same purpose (Törönen et al., 1999; Kaski et al., 2001a). The SOM (Kohonen, 1982; Kohonen, 1995) projects the genes or mutations onto a two-dimensional display which is ordered according to the similarity of the expression profiles. Perhaps the most immediate advantage over alternative methods is the easily understandable visualizability of the result: The SOM display provides an overview of the similarity relationships within the large genome-wide data set. Clusters are visualized with gray shades on the display, and for example mutation clusters can be characterized in terms of which kinds of genes the mutations affect.

This paper has two interrelated goals. We show how to use Self-Organizing Map-based data exploration tools to analyze a genome-wide expression data set and visualize its essential properties. By directly comparing the results with those obtained with hierarchical clustering we hope to convince experimentalists on the utility of the tools in other gene expression analyses as well. More specifically, we show that the same things that have been found using hierarchical clustering can be found easily with the SOM. The second goal is to demonstrate methods for exploring the gene expression data.

1. The Data

We used the data set analyzed earlier by Hughes et al. (2000). The set consists of expression measurements for all yeast (*Saccharomyces cerevisiae*) genes in 300 different treatments, in which one gene was either removed or its function was blocked. The expression of each gene in the mutated yeast is compared to the expression of a wild type yeast. The expression ratio and noise estimate for the ratio is measured for each gene-mutation pair. Thus the amount of data is 300×6297 ratios, and for each ratio there is an estimate for the variance of the measurement noise.

Hughes et al. analyzed first the probability that a given expression measurement deviates from pure background noise, in order to concentrate on expressed genes. They call the process significance-cutting. They constructed a gene-specific error model to calculate the P-value that a given measurement (for a gene-mutation pair) deviates from noise. Genes for which both the P-value and the actual measured expression ratio did not exceed the specified thresholds were left out of the analysis, and likewise for the mutations.

In our work the significance threshold was chosen to be $P < 0.01$, and two different thresholds are used for the expression ratios: When choosing the metric a looser test was used to leave enough data to make inferences: the ratio had to exceed 2. When clustering the data the threshold was stricter, 3.2, in accordance with the study of Hughes et al. The size of the resulting larger set was 179×1482, and the size of the smaller set was 127×578.

Each yeast strain was assigned a MIPS[1] functional class according to the class of the gene that had been mutated. The MIPS classification is hierarchical and allows a gene to belong to several functional classes. The original classification (249 classes) was reduced to 70 by picking the top

Absolute magnitudes \ Zero level	Reliable	Unreliable
Interesting	Euclidean metric	(Euclidean with mean subtracted)
Not interesting	Inner product	Correlation

Figure 1.1. Distance or similarity measures for gene expression profiles. Suitability of the measures depends on whether the zero level of the expressions is reliable or not, and whether the absolute or relative magnitudes are important. Inner product is taken of normalized vectors.

level of homogeneous main classes, and second level of more heterogeneous classes. The functional classes which we considered heterogeneous were metabolism, cell growth, transcription, cellular transport, cellular biogenesis and cellular organization. For example, in the cellular organization class the second level distinguished the genes based on which cellular components they affect, and for the metabolism class the second level subclassifies the genes to the metabolism of different biomolecules like fatty acids or amino acids.

2. Choosing the Metric

Clustering methods group together similar data items, here gene expression profiles, while maximizing dissimilarity between clusters. Hence, the results depend heavily on the metric. The most common distance or similarity measures for real-valued data, such as gene expression profiles, are based on either the Euclidean distance or the correlation coefficient. In this paper we will confine ourselves to variants of these two measures. To our knowledge, there are no comparisons of different metrics with gene expression data in the literature.

The suitability of different measures depends both on the accuracy of the data and on the goal of the analysis. The questions that need be answered include: (i) Is the zero level of the measurements reliable enough, i.e., is there possibly a (large) shift in the baseline? (ii) Which is interesting, the absolute magnitude of expression ratios or the relative values? Some suitable measures for each case are listed in Figure 1.1.

If the baseline is not very reliable then the measures should be invariant to it. In the correlation coefficient the average of the measurements is effectively subtracted before the analysis, and the same can in principle be done for the Euclidean measure as well.

The Euclidean distance measures both the absolute and relative differences in gene expression. If one is only interested in the latter, then the (Euclidean length of the) expression profiles can be normalized. It is easy to show that, for normalized profiles, similarity orderings according to increasing Euclidean distance and decreasing *inner products* $\mathbf{x}^T\mathbf{y}$ are equivalent. Here $\mathbf{x}$ and $\mathbf{y}$ are expression profiles considered as real-valued vectors. The correlation coefficient, on the other hand, is equivalent to the inner product of vectors from which the average has been subtracted.

In summary, correlation is invariant to both the zero level and the scale of gene expressions. Inner product (of normalized vectors) is invariant to scale. Euclidean distances takes into account differences in both.[2]

Note that the absolute value of correlation is sometimes used. This is very sensible within one time series, assuming that there is some common underlying factor that inhibits the activity of one gene and enhances the other. In this paper the advantages are not as salient, however, since the multitude of factors may cause different directions for each gene/treatment combination. Hence we decided to avoid potential interpretation difficulties in the later stages of analysis, and did not use the absolute value.

Plain Euclidean or correlation-based distances weight the expression of each gene at each treatment equally. A standard way of preprocessing variables in the absence of prior information is to scale their variance to unity. Hughes et al. (2000) have additionally measured the noise of each gene-mutation pair individually, and the expression values can be downscaled by the standard deviation of the noise.

In this section we measure empirically which distance measures are most appropriate for gene expression data. As a criterion we use the ability of a simple non-parametric classifier to distinguish between functional classes of the genes.

2.1 Methods

The K nearest neighbor (KNN) classifier stores its teaching data set, and classifies a new expression profile $\mathbf{x}$ according to majority voting between the classes of the K closest samples in the teaching set. That is, if there are more samples from class C than from the other classes, then the prediction will be that $\mathbf{x}$ belongs to class C. We used K=5.

The different metrics were compared by cross-validation: The data set was divided into 27 subsets, and the classification result was computed 27 times for each metric. Each subset was used for evaluating the accuracy

Metric	Preprocessing: None	Normalized noise	Normalized variance	Normalized noise & variance
Euclidean	0.9172	0.9138	0.9152	0.9186
Inner product	0.8898	0.8878	0.8899	0.8899
Correlation	0.8880	0.8866	0.8904	0.8905

Table 1.1. Comparison of different metrics for classifying mutations. The figures are estimated classification errors. For the metrics see Figure 1.1. Normalized noise: Each expression value is scaled by the inverse of the estimated standard deviation of the noise. Normalized variance: The variance of the expression of each gene is normalized to unity. Note that the number of classes is very large compared to the size of the data set; classification error of the naive classifier that predicts according to prior class probabilities is 0.9566.

of the classifier in turn, while the teaching set consisted of the remaining 26 subsets.

Since the genes may belong to several functional classes we had to modify the basic KNN classifier slightly. The new expression profile was considered to be randomly assigned to one of its classes, and the classifier chooses the predicted class randomly from a distribution it has estimated. The percentage of correct classification was then estimated by the probability that the two classes coincide. The classifier estimates the distribution by searching for the K closest samples and calculating the proportion of the classes in them. A gene $\mathbf{x}$ belonging to $N_C(\mathbf{x})$ functional classes contributes for each of the classes with a proportion of $1/N_C(\mathbf{x})$. The classification error is estimated by the average of $(1 - \hat{p}(C(\mathbf{x}))$ over $\mathbf{x}$ and its correct classes within the test set. Here $C(\mathbf{x})$ is the correct class of $\mathbf{x}$, and $\hat{p}(C)$ is the proportion of samples of class C within the neighborhood of K samples.

2.2 Results

The classification errors for the three metrics, with and without the different preprocessing methods are shown in Table 1.1. The best results were obtained by the correlation of expression profiles, which were normalized with respect to the measurement noise.

A paired t-test was used to measure significance of the difference between the best result, and the best results of the other two metrics. Only the difference from the Euclidean metric was significant ($P < 0.01$). The effects of preprocessing were additionally tested for the correlation measure, and they were not significant.

Although the correlation measure was slightly better than the inner product distance, we chose the latter one for the further analysis in the next section to keep the results comparable to the results of Hughes et al. (2000).[3]

3. Self-Organizing Map-Based Exploratory Clustering

3.1 Self-Organizing Maps

The Self-Organizing Map (SOM; Kohonen, 1982; Kohonen, 1995) is a discrete grid of map units. Each map unit represents certain kinds of data, in the present paper different mutations of yeast, the genes of which are expressed in similar ways. 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 an n-dimensional *model vector* $\mathbf{m}_i$ with each map unit i, and by mapping each expression profile $\mathbf{x} \in R^n$ to the map unit having the closest model vector. Here we use the inner product metric: Each expression measurement is first normalized by the estimate of the noise standard deviation, after which the length of the vectors is normalized to unity. The inner product is then used as the similarity measure, and the closest model vector is the vector c fulfilling the condition

$$\mathbf{x}^T \mathbf{m}_c \geq \mathbf{x}^T \mathbf{m}_i \tag{1.1}$$

for all i. The mapping becomes ordered and represents the data after 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 (1.1), and the model vectors are adapted according to

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{ci}(t)\mathbf{x}(t) \ , \tag{1.2}$$

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 (Kohonen, 1995; Kaski et al., 1998).

The SOM can be used both as a nonlinear projection method for dimensionality reduction, and as a clustering method. According to recent evidence (Venna and Kaski, 2001) the similarity diagrams formed by the SOM are more trustworthy than multidimensional scaling-based alternatives, 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.

Compared with most clustering methods the SOM has the advantage that instead of extracting a set of clusters it visualizes the cluster structures. The SOM display is an overview of the similarity relationships and cluster structures of the data set. Such visualizations will be used below to study clusters in gene expression patterns.

Several measurements are missing from the mutation data set, as is common for other real-word data sets as well. For SOM there exists a simple method for computing with missing values: In each step in the iteration, simply neglect the values missing from the current input vector. This approach works well at least if only a small percentage of the values is missing (Samad and Harp, 1992). For the inner product metric we have carried out all computations within the subspace defined by the available components.

3.2 Overview of the Cluster Structure of the Data

The self-organizing map used in this work was the inner-product version of the SOM consisting of 13 times 9 map units arranged in a regular hexagonal grid. The map was computed in two phases. In the first ordering phase of 120,000 iterations the standard deviation of the Gaussian-shaped neighborhood function decreased linearly from 5 to 2 while its height decreased from 0.2 to 0.02. In the second fine-tuning phase of 1,200,000 iterations the standard deviation decreased to 1 and the height to zero.

The SOM overview of the yeast gene mutation data is presented in Figure 1.2. Dense cluster-like areas are indicated in light shades and the sparser areas separating them in data space are indicated in dark shades. The lightness is proportional to the distances between the model vectors of neighboring map units in the data space. Because the density of the model vectors in the data space reflects the density of the data, the distances reflect the density structure of the data. The distance matrix is called the U-matrix (Ultsch, 1993).

The clusters formed by hierarchical clustering of gene expression patterns in mutated yeast strains by Hughes et al. (2000) have been manually drawn on the U-matrix in Figure 1.2. Overall, clusters derived from hierarchical clustering nicely fall into dense areas of the SOM-based dis-

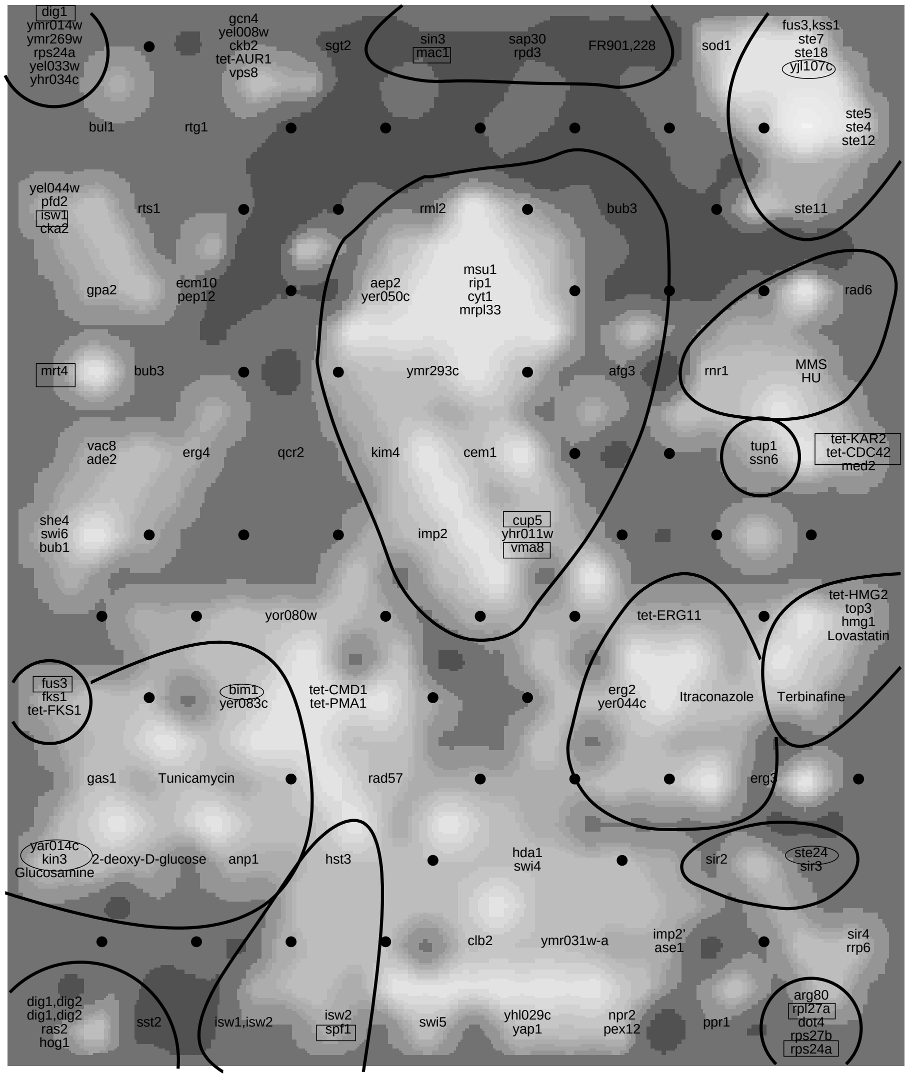

Figure 1.2. Overview of the cluster structure of the data: Light shades denote dense areas in the data space and dark shades sparse areas, respectively. Dots denote the map units without any data vectors. Clusters found by Hughes et al. (2000) with hierarchical clustering have been encircled with the thick lines. Yeast strains that did not belong to the original clusters have been encircled with the thin lines, and boxed strains have moved away from their original clusters.

play, indicating that these two clustering methods in general produce very comparable results.

Closer analysis of the SOM-based U-matrix reveals features which highlight the usefulness of the ability of SOM to visualize the spatial relationships between neighboring clusters. First, mutations in genes related to mitochondrial respiration are clustered in the upper middle region of the SOM display, perfectly agreeing with the cluster of mutations affecting mitochondrial function derived by Hughes et al. (2000). At the lower right corner of this cluster is a subcluster, which includes cup5 and vma8, genes which affect mitochondrial function indirectly by affecting iron homeostasis. Whereas these genes were clustered in a separate cluster from the other mitochondrial mutants by hierarchical clustering (Hughes et al., 2000), SOM places them into a separate group, yet in the close vicinity of the main cluster of mutations affecting mitochondrial respiration.

To the right of the mitochondrial respiration cluster is an area which in SOM analysis appears as a single cluster, but was divided into two distinct albeit spatially related clusters by hierarchical clustering (Hughes et al., 2000). All the mutations in this cluster are related to DNA damage or block of DNA replication (Zhou and Elledge, 1992), which indicates that clustering into a single cluster is functionally justified.

Finally, SOM places sir2, sir3 and sir4 mutants in neighboring neurons within a single cluster in the lower right end of the SOM display. These three genes directly interact to form a protein complex involved in telomeric silencing. Whereas sir2 and sir3 were clustered also by hierarchical clustering, sir4 was not clustered together with them.

In conclusion, in contrast to methods which simply extract a set of clusters, SOM-based display offers more information in the form of hints about the neighbor relations and topography of the data space.

3.3 Interpretation of the Clusters

The overview display (Fig. 1.2) reveals clusters and visualizes their similarity relationships: Clusters close-by on the map are similar. Ultimately the clusters need be analyzed based on biological knowledge and further experimentation, but there exist also SOM-based exploration tools that may help in understanding properties of the clusters.

Perhaps the most simple and intuitive method for characterizing the clusters is to compute the average expression profiles in each interesting map area, and then to visualize their differences in different map areas. It is of particular interest to measure what changes when moving along the map to neighboring map units; note that the map is a non-linear description of the data space and hence the coordinates are non-linear.

We have previously (Kaski et al., 2002) characterized differences between neighboring map units in terms of original data variables, here expression of different genes, using *local factors*. The Self-Organizing Map can be regarded as an approximative *principal surface*, a non-linear generalization of principal components, and measuring changes along the map thus reveals the local principal variation in the data.

In Figure 1.3 we have characterized two areas of the map. About the cluster area in the top right corner of the map we already have some biological knowledge based on the functional similarity of the mutated genes. In contrast, the cluster area in the lower end of the map is a previously unknown cluster structure found with the SOM-based overview in Figure 1.2. We characterize the clusters by analyzing which genes have the most different expression values in two neighboring clusters, and how those genes behave in other mutations. In this manner the similar, or dissimilar, behaviour of the genes in the previously identified cluster structures may clarify the functionality of the mutated genes in the still unknown clusters.

The results of the analysis of the original variables, the expressions of different genes, in the top right corner of the map are in accordance with the previous biological identification of these cluster structures. In Figure 1.3, panel **A** reveals that the components having the largest positive differences, when moving into the upper cluster from the lower cluster, characterize specifically the lower cluster. Indeed, the expression of these genes, many of which are related to metabolism of lipids and carbohydrates, is much lower in this cluster than anywhere else in the data space. Conversely, genes which are downregulated (panel **B**) when moving from the lower to the upper cluster and generally expressed at a lower level in the upper cluster than in other parts of the data space. Many of these downregulated genes are related to mating and pheromone response, which is consistent with the clustering of mating-related mutations in the upper right corner.

Figure 1.2 reveals also a new, possibly interesting cluster in the middle of the lower end of the map. This cluster was not analyzed in Hughes et al. (2000). The mutated genes in this cluster do not form a clear functional group, but many of them are related to transcriptional control, including transcription factors, histone deacetylases and cell-cycle regulated genes. In Figure 1.3 the panels **C** and **D** reveal the components that are up or downregulated when moving from the mid-left to the mid-right cluster at the lower part of the map. Upregulated genes are diffusely localized not only to the right cluster, but also to a large area above it, further indicating that the right cluster does not form a functionally uniform group. Downregulated genes highlight the left cluster, where the genes are expressed on average at a higher level than in the other parts of the data space. Many of the regulated genes encode for heat shock proteins indicating the activation of heat shock reaction in response to mutations clustered in the lower left region.

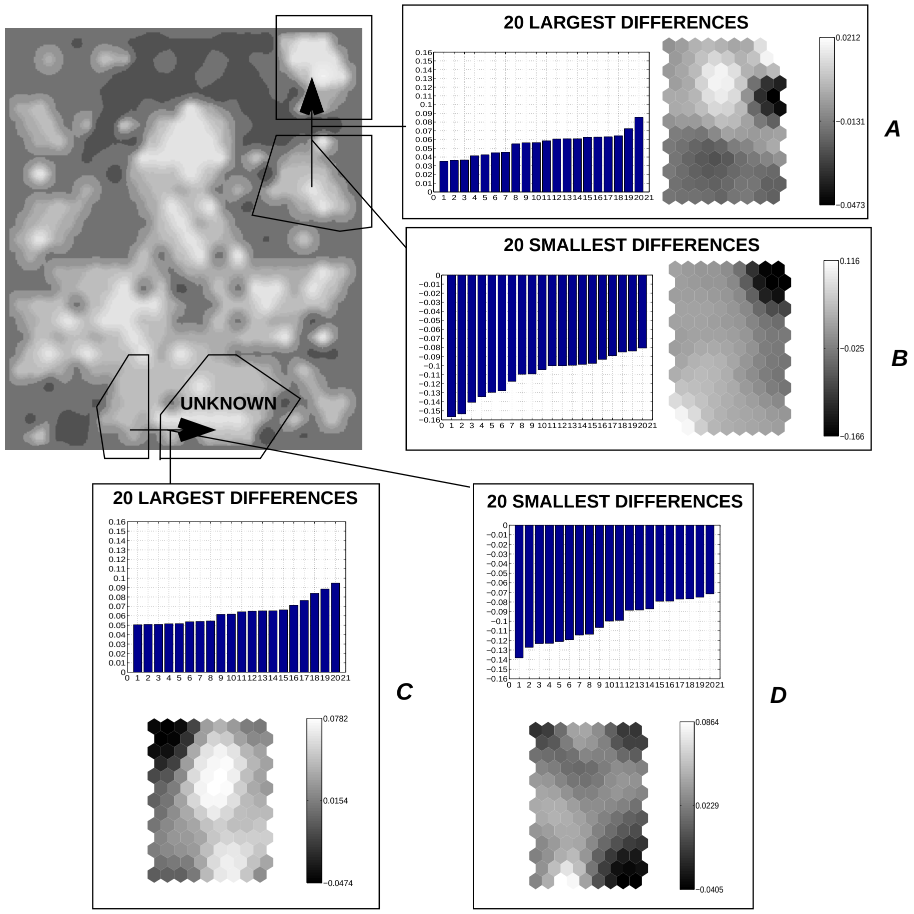

Figure 1.3. Characterization of a sample cluster (panels **A** and **B**) and a previously unknown cluster (panels **C** and **D**). The bar-plots represent changes in average gene expressions when moving according to the arrow from one cluster to another. The 20 largest positive differences and 20 largest negative differences between the average expression vectors of the clusters are shown. The grayshaded map grids in the same panels show the average expression of the same 20 components plotted on the whole map. Light shade denotes high average value and dark shade low average value.

4. Conclusions

We have shown that cluster displays constructed of self-organizing maps correspond well with clusterings formed earlier by hierarchical clustering (Hughes et al., 2000). The main difference between the methods seems to be in the ways they visualize the data: trees vs. 2-dimensional maps.

Map displays enable focusing of the further analysis and provide a framework for intuitive exploration of the data. We demonstrated such exploration by investigating known and still unknown cluster structures.

We compared the performance of different distance measures, Euclidean metric, inner product, and correlation coefficient, in functional classification of the expression profiles. The correlation coefficient was the best, although only the difference from Euclidean metric was significant. This suggests that the zero level in the signals may not be reliable, and that only the relative values of the expression ratios are important. Scaling the expressions by the inverse of a noise estimate improved the results, but not significantly.

The metrics we studied in this paper were global, i.e., the same everywhere in the data space. We will later apply metrics that learn to concentrate on the most important kinds of differences (Kaski et al., 2001b; Sinkkonen and Kaski, 2001). The importance is derived from auxiliary data.

Notes

1. http://www.mips.biochem.mpg.de/proj/yeast
2. Correlation can be viewed as a method of reducing both additive and multiplicative noise, whereas inner product reduces only multiplicative noise. Note, however, that all noise reduction models may also remove essential information.
3. Note that our terminology is different; they call the inner product measure correlation.

References

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffrey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M., and Friend, S. H. (2000). Functional discovery via a compendium of expression profiles. *Cell*, 102:109–126.

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

Kaski, S., Nikkilä, J., and Kohonen, T. (2002). Methods for exploratory cluster analysis. In P. S. Szczepaniak, J. Segovia, J. Kacprzyk, and

L. A. Zadeh, editors, *Intelligent Exploration of the Web*. Springer, Berlin. Accepted for publication.

Kaski, S., Nikkilä, J., Törönen, P., Castrén, E., and Wong, G. (2001a). 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*.

Kaski, S., Sinkkonen, J., and Peltonen, J. (2001b). Bankruptcy analysis with self-organizing maps in learning metrics. *IEEE Transactions on Neural Networks*, 12:936–947.

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).

Samad, T. and Harp, S. A. (1992). Self-organization with partial data. *Network: Computation in Neural Systems*, 3(2):205–212.

Sinkkonen, J. and Kaski, S. (2001). Clustering based on conditional distributions in an auxiliary space. *Neural Computation*. In press.

Törönen, P., Kolehmainen, M., Wong, G., and Castrén, E. (1999). Analysis of gene expression data using self-organizing maps. *FEBS Letters*, 451:142–146.

Ultsch, A. (1993). Knowledge extraction from self-organizing neural networks. In Opitz, O., Lausen, B., and Klar, R., editors, *Information and Classification*, pages 301 – 306. Springer Verlag.

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

Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J. D., Bussey, H., et al. (1999). Functional characterization of the S. Cerevisiae genome by gene deletion and parallel analysis. *Science*, 285:901–906.

Zhou, Z., and Elledge, S.J. (1992). Isolation of *crt* mutants constitutive for transcription of the DNA damage inducible gene *RNR3* in *Saccharomyces Cerevisiae*. *Genetics*, 131:851–866.