Chapter 11

Bioinformatics

Samuel Kaski, Janne Nikkilä, Merja Oja

11.1 Introduction

Bioinformatics refers to the study of biological data using methods from mathematics, statistics, and computer science. In particular, functional genomics experiments produce massive amounts of high-dimensional data that need to be analyzed and understood [7, 8]. Recently developed methods such as oligonucleotide arrays and DNA chips (microarrays) can be used to measure the activity or expression level of each gene in a genome as a function of time in various experimental settings.

The need for data driven methods for generating hypotheses about gene function is already acknowledged in the functional genomics community (see for example [11]). Exploratory methods for information visualization, developed in the Neural Networks Research Centre, would provide precisely the required tools in this first stage of analysis. Methods based on the learning metrics principle (Section 12), on their part, are needed to focus the analyses on the important parts of the very high-dimensional and noisy data, by finding the dependencies between the gene expression data, known functional classes of genes, and other kinds of biological databases including the gene sequences and properties of the corresponding proteins.

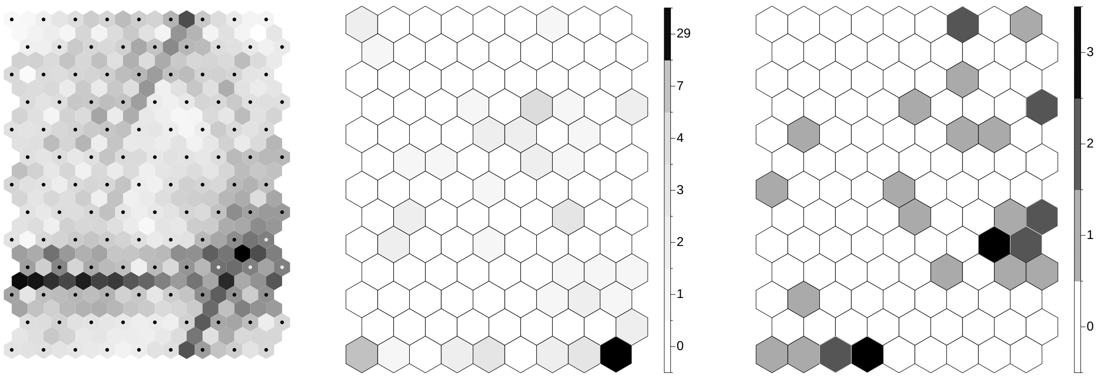
The project is carried out in collaboration with experts of the biological problem, and with the other bioinformatics group of the laboratory led by prof. Mannila and prof. Hollmén. The studies were started in late 2000 in collaboration with a group of the University of Kuopio led by Prof. Eero Castrén.

11.2 SOM-based exploration of gene function

We started by exploratory data analysis of gene expression [3, 4, 9]. The Self-Organizing Map (SOM) is particularly useful in this first stage of data analysis. The SOM constructs a nonlinear projection of the data to a map display which can be used for visualizing of similarity relationships and cluster structures, with methods developed in the Neural Networks Research Centre. Such a combination of non-parametric clustering and visualization distinguishes the SOM from the many clustering methods commonly applied to gene data.

The same SOM display can be used for visualizing the relationships between data sets, such as the gene expression and the functional classes of the genes below.

The genes of the yeast *Saccharomyces Cerevisiae* were first clustered based on their expression in a set of different conditions and treatments such as the diauxic shift and a heat shock (in a public-domain data set). Visualizations of the cluster structure and the relationships of the clusters to the functional classes of the genes were constructed, and the clusters were interpreted in terms of the functional classes of the genes and their activity in the treatments. Figure 11.1 shows, for instance, that most genes related to cytoplasmic degradation form a cluster and hence are expressed similarly in the set of treaments. We demonstrated how to use SOMs in the exploratory task and proposed new hypotheses on the relationships between some functional classes.



a **b** **c**

Figure 11.1: **a** SOM-based visualization of the cluster structure in yeast gene expression data (from [1]). Light shades: clusters; dark shades: sparser areas or gaps in between clusters. The dots denote map units. Note that in this display there is a hexagon in between each pair of map units, whereas on the **b** and **c** figures only the map units themselves are shown. The same SOM is shown in all figures. **b**: Distribution of the functional class 'cytoplasmic degradation' (87 genes) on the SOM. **c**: Distribution of the functional class 'sugar and carbohydrate transporters' (32 genes).

One of the ultimate goals in functional genetics is understanding the regulatory pathways of the genes. This is the key to understanding the dependencies between the genes and to controlling the processes within the cell. The pathways have been studied by "knocking out" one gene at a time by mutations and inspecting the effects on gene expression. If the mutated gene is a vital part of some pathway, then the whole pathway will be blocked. We would thus expect to see all the critical genes in one pathway cluster together because they cause similar effects for the expressions of all the other genes in the cell.

We analyzed the similarities of mutated yeast strains with Self-Organizing Maps [9].

The clusters that had earlier been found by hierarchical clustering [2] were found by the SOM as well, verifying the viability of the method (Fig. 11.2). We were additionally able to propose some new groupings for the mutated yeast strains.

The conclusion from these first studies is that the SOM is a valuable addition to the toolbox of bioinformaticians.

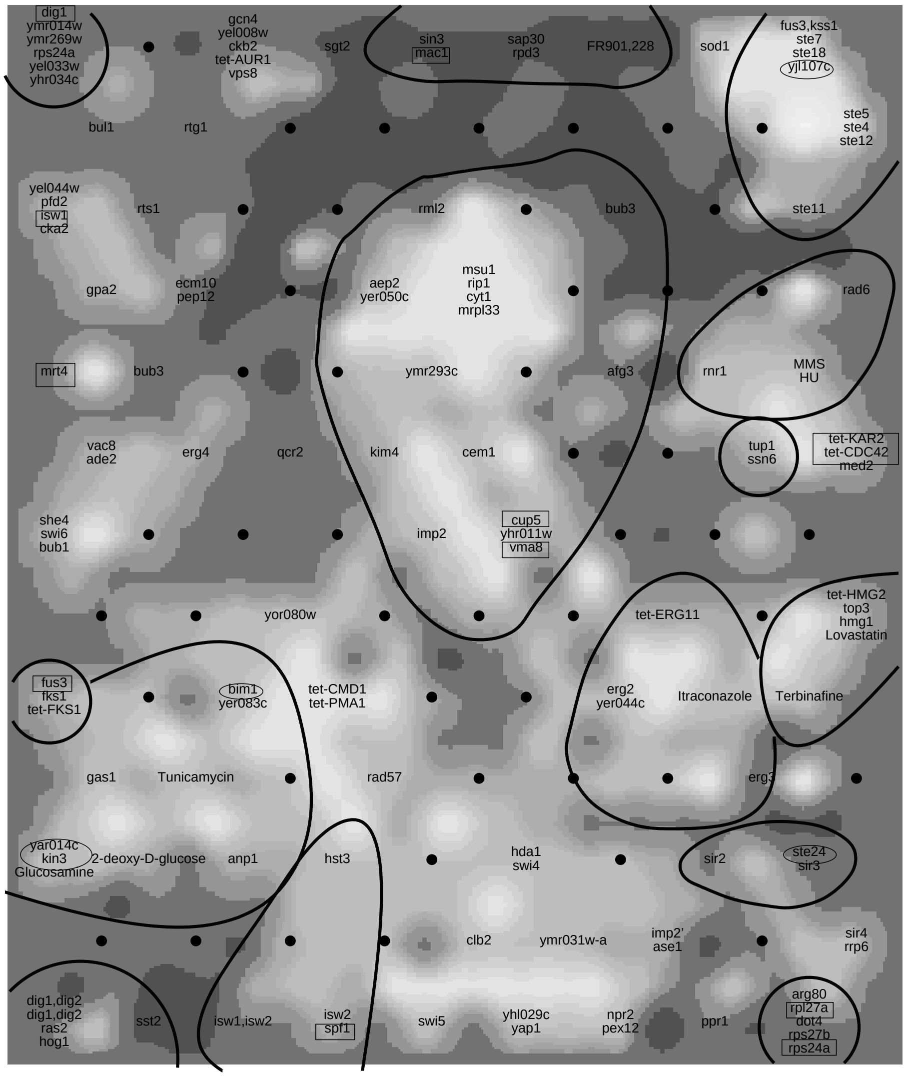


Figure 11.2: The Self-Organizing Map of the mutated yeast strains. Light shades denote dense areas in the expression space and dark sparse areas. The strains located near each other on the map are also nearby in the expression space. The manually drawn lines group the clusters derived earlier by hierarchical clustering [2] (small circles and boxes denote exceptions). As can be seen, approximately the same clusters can also be found based on this SOM-display. The SOM display additionally visualizes the similarities of the *clusters* and the data *between* them.

11.3 Discriminative clustering of genes

Proper selection of the metric of the gene expression space is a well-acknowledged problem since the data is high-dimensional, noisy, and contains uninteresting variation due to the several biological processes going on simultaneously within the cells. Our solution is to derive the metric from other biological data sets such as the functional classification (for details see Section 12).

More generally, there often exist several datasets derived from proteomics, gene expression, and genetic sequences, that could be combined to yield a more accurate picture of cell function.

Discriminative clustering (DC) [10] is a principled way to derive the metric used in clustering from combine auxiliary information. The first preliminary results of this approach applied to gene expression data were presented in [5, 10]. The DC was able to form clusters in the gene expression space that were more homogeneous with respect to the distribution of functional classes than the other methods. More detailed biological analysis and interpretation of these results is in progress.

The learning metrics principle can be applied to Self-Organizing Maps as well [6]. An application to gene expression data is in progress.

References

[1] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences, USA*, 95:14863–14868, 1998.

[2] Timothy R. Hughes, Matthew J. Marton, Allan R. Jones, et al. Functional discovery via a compendium of expression profiles. *Cell*, 102:109–126, 2000.

[3] Samuel Kaski, Janne Nikkilä, Petri Törönen, Eero Castren, and Garry Wong. Analysis and visualization of gene expression data using self-organizing maps. In *Proceedings of NSIP-01, IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing.* 2001.

[4] Samuel Kaski and Janne Sinkkonen. A topography-preserving latent variable model with learning metrics. In Nigel Allinson, Hujun Yin, Lesley Allinson, and Jon Slack, editors, *Advances in Self-Organizing Maps*, pages 224–229. Springer, London, 2001.

[5] Samuel Kaski, Janne Sinkkonen, and Janne Nikkilä. Clustering gene expression data by mutual information with gene function. In Georg Dorffner, Horst Bischof, and Kurt Hornik, editors, *Artificial Neural Networks—ICANN 2001*, pages 81–86. Springer, Berlin, 2001.

[6] Samuel Kaski, Janne Sinkkonen, and Jaakko Peltonen. Bankruptcy analysis with self-organizing maps in learning metrics. *IEEE Transactions on Neural Networks*, 12:936–947, 2001.

[7] Nature Genetics Supplement. The chipping forecast. *Nature Genetics*, 21(1):1–60, 1999.

[8] Nature Insight. Functional genomics. *Nature*, 405:819–846, 2000.

[9] Merja Oja, Janne Nikkilä, Petri Törönen, Garry Wong, Eero Castrén, and Samuel Kaski. Exploratory clustering of gene expression profiles of mutated yeast strains. In

Wei Zhang and Ilya Shmulevich, editors, *Computational And Statistical Approaches To Genomics*. Kluwer, 2002. In press.

[10] Janne Sinkkonen and Samuel Kaski. Clustering based on conditional distributions in an auxiliary space. *Neural Computation*, 14:217–239, 2002.

[11] Ognjenka Goga Vukmirovic and Shirley M. Tilghman. Exploring genome space. *Nature*, 405:820–822, 2000.