Chapter 10

Bioinformatics

Samuel Kaski, Janne Nikkilä, Merja Oja, Leo Lahti, Jarkko Venna, Eerika Savia, Janne Sinkkonen, Jaakko Peltonen

10.1 Introduction

Bioinformatics refers to the study of biomedical data using methods from mathematics, statistics, and computer science. In particular, gene sequencing and functional genomics experiments produce massive amounts of high-dimensional data that are being collected into community resource databases. The data needs to be analyzed, mined, understood, and taken into account in further experiments, which makes data analysis an integral part of biomedical research.

Mining the data to generate hypotheses about gene function and regulation is the next big challenge. Statistical machine learning and mining methods can contribute by learning flexible models of regularities in data. Our research has had two interlinked goals: (1) to develop and apply information visualization and clustering methods for exploring the functional genomics data sets, and (2) to develop methods for focusing the analysis to interesting variation in data. The key assumption is that structures appearing in several data sets are more relevant, and hence dependencies between data sets may reveal interesting findings. Methods of learning metrics and dependency modeling (Section 11) will be used.

The project is carried out in collaboration with experts of the biomedical areas and with the other bioinformatics group of the laboratory that belongs to the From Data to Knowledge research unit.

First steps of a new project in analyzing human endogenous retroviruses were described in Section 8.3.

10.2 Exploratory analysis of gene expression

Exploratory analysis is an irreplaceable first step in bioinformatics research, in particular of gene expression data. Interesting findings need to be made amidst the unknown interactions of thousands of genes, and distinguished from biological and measurement noise.

Visualization plays an important role in the exploratory analysis. The Self-Organizing Map (SOM), developed in the Neural Networks Research Centre, is particularly useful since it constructs a nonlinear projection of the data to a map display which can be used for visualizing of similarity relationships and cluster structures. The SOM has been used successfully to generate hypotheses about regularities in gene expression data [1, 2, 3]. Figure 10.1 shows an example where a SOM-based display revealed the density structure in yeast *Saccharomyces cerevisiae* gene expression measurements [4]. As a result, the biological experiments producing locally the most variation could be discerned. Moreover, we found functionally meaningful subsets of genes, for example ribosomal proteins, and also confirmed that clsses of an existing functional classification are homogeneous in terms of their expression.

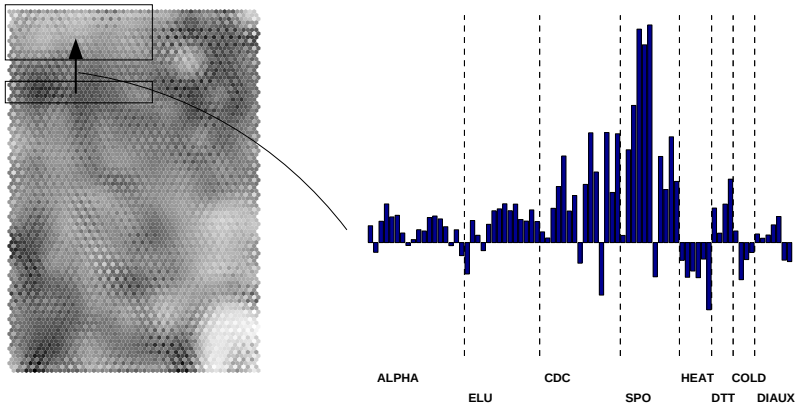

Figure 10.1: Left: SOM representing yeast gene expressions. Right: Difference between data in the cluster in the upper left corner and the area below it. Strong expression in the sporulation (SPO) is the main characteristic distinguishing the cluster.

A key question in visualization is the preservation of the original similarity relationships. In general, it is impossible to preserve all the similarites in the data set, when projecting it to a lower dimensional display. Hence, all visualization methods make a compromise between two goals. On the one hand the visualizations should be *trustworthy*, in the sense that samples that are near each other, i.e. in the same neighborhood, in the visualization can be trusted to actually be similar. On the other hand all the original similarities should become visualized. We argue that, for data exploration, it is more important that the visualizations are trustworthy.

We studied the trustworthiness of SOM and other visualization methods with gene expression data [1, 3]. The SOM was found to be more trustworthy than other methods, except for the smallest neighborhoods (Figure 10.2).

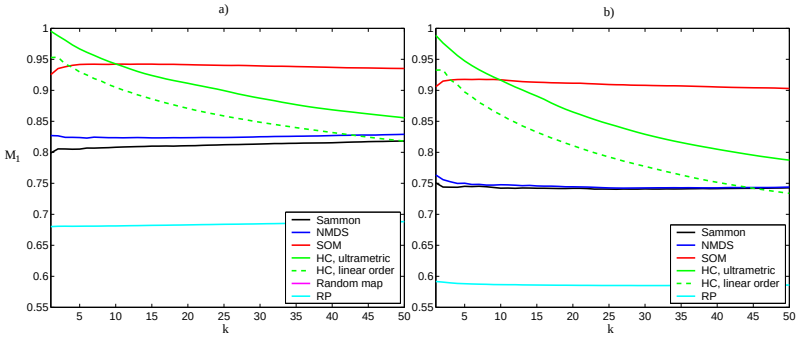

Figure 10.2: Trustworthiness of the visualized similarities (neighborhoods of k nearest samples) for methods that visualize similarity relationships in data. Sammon: Sammon's mapping, NMDS: non-metric multidimensional scaling, SOM: self-organizing map, HC: hierarchical clustering, with the ultrametric distance measure and with the linear distance measure. RP: Random linear projection is the approximate worst possible practical result (the small standard deviation over different projections, approximately 0.01, is not shown). The theoretical worst case, estimated with random neighborhoods, is approximately $M_1 = 0.5$. **a)** Yeast data. **b)** Mouse data.

10.3 Exploratory analysis of dependencies between functional genomics data sets

Exploratory analysis can be enhanced by focusing on relevant variation in the data set, the relevance being determined by another, auxiliary data set. Dependency modeling and learning metrics methods (Section 11) provide a state of the art tool for this, and we have developed and applied them for visualization and clustering in bioinformatics problems.

Visualizations by Maximizing Dependency

A main problem in gene expression analysis is the correct choice of similarity measure, or the metric. It can be learned automatically with the *learning metrics principle* (see Section 11).

We have visualized similarity relationships of gene expression profiles with SOMs in learning metrics. For instance, the metric used in visualizing yeast gene expression was supervised by functional classes of the genes. Visualization of human gene expression was supervised by better-known homologous mouse genes.

Alternatively, expression data can be visualized with a linear projection that generalizes classical methods (Section 11). The results of supervising the projection by functional classes of yeast genes suggest that most functional classes are not strongly differentiated in the expression data, while some information about the overlap of the classes and about their division into subclasses can be found (Fig. 10.3).

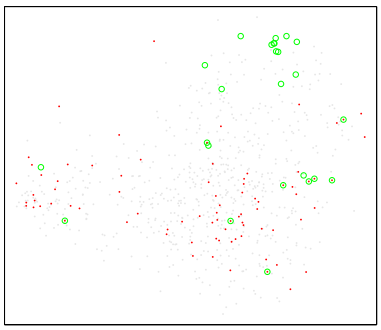

Figure 10.3: Yeast gene expression profiles projected onto two informative components, with protein synthesis (green circle) and mitochondrial organization (red dot) functional classes shown. The protein synthesis class has a subclass at the top.

Clusters by Maximizing Dependency

Dependency maximizing clustering methods (Section 11) are a principled way of finding dependencies between data sets, and presenting them in the form of clusters.

Discriminative clustering (DC) (Section 11) was applied to search for yeast stress genes [7]. We tested the method by replicating the findings of an earlier study. Stress genes should be active in all stress treatments, and additionally potentially regulated by certain regulators (MSN2/4). We clustered yeast gene expression profiles measured in stress treatments, and supervised the clustering by the change of the behavior after the potential regulators were knocked out. This should focus the clusters on behavior regulated by MSN2/4.

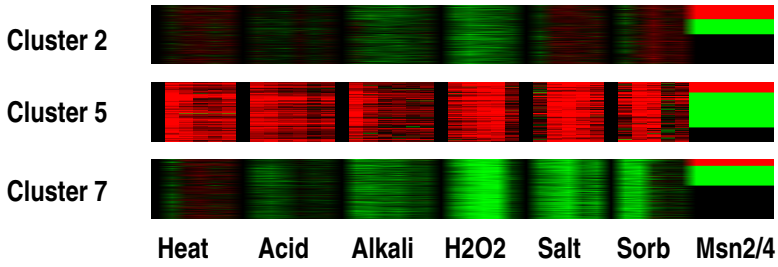

Figure 10.4: Discriminative clusters revealing a group of yeast stress genes (cluster 5) that are putatively regulated by transcription factors MSN2 and MSN4, that is, they are upregulated normally, but downregulated when *Msn2* and *Msn4* are mutated to non-functional. The six leftmost columns are the gene expressions of unmutated yeast and the rightmost column is the discretized gene expression of mutated yeast under stress. Clusters 2 and 7 are examples of an expression cluster without dependency to transcription factors. (Red = upregulation, green= downregulation).

We identified a subset of genes that are upregulated in all stress conditions, but only when regulators MSN2 and MSN4 are functional. Figure 10.4 presents both the gene expressions of normal yeast and the discretized expression of the mutated yeast genes in all DC clusters. Stress genes found in an independent study were strongly enriched in the discovered subset.

Yeast gene regulatory mechanisms were explored with *associative clustering (AC)* (Section 11), by searching for gene groups that are maximally dependent by expression [5] and by transcription factor binding [8]. We found statistically significant dependency, confirmed the results with known regulatory mechanisms, and generated hypotheses for new regulatory interactions.

In a novel application we explored the dependency between expression of human and mouse genes that are putatively orthologous, that is, similar by their sequences [6]. Associative clustering summarizes the data as sets with regularities in their behavior in the two organisms, and outlier sets (Fig. 10.5).

References

[1] Janne Nikkilä, Petri Törönen, Samuel Kaski, Jarkko Venna, Eero Castrén, and Garry Wong. Analysis and visualization of gene expression data using self-organizing maps. *Neural Networks*, 15:953–966, 2002.

[2] 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*, pages 65–78. Kluwer, Boston, MA, 2002.

[3] Samuel Kaski, Janne Nikkilä, Merja Oja, Jarkko Venna, Petri Törönen, and Eero Castrén. Trustworthiness and metrics in visualizing similarity of gene expression. *BMC Bioinformatics*, 4:48, 2003.

[4] 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*, 95:14863–14868, 1998.

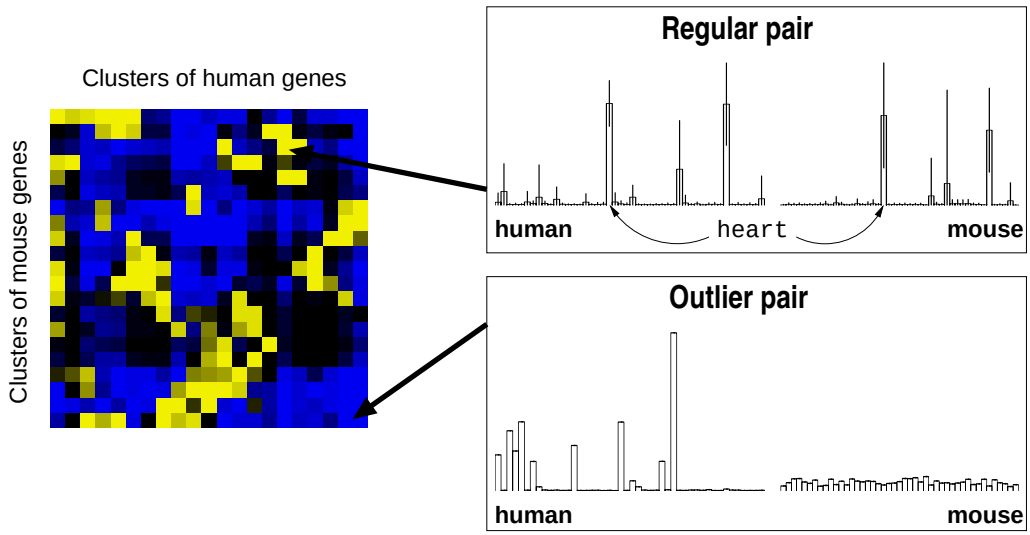

Figure 10.5: Contingency table from AC trained with gene expressions of orthologous genes of human and mouse, that reveals unexpectedly common (yellow) gene pairs and unexpectedly rare gene pairs (blue). An example of both cases is given. **Regular pair:** orthologous genes are expressed strongly in the same tissue, heart, in both organisms. **Outlier pair:** A possibly interesting case where the gene is expressed in human but not at all in mouse, due to functional differences or measurement errors.

[5] Timothy R. Hughes, Matthew J. Marton, Allan R. Jones, Christopher J. Roberts, Roland Stoughton, Christopher D. Armour, Holly A. Bennett, Ernest Coffrey, Hongyue Dai, Yudong D. He, Matthew J. Kidd, Amy M. King, Michael R. Meyer, David Slade, Pek Y. Lum, Sergey B. Stepaniants, Daniel D. Shoemaker, Daniel Gachotte, Kalpana Chakraburtty, Julian Simon, Martin Bard, and Stephen H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102:109–126, 2000.

[6] Andrew I. Su, Michael P. Cooke, Keith A. Ching, Yaron Hakak, John R. Walker, Tim Wiltshire, Anthony P. Orth, Raquel G. Vega, Lisa M. Sapinoso, Aziz Moqrich, Ardem Patapoutian, Garret M. Hampton, Peter G. Schultz, and John B. Hogenesch. Large-scale analysis of the human and mouse transcriptomes. *PNAS*, 99:4465–4470, 2002.

[7] Helen C. Causton, Bing Ren, Sang Seok Koh, Christopher T. Harbison, Alanita Kanin, Ezra G. Jennings, Tong Ihn Lee, Heather L. True, Eric S. Lander, and Richard A. Young. Remodeling of yeast genome expression in response to environmental changes. *Molecular Biology of Cell*, 12:323–337, February 2001.

[8] T.I. Lee, N.J. Rinaldi, F. Robert, D.T. Odom, Z. Bar-Joseph, G.K. Gerber, N.M. Hannett, C.T. Harbison, C.M. Tomphson, I. Simon, J. Zeitlinger, E.G. Jennings, H.L. Murray, D.B. Gordon, B. Ren, J.J. Wyrick, J.-B. Tagne, T.L. Volkert, E.Fraenkel, D.K Gifford, and R.A. Young. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298:799–804, 2002.