Chapter 5

Bioinformatics

Samuel Kaski, Janne Nikkilä, Merja Oja, Leo Lahti, Jarkko Venna, Eerika Savia, Arto Klami

5.1 Introduction

New so-called high-throughput measurement techniques have made possible genome-wide studies of gene function. Gene expression, gene regulation, protein content, protein interaction, and metabolic profiles can be measured and combined with the genetic sequence. The current challenge is to extract meaningful findings from the noisy and incomplete data masses, collected into both community resource and private data banks. 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. Successful genome-wide analyses would allow a completely novel systems-level view into a biological organism.

Combining the different kinds of data produces new systems-level hypotheses about gene function and regulation, and ultimately functioning of biological organisms. We develop probabilistic modeling and statistical data analysis methods to advance this field.

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.

5.2 Yeast systems biology

A major component of systems biology is integration of information from multiple sources. The integration is far from trivial since the data types and scales can vary dramatically. We propose a new framework for systems biology that enables focusing on relevant variation in the data sets, the relevance being determined by other, auxiliary data sets. Dependency modeling and learning metrics methods (Section 6) provide a state of the art tool for this. Together with Docent Christophe Roos from Medicel Ltd. we have developed and applied them for yeast systems biology, especially for exploring yeast stress reaction and its regulation by transcription factor proteins.

Defining yeast stress reaction

Yeast is a key model organism in biological research and process industry. The majority of experiments and utilization of yeast take place by modifying yeast's environmental conditions. A problem is that this always shifts the yeast's biological state from the optimal to more or less stressful state that affects yeast's behaviour. Understanding the yeast stress is thus of crucial importance.

It is believed that yeast as a unicellular organism has a special set of genes that are always activated under any environmental stress. Still, it is very difficult to define an explicit, parametric model for the stress reaction. We have used a novel method to define the stress behavior of the yeast in a data driven way: stress reaction are the effects that are common across different stress treatments.

We have applied a version of gCCA (Section 6) in a novel way for data-driven extraction of the stress effect from the multiple gene expression data sets measured under various stress treatments [5, 6]. The use of gCCA for feature extraction produces a lower dimensional subspace of the original joint data space of all the data sets. That subspace maximally preserves the dependencies between the original data sets. Figure 5.1 presents an overview of the application for gCCA to yeast stress extraction.

In [3] it was demonstrated that even better representation can be achieved by using a non-parametric measure for the dependency in place of the correlation used in gCCA, but with an increased computational cost.

Exploring regulation of yeast gene expression

The biological state of the cell is for a large part defined by which genes are expressed at a certain moment. The regulation of gene expression is thus the key for understanding, for example, the reasons why some cells are transformed to cancer cells. The regulation of expression has been under intensive study during past five years, but it has proved to be extremely difficult to model with detailed statistical models because of the small sample sizes. We search for the hints which genes are dependent on regulating proteins by general purpose methods that work in a data-driven way and utilize the latent group structure of the genes by clustering.

Gene expression is largely regulated by a set proteins called *transcription factors*. They affect gene expression by binding a gene's promoter region, and their type and configuration on the promoter determines in part the activity of the gene. We explored yeast gene regulatory mechanisms with *associative clustering (AC)* (Section 6), by searching for gene groups that are maximally dependent by expression and by transcription factor binding (see Figure 5.2) [4]. We found statistically significant dependency, confirmed the results with known regulatory mechanisms, and generated hypotheses for new regulatory

interactions for many expression data sets, including the expression under environmental stress [1, 7, 6].

Additionally, *discriminative clustering (DC)* (Section 6) was applied to explore the regulation of yeast stress genes [2]. Stress genes should behave similarly in all stress treaments, and potentially be 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 focused the clusters on gene expression regulated by MSN2/4.

We identified a subset of genes that are upregulated in all stress conditions, but only when regulators MSN2 and MSN4 are functional. Stress genes found in an independent study were strongly enriched in the discovered subset.

References

[1] Samuel Kaski, Janne Nikkilä, Janne Sinkkonen, Leo Lahti, Juha Knuuttila, and Christophe Roos. Associative clustering for exploring dependencies between functional genomics data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, Special Issue on Machine Learning for Bioinformatics – Part 2*, 2(3):203–216, 2005.

[2] Samuel Kaski, Janne Nikkilä, Eerika Savia, and Christophe Roos. Discriminative clustering of yeast stress response. In Udo Seiffert, Lakhmi Jain, and Patric Schweizer,

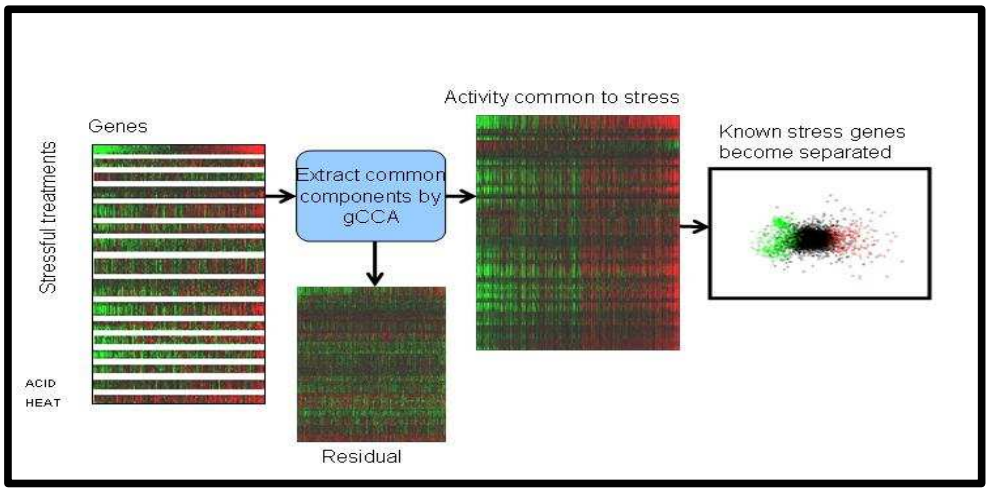

Figure 5.1: Dimensionality reduction by gCCA extracts common properties of data sets. On the left are the original expression data sets from various stress treatments. Red denotes the genes upregulated during stress and, respectively, green the genes that are downregulated. Applying gCCA to the concatenated data sets in a novel way finds a linear subspace that maximally preserves the dependencies between the original variables (in the middle). In the rightmost figure, when the data is projected on the two first components, the known stress genes (red and green dots) become separated from the rest (black dots).

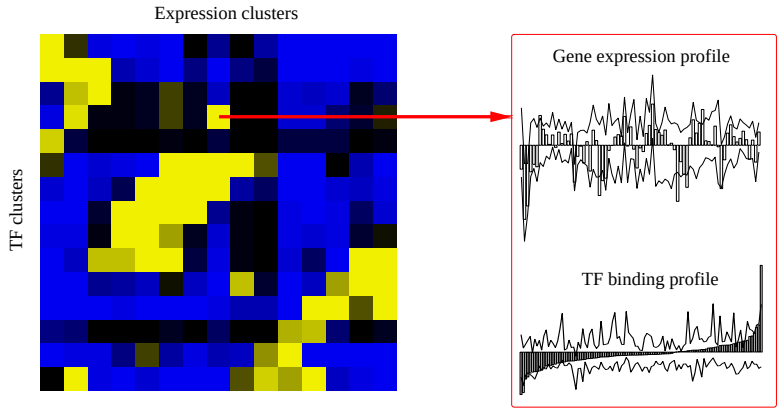

Figure 5.2: Example of a significant association between transcription factor binding and gene expression patterns identified by AC. The yellow cross cluster is associated with cell cycle, and reveals both known and novel dependencies between gene expression and TF binding. The upper profile shows the average expression profile (bars) of the cluster, and the confidence intervals (curves). The lower figure show the average TF-binding profile of the clusters with confidence intervals. In this cluster there were two reliable TF bindings (the rightmost bars in the lower figure), SIP4 and SFL1, of which SIP4 could be verified from the literature, and SFL1 is a new potential regulator for the genes in this cluster.

editors, *Bioinformatics using Computational Intelligence Paradigms*, pages 75–92. Springer,Berlin, 2005.

[3] Arto Klami and Samuel Kaski. Non-parametric dependent components. In *Proceedings of ICASSP'05, IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages V–209–V–212. IEEE, 2005.

[4] 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.

[5] J. Nikkilä, C. Roos, and S. Kaski. Exploring dependencies between yeast stress genes and their regulators. In Zheng Rong Yang, Richard Everson, and Hujun Yin, editors, *Proceedings of International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2004)*, pages 92–98. Springer, 2004.

[6] Janne Nikkilä, Christophe Roos, Eerika Savia, and Samuel Kaski. Explorative modeling of yeast stress response and its regulation with gCCA and associative clustering. *International Journal of Neural Systems*, 15(4):237–246, 2005.

[7] Janne Sinkkonen, Janne Nikkilä, Leo Lahti, and Samuel Kaski. Associative clustering. In Boulicaut, Esposito, Giannotti, and Pedreschi, editors, *Machine Learning: ECML2004 (Proceedings of the ECML'04, 15th European Conference on Machine Learning)*, pages 396–406. Springer, Berlin, 2004.

5.3 Comparative functional genomics

Comparative genomics studies the similarity of the genes in different species. This is of utmost importance when animal models are used to study human diseases and the inferences based on an animal models are translated to human. The basis of this translation are usually the *orthologous genes*.

Determination of orthologous genes, between organisms, relies on similarities in their DNA sequence. However, this does not guarantee their functional similarity. We explored the *functional* dependencies between human and mouse orthologous gene expression by clustering them with associative clustering [1, 2]. In collaboration with the group of Eero Castrén from Neuroscience Center, University of Helsinki, we confirmed the expected functional similarity for many orthologous genes, but we also found some unexpected differences. These differences suggest deviations of genes' functions during evolution and show evidence in favour of using both functional and sequence information when determining homologous genes.

References

[1] Samuel Kaski, Janne Nikkilä, Janne Sinkkonen, Leo Lahti, Juha Knuuttila, and Christophe Roos. Associative clustering for exploring dependencies between functional genomics data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, Special Issue on Machine Learning for Bioinformatics – Part 2*, 2(3):203–216, 2005.

[2] Janne Sinkkonen, Janne Nikkilä, Leo Lahti, and Samuel Kaski. Associative clustering. In Boulicaut, Esposito, Giannotti, and Pedreschi, editors, *Machine Learning: ECML2004 (Proceedings of the ECML'04, 15th European Conference on Machine Learning)*, pages 396–406. Springer, Berlin, 2004.

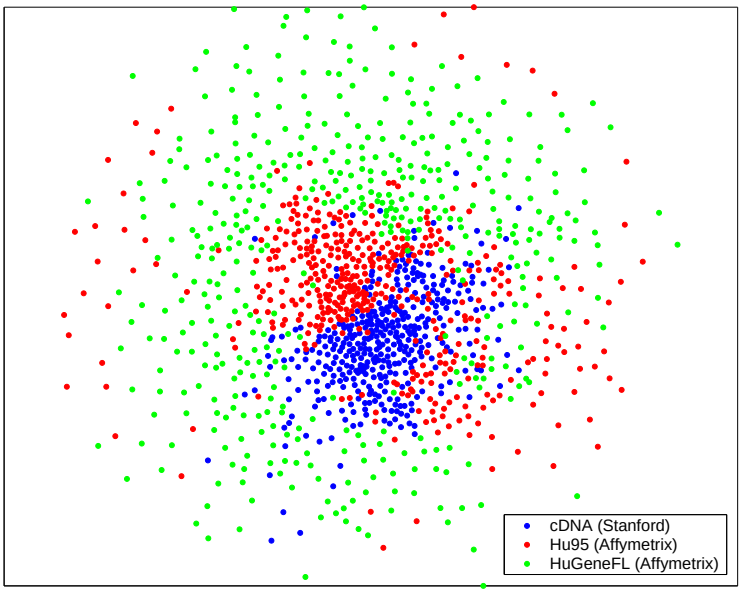

Figure 5.3: Sample visualization of the gene expression atlas by curvilinear component analysis. Each dot denotes one microarray; the colors show the measurement platform.

5.4 Gene expression atlas

A large community-resource or private gene expression databank consists of numerous data sets submitted by several parties. They may have been measured for different purposes, with different treatments and methods in different laboratories. Several such databanks have been established and they continue to grow. A key challenge is how to best use the databanks to support further research. Currently information in these databanks is accessed using queries on the imperfect meta-data, that is, textual annotations and descriptions. In the future more sophisticated search methods, that take the actual data into account, are needed. Our first study [1] aims at creating a visual interface that reveals similarities of data sets.

We compared several different visualization methods in the task of visualizing a large collection of gene expression arrays. A nonlinear dimensionality reduction algorithm, curvilinear component analysis, had the best performance of the methods compared. We also verified that the main sources of variance in the data were also evident from the visualization. In this case the data turned out to be mostly organized based on the type of platform used to perform the experiment. Thus the simple preprocessing method used was not able to make the data commensurable. This is also very evident in Fig. 5.3; the experiments done on each platform create coherent groups and there is only a small amount of mixing of arrays from different platforms. Thus, we were able to produce a visualization organized according to the main sources of variation in the data, and when better preprocessing methods are developed, it will be possible to produce a useful visual interface to a gene expression data bank.

References

[1] Jarkko Venna and Samuel Kaski. Visualized atlas of a gene expression databank. In *Proceedings of Symposium of Knowledge Representation in Bioinformatics*, pages 30–36, Espoo, Finland, 2005.

5.5 Genomics of human endogenous retroviruses

About eight per cent of human DNA consists of remains of specific kinds of transposons[1], called *human endogenous retroviruses (HERV)*. Human retroviruses, such as HIV, in general are viruses capable of copying their genetic code to the DNA of humans, and they become endogenous once they have been copied to the germ-line. Human endogenous retroviruses are remains from ancient infections.

Human endogenous retroviruses, in contrast to some other human transposons, are not capable of moving any longer but it has been suggested that they may have functions in regulating the activity of human genes, and may produce proteins under some conditions [1].

One of the first steps in understanding HERV function is to classify HERVs into families. We have studied the relationships of existing HERV families and tried to detect potentially new HERV families in co-operation with the group of Professor Blomberg, University of Uppsala, [3, 4]. A Median Self-Organizing Map (SOM) [2], a SOM for non-vectorial data, was used to group and visualize a collection of 3661 HERV protein sequences.

The SOM-based analysis was complemented with estimates of the reliability of the results [4]. A novel trustworthiness visualization method was used to estimate which parts of the SOM visualization are reliable and which not. The reliability of extracted interesting HERV groups was verified by a bootstrap procedure suitable for SOM visualization-based analysis. The SOM detected a completely new group of epsilonretroviral sequences and was able to shed light into the relationships of three pre-existing HERV families. The SOM detected a group of ERV9, HERVW, and HUERSP3 sequences which suggested that ERV9 and HERVW sequences may have a common origin.

References

[1] David J. Griffiths. Endogenous retroviruses in the human genome sequence. *Genome Biology*, 2:1017.1–1017.5, 2001.

[2] Teuvo Kohonen and Panu Somervuo. How to make large self-organizing maps for nonvectorial data. *Neural Networks*, 15:945–52, 2002.

[3] Merja Oja, Göran Sperber, Jonas Blomberg, and Samuel Kaski. Grouping and visualizing human endogenous retroviruses by bootstrapping median self-organizing maps. In *CIBCB 2004. IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, 7-8 October, San Diego, USA.*, 2004.

[4] Merja Oja, Göran O. Sperber, Jonas Blomberg, and Samuel Kaski. Self-organizing map-based discovery and visualization of human endogenous retroviral sequence groups. *International Journal of Neural Systems*, 15(3):163–179, 2005.

[1] parts of genome capable of moving or copying themselves in the genome