Chapter 5

Bioinformatics

Samuel Kaski, Janne Nikkilä, Merja Oja, Jaakko Peltonen, Jarkko Venna, Antti Ajanki, Andrey Ermolov, Ilkka Huopaniemi, Arto Klami, Leo Lahti, Jarkko Salojärvi, Abhishek Tripathi

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 methods are used routinely in modern biology and medicine, and now 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. Our main novel contributions stem from the cross-breeding of the methodological basic research, in particular on Modeling of Relevance, and collaboration with top groups in Biology and Medicine. We have had long-standing collaboration with Laboratory of Cytomolecular Genetics (Prof. S. Knuutila) and Neuroscience Center (Prof. E. Castrén), University of Helsinki, University of Uppsala (Prof. J. Blomberg), Turku Centre for Biology (Doc. T. Aittokallio), VTT (Prof. M. Oresic), and smaller-scale collaboration with several other groups. During 2007 we started new projects with EBI, UK (A. Brazma) and Finnish CoE in Plant Signal Research, University of Helsinki (Prof. J. Kangasjärvi) with promising results that will be reported in the next biennial report.

In 2006 we started a new conference series in collaboration with Prof. E. Ukkonen and J. Rousu of University of Helsinki. The conference "Probabilistic Modeling and Machine Learning in Structural and Systems Biology" inspired a special issue in a main journal, and yearly conferences in Evry, France, in 2007, and in Belgium in 2008.

References

[1] Juho Rousu, Samuel Kaski, and Esko Ukkonen, editors. *Probabilistic Modeling and Machine Learning in Structural and Systems Biology. Workshop Proceedings; Tuusula, Finland, June 17-18.* Helsinki, Finland, 2006.

[2] Samuel Kaski, Juho Rousu, and Esko Ukkonen. Probabilistic modeling and machine learning in structural and systems biology. *BMC Bioinformatics*, 8(Suppl 2):S1, 2007.

5.2 Translational medicine on metabolical level

Translational medicine is a research field which attempts to more directly bring basic research findings to clinical practice. One of the necessary steps of this process is to translate inferences made on the molecular level, for example about metabolites, in model organisms into inferences about humans. Such translation is extremely challenging and the existing knowledge, if there is any, is currently largely tacit and only known to experts of the specific disease and model organism.

Metabolomics is the study of the set of all metabolites found in a sample tissue. Metabolite concentrations are affected strongly by diseases and drugs, and hence they complement the genomic, proteomic, and transcriptomic measurements in an excellent way, in studies of the biological state of an organism.

We are in the process of developing new computational methods for translational medicine, for mapping between the observed metabolomics data from model organisms and humans. In project TRANSCENDO we apply the methods to studies of the emergence of Type I diabetes, by computing mappings between non-obese diabetic (NOD) mice and children, and between the effects of a disease in several tissues. The project is collaboration within a consortium involving computational systems biology (Matej Oresic, VTT), semantic modelling (Antti Pesonen, VTT), probabilistic modelling (us), and pharmacology and animal models of metabolic disease (Eriika Savontaus, University of Turku).

Metabolomic development in humans

Metabolic development of children and its differences between the genders is not yet well understood. These dynamic changes may, however, affect strongly the susceptibility to diseases and the responses to drugs.

We are studying a metabolomic data set derived from a collection of blood samples collected during the first years of life from boys and girls. We assume that the metabolic profiles are generated by a set of unobserved metabolic states, and we model those states and the data with a Hidden Markov Model (HMM). HMM fits the assumption of latent states very well and is easy to compute and interpret. Moreover, HMM provides a way for probabilistic re-alignment of the time series, which takes into account the individual variation in the dynamics. Simulations have indicated that HMMs can separate the boys' and girls' metabolic states more efficiently apart than traditional linear method; classification accuracy is 73% for HMM, and under 60% for linear methods. Figure 5.1 presents the model structures for girls and boys.

Disease-related dependencies between multiple tissues

A common setting in medical research is that a disease may be mainly located in a specific organ, for example in lungs, but it indirectly affects multiple tissues. Giving drugs to patients induces an analogous setup: the drugs may affect multiple other tissues in addition to the target tissue (and hence disease). We are developing new methods for discovering the disease-related metabolic dependencies between the multiple tissues, with the goal of revealing potential side effects of the diseases and drugs.

In practice, we have metabolomics data from mice belonging to 4 classes: healthy and untreated, sick and untreated, healthy and treated, sick and treated. A fast and straightforward way of digging out disease-related dependencies is to first find disease-related aspects with partial least-squares classifiers, and then dependencies with canonical correlation analyses and more straightforward correlations between contributing metabolites.

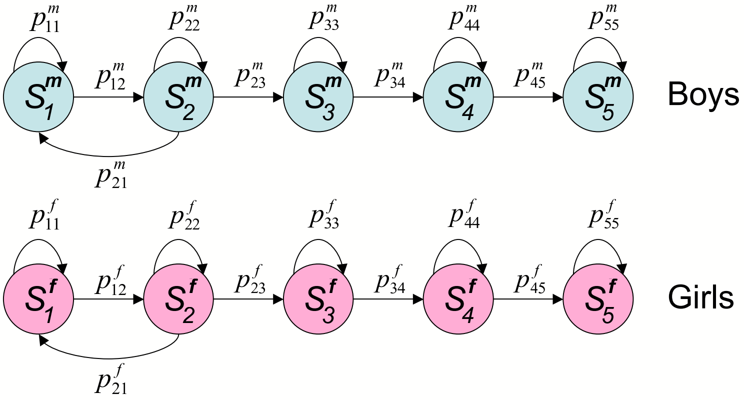

Figure 5.1: HMM models for metabolic states in boys and girls. The nodes represent hidden metabolic states, and the arrows possible transitions. Note that the states form a chain in order to force the models to focus on progressive changes in metabolite concentrations.

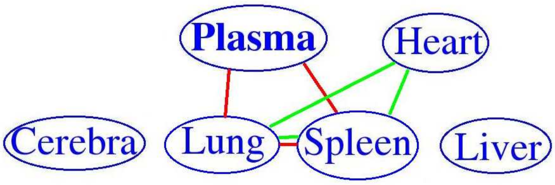

Figure 5.2: Disease-related dependencies between tissues before treatment (red), and after treatment (green). The disease is located in the lungs so the dependencies between lungs and plasma and spleen are logical, but note that after the treatment the dependency with plasma disappears and a dependency to heart emerges. This might be a sign of a side effect of the treatment.

This multivariate approach complements the traditional metabolite-wise linear models. Figure 5.2 shows the dependencies found between tissues before and after drug treatment.

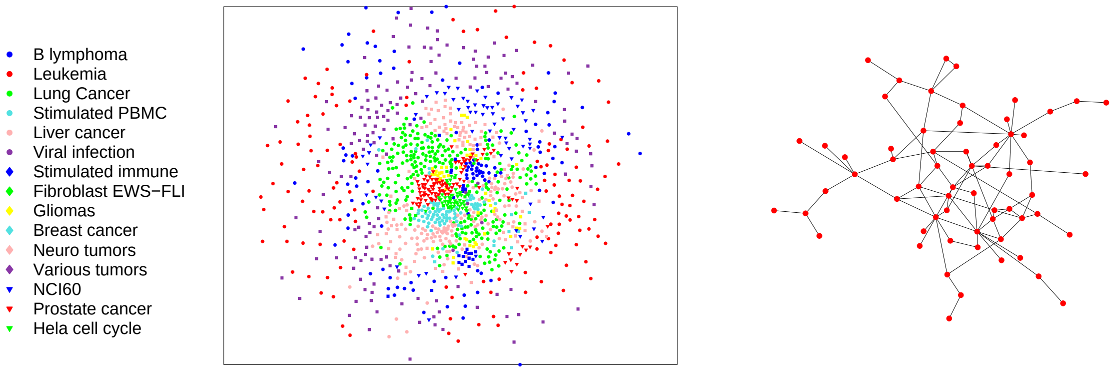

Figure 5.3: *Left:* Sample visualization of a gene expression atlas of cancer samples by curvilinear component analysis. Each dot denotes one microarray; the colors show the cancer class of the sample. *Right:* Part of yeast gene regulatory interaction network visualized by local multidimensional scaling.

5.3 Visualizing gene expression and interaction data

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 study [2] aimed at comparing the different methods applicable as 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. Several new methods have been recently proposed for the estimation of data mani- folds or embeddings, but they have so far not been compared in the task of visualization. In visualizations the dimensionality is constrained, in addition to the data itself, by the presentation medium. It turned out that an older method, curvilinear components analysis, outperforms the new ones in terms of trustworthiness of the projections. Even though the standard preprocessing methods still need to be improved to make measurements of different labs and platforms more commensurable, the good news is that the visualized overview, expression atlas, reveals many of the cancer subsets (Fig. 5.3). Hence, we conclude that dimensionality reduction even from 1339 to 2 can produce a useful interface to gene expression databanks.

Biological high-throughput data sets can also be visualized as graphs that represent the relations between the biological entities. We applied our visualization methods for visualizing gene interaction graphs, and showed that Local Multidimensional Scaling performs very well in this task (Fig. 5.3; [1]).

References

[1] Jarkko Venna and Samuel Kaski. Visualizing Gene Interaction Graphs with Local Multidimensional Scaling In *Proceedings of ESANN'06, 14th European Symposium on Artificial Neural Networks*, pages 557–562, d-side, Evere, Belgium, 2006.

[2] Jarkko Venna and Samuel Kaski. Comparison of visualization methods for an atlas of gene expression data sets *Information Visualization*, 6:139–154, 2007.

5.4 Fusion of gene expression and other biological data sets

While analysis of gene expression data is a corner stone in modern bioinformatics, it is not a sufficient description of cellular state. The cell is an extremely complex system, and gene expression is only a partial view, among all the other omics. Only integration of information from multiple sources can reveal the true potential of the modern high-throughput measurement methods, such as gene expression data.

Integration is not trivial since the data types and scales can vary dramatically. Moreover, what is a proper way of doing the integration depends on the analysis task. Our main novel contribution has been to develop and apply new methods for searching for relevant features by combining data sources (described in Section Modeling of Relevance). We have additionally developed more specific methods for taking into account the known regulatory and context variables in modeling gene expression.

Relevant features through data fusion

We consider a data fusion problem of combining two or more data sources where each source consists of vector-valued measurements from the same object or entities but on different variables. The task is to include only those aspects which are *mutually* informative of each other. This task of including only shared aspects of data sources is motivated through two interrelated lines of thought. The first is noise reduction. If the data sources are measurements of the same entity corrupted by independent noise, discarding source-specific aspects will discard the noise and leave the shared properties that describe the shared entity. The second motivation is to analyze what is interesting in the data. One example is the study of activation profiles of yeast genes in several stressful treatments in the task of defining yeast stress response. In this example what is in common in the sources is what we are really interested in. The “noise” may be very structured; its definition is simply that it is source-specific.

A recent application is search for asbestos-related effects in gene expression by combining several cell lines [3].

We recently showed that there is a simple and computationally fast way of doing data fusion such that shared, relevant features are retained and source-specific noise is discarded [2]. The method is based on the classical canonical correlation analysis; it is surprising that there are still new practically important uses for so old methods! The method has been applied to several gene expression studies: classification of cell cycle regulated genes in yeast, identification of differentially expressed genes in leukemia, and defining stress response in yeast. The software package is available at `http://www.cis.hut.fi/projects/mi/software/drCCA/`.

Modeling context specific gene expression regulation

The biological state of the cell is to a large part defined by which genes are expressed at a certain moment or under certain environmental conditions. Regulation of gene expression is thus the key to understanding, for example, the reasons why some cells become transformed to cancer cells. Regulation of expression has been under intensive study during the past years, but analysis with statistical models has proved to be extremely difficult because the sample sizes are always small due to high measurement costs. The effective sample sizes become even smaller when analyzing context specific regulation, where the data becomes divided according to the context or experimental setup. We have introduced ways of context-specific modeling with one of the most often used model families, the Bays networks.

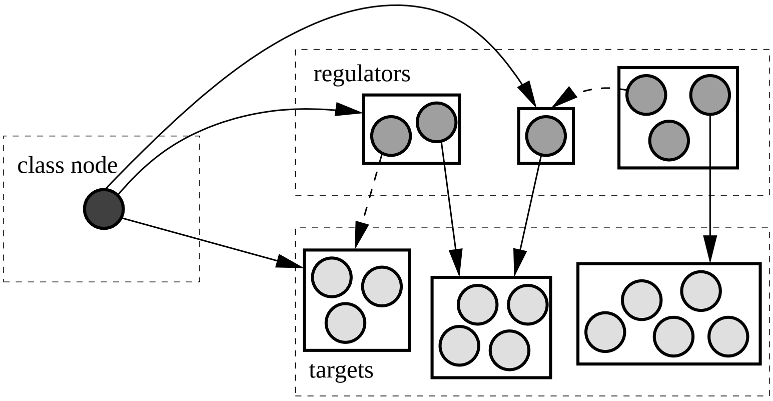

Figure 5.4: The structure of condition-dependent Bayesian network. Similarly behaving genes are grouped into modules. The edges depict the regulatory interactions. The dashed edges indicate interactions that are active in one of the conditions only. Genes linked from the class node may behave differently in different conditions.

The gene regulatory relationships form a complex network in which genes can be regulated by multiple regulators or through long chains of regulatory interactions. The regulatory network adapts to the conditions outside the cell by activating or stopping regulatory interactions in response to changes in the environment. We have studied regulatory networks in yeast with new *condition-dependent Bayesian network* [1]. The data has been divided into several conditions or contexts indicated by a context or class variable that is treated as a covariate. The model has the novel capability of identifying interactions that are active only in subset of conditions. The output of the method is a graphical representation of the network where the possible condition-dependent interactions are highlighted. Figure 5.4 depicts a conceptual example of a condition-dependent network.

We analyzed the regulation in yeast cultures which had been subjected to normal and stressful growth conditions. The method identified 25 regulators which are active only in stressful conditions. The majority of them (20 out of 25) have been annotated stress-related in the literature. The rest are new potential stress regulators.

References

[1] Antti Ajanki, Janne Nikkilä, and Samuel Kaski. Discovering condition-dependent Bayesian networks for gene regulation. In *Proceedings of Fifth IEEE International Workshop on Genomic Signal Processing and Statistics*, 2007.

[2] Abhishek Tripathi, Arto Klami, and Samuel Kaski. Simple integrative preprocessing preserves what is shared in data sources. *BMC Bioinformatics*, 9:111, 2008.

[3] Penny Nymark, Pamela M Lindholm, Mikko V Korpela, Leo Lahti, Salla Ruosaari, Samuel Kaski, Jaakko Hollmen, Sisko Anttila, Vuokko L Kinnula, and Sakari Knuutila. Gene expression profiles in asbestos-exposed epithelial and mesothelial lung cell lines. *BMC Genomics*, 8:62, 2007.

5.5 Human endogenous retroviruses

The human genome includes surviving traces of ancient infections by retroviruses that have become fixed to human DNA. These surviving traces are called *human endogenous retroviruses* (HERVs). HERVs are interesting because they can express viral genes in human tissues, and because their presence in the genome may affect the functioning of nearby human genes. If ancient highly mutated elements are included, HERV sequences form 8% of the human genome [1].

In earlier research we had used Self-Organizing Maps to analyze the classification of HERVs into families [2]. In recent research we have moved to estimating the relative activities (expression levels) of the HERVs across several human tissues. We analyze activity for individual HERV sequences (rather than groups of sequences); this is vital for analyzing their individual control mechanisms and their possible roles in diseased and normal cell functions.

To find evidence of HERV activity, we use probabilistic modeling methods for expressed sequence tags (ESTs) gathered from public databases. We introduced a generative mixture model for EST sequences where each component of the mixture was associated with a particular HERV (see the top subfigure of Fig. 5.5). In our experiments we compared this rigorous model with a fast heuristic method; it turned out that the fast method performed reasonably accurately on simulated data, which made it possible to analyze very large HERV collections.

We first used the models to analyze overall activities across different tissues and conditions. In addition to comparisons on simulated data, we performed several experiments on real HERV data; the probabilistic method for a smaller and the fast method for a larger set having 2450 HERVs [3]. Lastly the probabilistic model was used to estimate tissue-specific expression of HERVs from the HML2 family [4].

Overall, 7% of the HERVs were estimated to be active; the majority of the HERV activities were previously unknown. HERVs with the retroviral *env* gene were found to be more often active than HERVs without *env*. We were also able to analyze which parts of the HERV sequences the EST data match to; see [4] and its supplementary material for figures. For the HERV family HML2, activity profiles of HERVs over tissues are shown in the bottom subfigure of Fig. 5.5; some of the HML2 HERVs display tissue-specific expression (e.g. activity in male reproductive tissues or in the brain).

References

[1] Eric S. Lander *et al.* Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[2] 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.

[3] Merja Oja, Jaakko Peltonen, Jonas Blomberg, and Samuel Kaski. Methods for estimating human endogenous retrovirus activities from EST databases. *BMC Bioinformatics*, 8(Suppl 2):S11, 2007.

[4] Merja Oja. In silico expression profiles of human endogeneous retroviruses. In *Proceedings of the Workshop on Pattern Recognition in Bioinformatics (PRIB 2007)*, 2007.

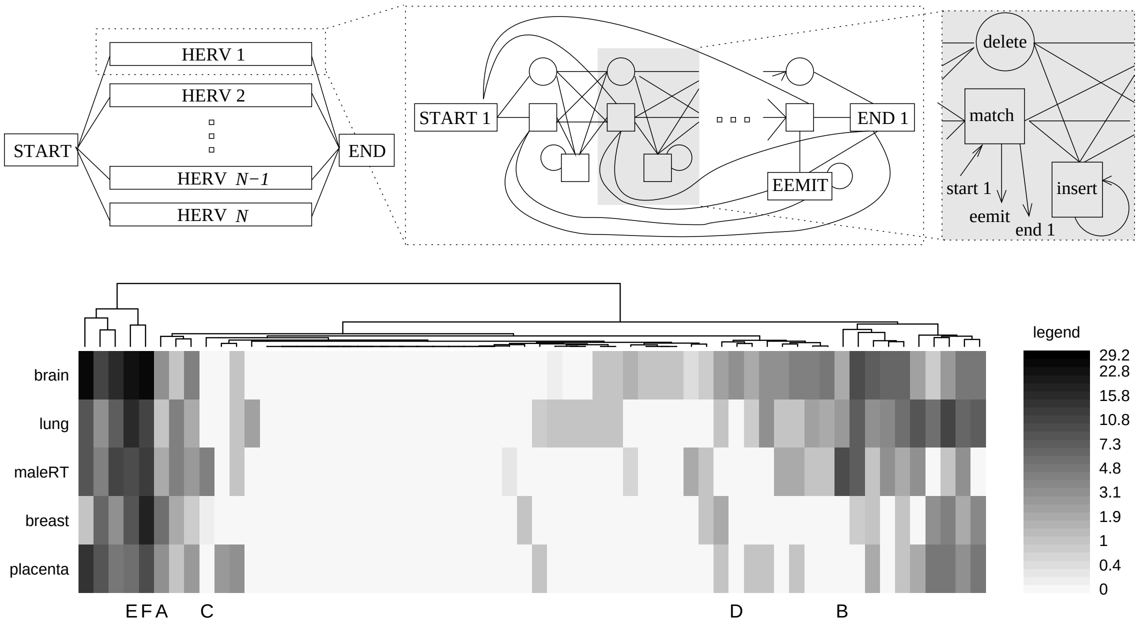

Figure 5.5: Top: the new probabilistic mixture model introduced for estimating activity of human endogeneous retroviruses (HERVs) from expressed sequence tags (ESTs). Bottom: activities of HERVs from the HML2 family in different tissues. Each column depicts the expression profile of an individual HERV sequence; the columns have been ordered by hierarchical clustering based on the profiles. Numbers next to the legend are probabilistic EST counts. Letters A-F at the bottom identify individual HERVs that have been analyzed in [4].