Chapter 5

Bioinformatics

Samuel Kaski, Jarkko Salojärvi, Gayle Leen, Arto Klami, Jaakko Peltonen, José Caldas, Andrey Ermolov, Ali Faisal, Ilkka Huopaniemi, Leo Lahti, Juuso Parkkinen, 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.

We develop new modeling and data analysis principles needed for discovering the relevant signals and patterns from among the several measurement sources, and numerous earlier experiments collected into measurement databases. Our multi-source machine learning methods have proven to be very useful here, and new methods for retrieving and analyzing relevant experiments have promise for breakthroughs in making the data-driven sciences, biology and medicine, more cumulative. We have long-standing collaboration with European Bioinformatics Institute EBI (prof. A. Brazma), Laboratory of Cytomolecular Genetics (Prof. S. Knuutila), Department of Biological and Environmental Sciences, University of Helsinki (Prof. J. Kangasjärvi), VTT (Prof. M. Orešič), Finnish Institute for Molecular Medicine FIMM (prof. O. Kallioniemi), and smaller-scale or preliminary collaboration with several other groups.

5.2 Translational medicine on metabolic 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.

Our mission is to develop the computational methods needed for making molecular level translational medicine possible. We have developed new computational methods for mapping between the observed metabolomics data from model organisms and humans. In project TRANSCENDO we applied 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 (M. Orešič, 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 in children developing into Type 1 diabetes is not well understood, and we develop computational methods in order to shed more light into it. We work on a unique data set of our collaborators [11], of metabolomic profiles derived from time series of blood samples of a large cohort of children.

We developed computational methods for studying dynamic differences between time-series measurements of two populations. In the first phase, differences between healthy boys and girls were studied [10], and at the moment we are moving forward to actual translational medicine.

The models operate under the assumption that the metabolic profiles are generated by a set of unobserved metabolic states, which can as the first approximation be modelled with Hidden Markov Models (HMM). HMM fits the assumption of latent states very well, is easy to compute and interpret, and can be extended into more flexible and expressive models. Moreover, HMM provides a way for probabilistic re-alignment of the time series, which takes into account the individual variation in the dynamics. The HMMs were able to separate the boys' and girls' metabolic states (Figure 5.1) more efficiently apart than traditional linear methods.

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).

A typical analysis setup in any one-tissue (typically blood) biological experiments, looking for potential biomarkers for disease, is the diseased-healthy differential analysis. Biological experiments often contain additional covariates, such as drug treatment groups, gender or time-series, resulting in a multi-way experimental setup. Finding effects of multiple covariates from the data is a traditional statistical problem dealt with by Analysis

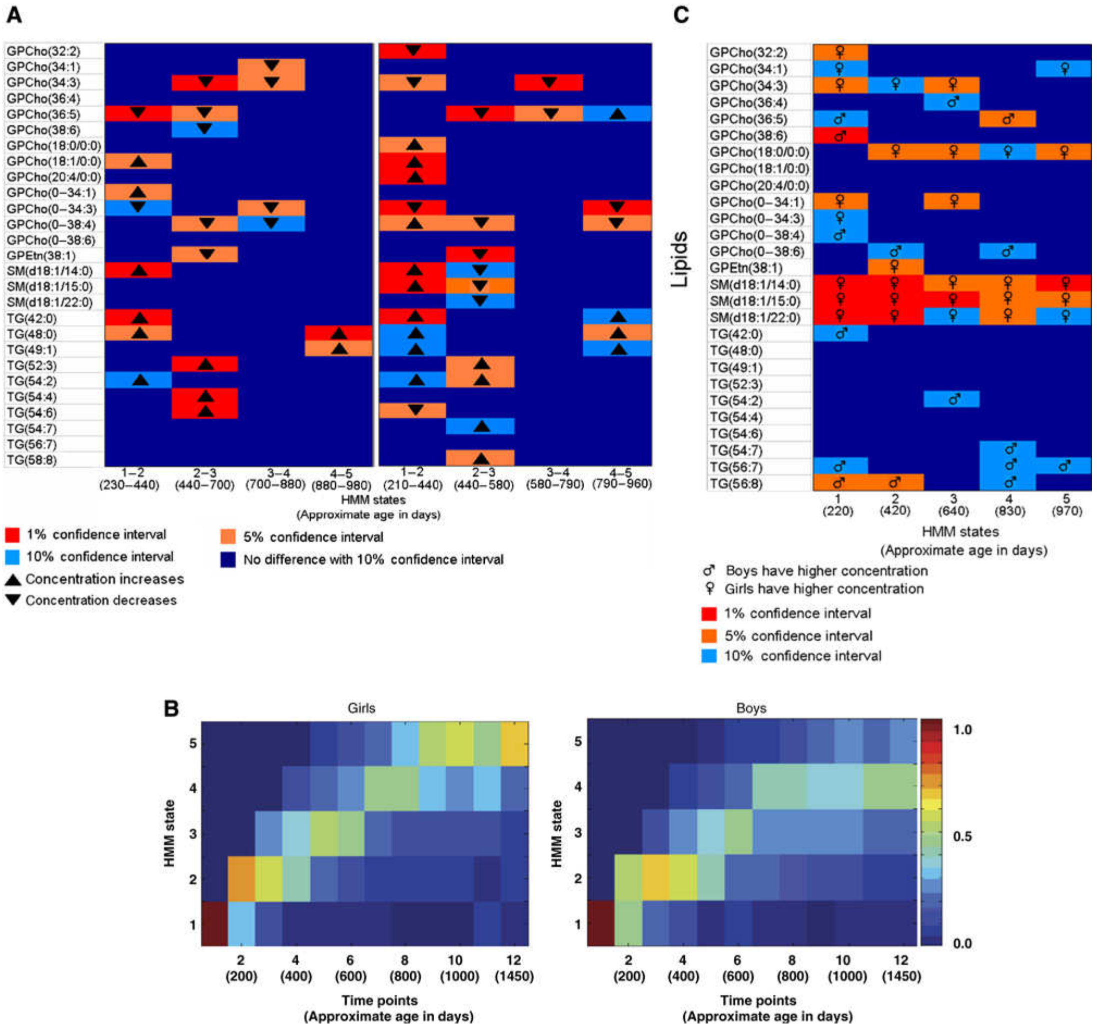

Figure 5.1: Boys and girls have different development of metabolic states.

of Variance (ANOVA) or in general by linear models. However, the main challenge in modern molecular biology is that the number of samples (such as mice or human patients) is small due to economical and ethical reasons, and the number of variables (genes or metabolites) is huge. Due to this, the traditional multivariate methods cannot be used, and few modern methods exist for this task. To address this broad and common set of problems, we developed a Bayesian model family for multi-way analysis of small sample-size, high-dimensional datasets [5].

The problem becomes even more interesting when the different data sources (here tissues) form different variable-spaces. Then standard approaches are not applicable even in principle. Our model can be extended even to this case, by considering the different sources as different "views" in the sense of multi-view machine learning. The extended model is able to find the multi-way covariate-effects and to partion them into shared and source-specific effects. The method is applicable to any small sample-size, multi-source experiments, currently very popular in biological research. We call the general problem (Figure 5.2) Multi-Way, Multi-View Learning [6].

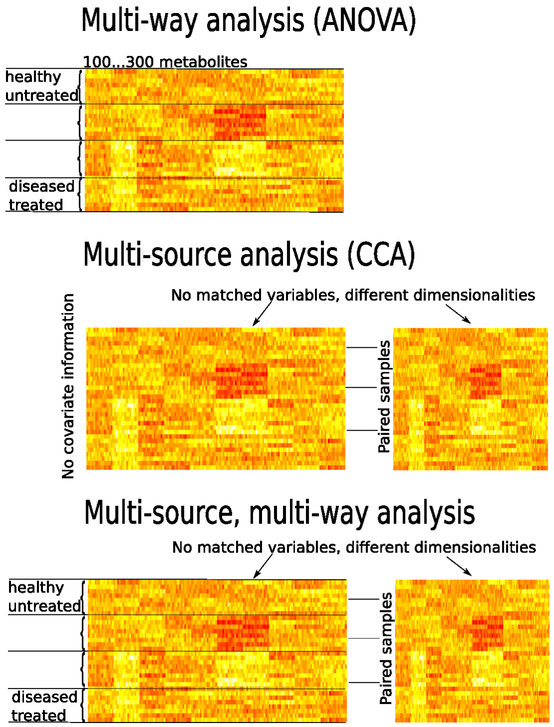
Multi-way analysis (ANOVA)
100...300 metabolites
healthy
untreated
diseased
treated
Multi-source analysis (CCA)
No matched variables, different dimensionalities
No covariate information
Paired samples
Multi-source, multi-way analysis
No matched variables, different dimensionalities
healthy
untreated
diseased
treated
Paired samples

Figure 5.2: (a) Multi-way analysis studies datasets with two or more covariates for each sample. The task is to find the effects of the covariates in the data (b) Multi-source analysis studies dependencies between two or more datasets with paired samples without covariate information. (c) Multi-way, multi-source analysis combines both tasks. The task is to find shared and source-specific covariate effects. In the data matrices, rows represent samples, and columns represent variables.

5.3 Retrieval and visualization of relevant experiments

Repositories of genome-wide expression studies such as ArrayExpress [12] are becoming mature both in size and data annotation quality. This brings in the research question of how to systematically relate studies contained within those repositories. By allowing data to be re-used on a mass scale, researchers will be able to access a meaningful biological context to aid in the planning and analysis of new studies. This will in turn increase the statistical power of novel studies and put biological results in the context of previous studies. Most repositories contain basic text search functionalities that allow retrieving studies whose textual descriptions contain certain keywords (e.g. 'cancer'). This paradigm has several shortcomings. First, the textual description of an experiment or its results is not as information-rich as the actual data itself. Secondly, it does not provide any solution with respect to analyzing the retrieved study and rigorously comparing it to a novel study. In collaboration with the Brazma group at The European Bioinformatics Institute, which has created and maintains the ArrayExpress database, we are working towards developing machine learning methods that relate studies through their actual expression data, along with visualization tools that allow exploring and interpreting the results.

Content-based information retrieval for differential expression. Gene expression studies often involve differential expression analyses that allow assessing genes or pathways for consistent changes in expression in a phenotype of interest (e.g. cancerous tissue

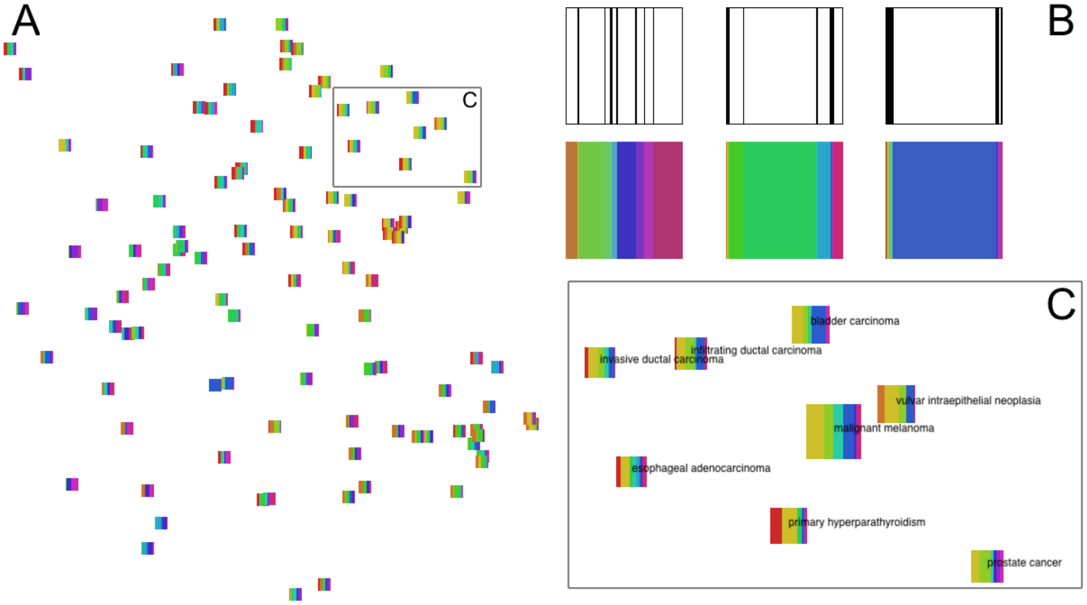

Figure 5.3: Figure taken from Caldas *et al.* [2]. (A) The experiment collection visualized as glyphs on a plane. (B) The method represents each experiment as a collection of so-called biological topics or components. Slice color and width represent the importance of each component in an experiment. (C) Enlarged region from (A) where glyphs have additionally been scaled according to their relevance to a query with a malignant melanoma experiment shown in the center. The surrounding experiments are either from cancer or from primary hyperparathyroidism, which is known to be associated with a higher incidence of cancer.

versus healthy tissue). Recently, it has been shown that differential expression analysis at the level of pathways or gene sets leads to improved and more robust results than differential expression at the gene level [15]. As the first prototype of our biological content-based information retrieval paradigm, we have developed a method that allows relating studies in a repository through shared patterns of gene set differential expression, using a combination of state-of-the-art nonparametric statistics and machine learning approaches [2]. We have also developed novel visualization tools that allow exploring both the data and the retrieval results. Our results show that, given a so-called query study, the method provides a set of other studies where most target the same biological question (e.g. cancer studies) (Fig.5.3). It is also able to find highly non-trivial relations between significantly different pathological entities which were confirmed in the literature. Finally, the retrieval results are interpretable, in the sense that the method provides the patterns of differential expression that are responsible for the inferred relevances (Fig.5.4). Although there has been previous work related to large-scale analysis of differential expression, ours is the first to highlight the potential of performing content-based, interpretable information retrieval in a rigorous and principled manner.

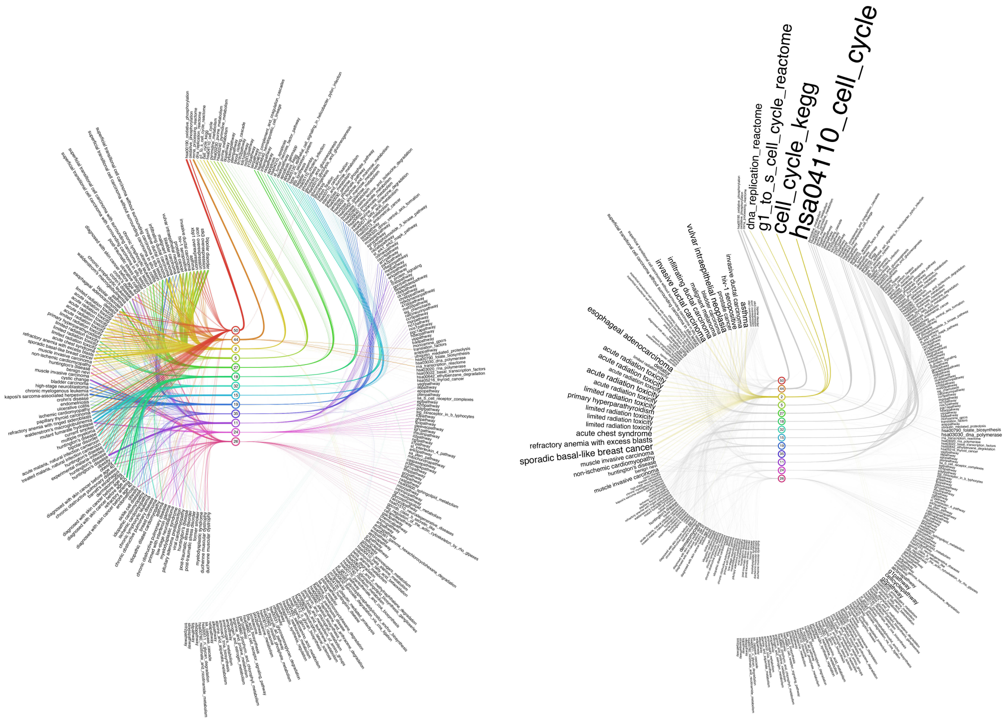

Figure 5.4: Figure taken from Caldas *et al.* [2]. A circular visualization that associates biological studies (labels on the left) to recurrent patterns (circles in the middle), and recurrent patterns to known biological pathways (labels on the right). The left figure shows the general visualization; the degree of association between studies, patterns, and pathways is encoded through edge opacity, where each color is specific to one pattern. In the right figure, the labels of both biological experiments and pathways are scaled according to the degree of association with recurrent pattern number 2. The figure shows an association between several cancer studies (e.g. sporadic basal-like breast cancer) and a collection of cell cycle-related biological pathways.

5.4 Fusion of heterogeneous biomedical data

A living cell is an extremely complex system, and hence integration of information from multiple sources is needed for revealing the true potential of the modern high-throughput measurement methods, such as gene expression or micro-RNA data, combined with relational information of the genes, environmental factors, and disease.

Much of the blooming data integration literature focuses either on well-targeted combinations of sources, such as using sequence-based regulators for explaining gene expression, or on well-focused prediction tasks such as predicting molecular interactions from several data sources. We have focused on knowledge discovery-types of problems where the goal is to discover what is relevant in massive data sets by searching for connections between data sources. This will become more concrete below. Additionally, we have worked on more specific but application-wise interesting problems, such as detection and analysis of deficiencies in the measurements [8].

Dependency modeling. We consider the 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

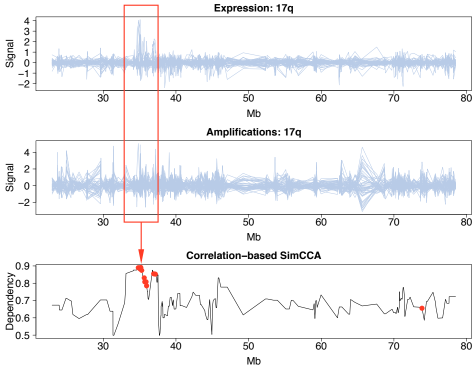

Figure 5.5: Gene expression, copy number signal, and the dependency between the two data sources along chromosome arm 17q in gastric cancer patients. The model detects known cancer associated genes (red dots) with high sensitivity. The Figure is from [9].

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. We have investigated, for example, functional effects of DNA mutations by observing dependencies between gene expression and copy number levels across cancer genomes [9]. In these works, the shared variation of the data sources is of primary interest (Figure 5.5). The data set specific effects, often regarded as "noise", may have specific structure; its definition is simply that it is source-specific. The data set specific effects can also be of interest in certain applications. In [14], for example, the decomposition of gene and protein expression levels into shared and data set specific effects was used to distinguish between pre- and post-translational regulation.

We have developed novel ways to bring in prior information to dependency modeling tasks [9]. This helps to reduce overfitting and focus the modeling on the interesting parts of the data, which is critical in many biomedical applications with small sample sizes. We have also released an open-source software package for general fusion of biological data sets, using generalized canonical correlation analysis for both combining the data sets and finding a lower-dimensional representation for them [16].

Dependency models are potentially applicable for modeling other regulatory mechanisms such as transcription factors [7], or micro-RNAs that form a recently discovered and central class of cellular regulators. While causality and confounding factors are often unknown in these studies, detection of statistical dependencies provides a useful proxy for such effects. Future extensions of the dependency models will provide tools to detect multi-level relations between various regulatory mechanisms and gene activity.

Matching of entities. Most data fusion approaches assume co-occurring data sources, in the sense that all sources are different representations of the same entities. For example, in joint analysis of several mRNA experiments we assume the same set of genes has been measured in each experiment. Due to heterogenity of the biological data sources this assumption, however, does not always hold. If the experiments have been measured with different platforms the mapping between the probes might not be perfectly known, or in translational medicine the sources (tissues or species) might even have different entities altogether.

We have introduced a novel method that learns the matching of the entities in a data-driven way, using the actual measurements to find the co-occurrences [17]. The method is based on a very intuitive principle: The matching that gives maximal statistical dependency between the sources is most likely to be correct. The approach can either be used to complement a partial match found based on auxiliary data sources (such as sequence similarity when matching probes of two microarrays), or even to learn the matching from scratch.

Bayesian biclustering. Biclustering is the computational task of simultaneously clustering objects and inferring which features of the objects contribute to the grouping. It is a highly relevant area in gene expression bioinformatics, when one aims at finding restricted biological conditions where certain genes exhibit similar behavior, or alternatively at finding groups of genes with respect to which a set of biological conditions is similar. It is also deeply connected to the fields of content-based information retrieval and data fusion.

We have first adapted an existing promising model to the Bayesian framework, allowing the model to handle noise and endowing it with a rigorous inference engine [1]. More recently, we have developed a hierarchical nonparametric biclustering method [4]. Using recent advances in probabilistic machine learning, our method is able to generate a flexible tree structure of biclusters while keeping computations feasible. We showed that the model achieves state-of-the-art performance on a large data set, and that the model naturally lends itself to hierarchical content-based information retrieval. Finally, we highlighted how the information retrieval functionality can be used to mine for novel biological knowledge, via a case study that provides insight into the potential novel role of miR-224 in the association between melanoma and non-Hodgkin lymphoma.

Searching for functional modules. Functional gene modules and protein complexes have been sought from both protein-protein interaction and gene expression data with various clustering-type methods. We have devised a combined generative model for these data that directly models noise in both data types [13]. The model outperforms other state-of-the-art methods in the task of discovering functional modules. In addition, it is able to detect overlapping modules, in which proteins may have different roles.

References

[1] J. Caldas and S. Kaski Bayesian Biclustering with the Plaid Model. In J. Príncipe, D. Erdogmus and T. Adali, editors, *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing XVIII*, 2008.

[2] J. Caldas, N. Gehlenborg, A. Faisal, A. Brazma, and S. Kaski. Probabilistic retrieval and visualization of biologically relevant microarray experiments. *Bioinformatics*, 25(12):i145–i153, 2009.

[3] J. Caldas, N. Gehlenborg, A. Faisal, A. Brazma, and S. Kaski. Probabilistic retrieval and visualization of biologically relevant microarray experiments. *BMC Bioinformatics*, 10(suppl.13):P1, 2009.

[4] J. Caldas and S. Kaski. Generative tree biclustering for information retrieval and microrna biomarker discovery. In *Proceedings of the 14th International Conference on Research in Computational Molecular Biology*, 2010. To appear.

[5] I. Huopaniemi, T. Suvitaival, J. Nikkilä, S. Kaski, and M. Orešič. Two-way analysis of high-dimensional collinear data. *Data Mining and Knowledge Discovery*, 19(2):261–276, 2009.

[6] I. Huopaniemi and T. Suvitaival and J. Nikkilä and M. Orešič and S. Kaski. Multi-Way, Multi-View Learning. In *NIPS 2009 workshop on Learning from Multiple Sources with Applications to Robotics*, 2009

[7] P. Jaspers, T. Blomster, M. Brosché J. Salojärvi, R. Ahlfors, J. Vainonen, R. Reddy, R. Immink, G. Angenent, F. Turck, K. Overmyer, and J. Kangasjärvi. Unequally redundant rcd1 and sro1 mediate stress and developmental responses and interact with transcription factors. *The Plant Journal*, 60(2):268–279, 2009.

[8] L. Lahti, L.L. Elo, T. Aittokallio, and S. Kaski. Probabilistic analysis of probe reliability in differential gene expression studies with short oligonucleotide arrays. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, to appear.

[9] L. Lahti, S. Myllykangas, S. Knuutila, and S. Kaski. Dependency detection with similarity constraints. In *Proc. MLSP 2009 IEEE International Workshop on Machine Learning for Signal Processing*, 2009.

[10] J. Nikkilä, M. Sysi-Aho, A. Ermolov, T. Seppänen-Laakso, O. Simell, S. Kaski, and M. Orešič. Gender dependent progression of systemic metabolic states in early childhood. *Molecular Systems Biology*, 4:197, 2008.

[11] M. Orešič *et al.* Dysregulation of lipid and amino acid metabolism precedes islet autoimmunity in children who later progress to type 1 diabetes. *Journal of Experimental Medicine*, 205(13):2975–2984, 2008.

[12] H. Parkinson *et al.* Arrayexpress update – from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Research*, 37(Database issue):D868–D872, 2009.

[13] J. Parkkinen and S. Kaski. Searching for functional gene modules with interaction component models. *BMC Systems Biology* 2010 4:4.

[14] S. Rogers, A. Klami, J. Sinkkonen, M. Girolami, and S. Kaski. Infinite Factorization of Multiple Non-parametric Views. *Machine Learning*, 79(1–2):201–226, 2010.

[15] A. Subramanian *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, page 0506580102, 2005.

[16] A. Tripathi, A. Klami, and S. Kaski. Simple integrative preprocessing preserves what is shared in data sources. *BMC Bioinformatics*, 9:111, 2008.

[17] A. Tripathi, A. Klami, and S. Kaski. Using dependencies to pair samples for multi-view learning. In *Proceedings of ICASSP 09, the International Conference on Acoustics, Speech, and Signal Processing*, pages 1561–1564, 2009.