Chapter 5

Bioinformatics

Samuel Kaski, Elisabeth Georgii, Arto Klami, José Caldas, Ali Faisal, Ilkka Huopaniemi, Suleiman Ali Khan, Leo Lahti, Juuso Parkkinen, Tommi Suvitaival

5.1 Introduction

The accumulation of different types of high-throughput measurement data yields unprecedented opportunities to study specific biological questions in context of the big picture of genome biology. It remains a major challenge how to process, analyze, and exploit this wealth of data such that the findings generate useful biomedical hypotheses and advance our understanding of cellular processes. As the number of molecular players, such as genes and metabolites, is extremely large compared to the number of available measurements, deciphering their function and functional interactions is highly non-trivial. In addition, regulatory elements within the genomic sequence are known only to a small extent, and the number of potential candidate regions is enormous. Thus, a main concern of computational systems biology is to detect statistical relationships between data points as well as variables in high-dimensional and potentially heterogeneous data spaces.

Our research focuses on three major topics. The first theme is translational modeling in medical studies; the ultimate goal is to predict biological responses to treatment across different tissues as well as from model organisms to human. The second theme is data-driven comparison and retrieval of gene expression experiments; measurements from different laboratories and different biological conditions are brought together in a common modeling framework to discover similarities or dissimilarities of samples regarding their transcriptional characteristics. The third theme is data integration, modeling of heterogeneous data sets that provide multiple views on the same biological samples or entities, e.g., gene expression measurements, genome methylation profiles, and copy number changes. The task is to detect shared aspects as well as source-specific aspects by looking at dependencies between the views. The three topics are described in more detail in the following sections. In addition, we have developed probabilistic models for decomposing biological networks into functional modules [12], and for estimating probe reliability in microarray measurements [8].

We have worked in close collaboration with VTT (Prof. M. Orešič), Haartman Institute (Prof. S. Knuutila), European Bioinformatics Institute EBI (Prof. A. Brazma), Department of Biological and Environmental Sciences at University of Helsinki (Prof. J. Kangasjärvi), Institute for Molecular Medicine Finland FIMM (Prof. O. Kallioniemi), and Institute of Biomedicine (Dr. Sampsa Hautaniemi).

5.2 Translational modeling for molecular medicine

We develop probabilistic machine learning methods for translational tasks motivated by research questions in molecular medicine. We are addressing computational modeling problems, with the aim to ultimately assist in approaching the following tasks: (i) to predict the response to disease and its medical treatments in the complex biological system of a human being, based on experiments on model organisms and cell lines, and (ii) to decrease the need for invasive operations on human patients, by detecting dependencies between views that are hard and easy to observe (e.g., study of the state of an inner organ based on blood levels).

In our research, we have developed machine learning methods for estimating the effects of multiple experimental factors. These solutions take ANOVA-type modeling beyond the possibilities of the classical approaches. We have presented ways of detecting similar responses between multiple tissues of a patient, and between the patient and a model organism. We have utilized the novel methods in current metabolomic studies of human diseased and their medical treatments.

Disease-related dependencies between multiple tissues. We developed a data fusion method [5], which allows us to detect dependencies between multiple tissues of a biological organism that are related to known experimental factors, such the disease status and medical treatment (Figure 5.1b). Many diseases such as cancer may be located in a specific organ whose state is not directly observable without invasive operations. Our method provides a way of making predictions of the hard-to-measure tissue via more readily observable samples, such as from the blood.

Multi-way modeling made possible for heterogeneous clinical data sets. We have developed ANOVA-type modeling of responses to multiple experimental factors further for heterogeneous time series data [6] (Figure 5.1c). One of the major complications in the analysis of clinical studies of humans has been the heterogenity of individual histories in the medical records. By utilizing dynamical generative models, we separated progression into disease from normal aging-related development of individuals. For another clinical study [18], our ANOVA-type modeling approach for high-dimensional data was extended to the repeated measures setting, where the blood levels of each patient are observed both before and after the medical treatment.

Matching objects of multiple views. The major obtacle for translational studies is the lack of one-to-one mapping between the biological systems of the different organisms. Even the approximate mapping is often unknown. We have developed a novel matching algorithm for simultaneously (i) learning a metric to maximize the dependency between two data sets, and (ii) matching the objects between the data sets [16].

Disease-related responses across several species. The goals of translational cross-species modeling are to (i) find similarities between the responses measured in two domains (human, model organism), and (ii) predict the outcome of a new intervention in one of these domains based on a similar realized experiment in the other domain. An important application lies in pharmaceutics, where the effect of a new drug on the development of

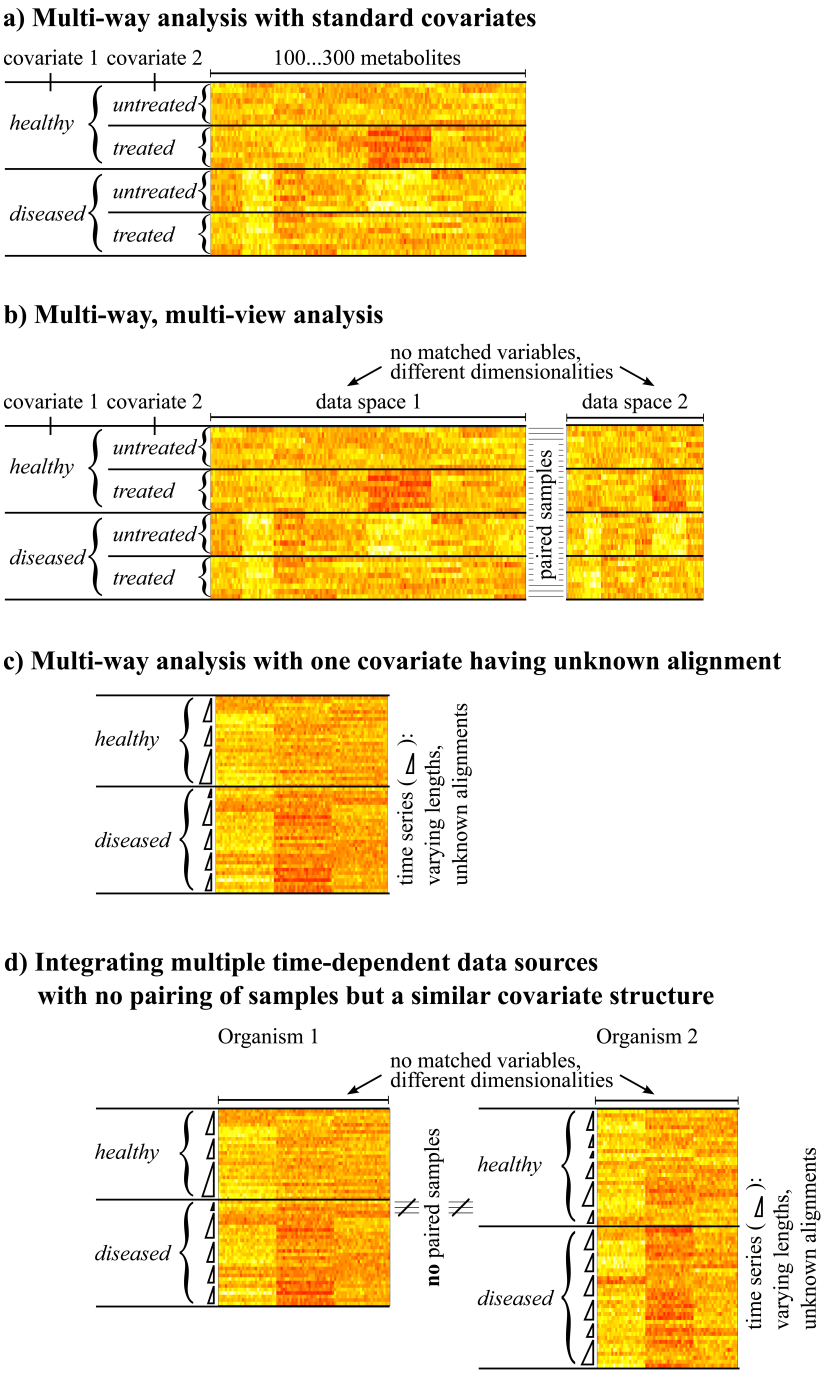

Figure 5.1: Illustration of the four data analysis tasks common in translational modeling for molecular medicine.

a human disease is studied with animals or tissues grown in laboratory, and later further tested with clinical studies on human patients. From the computational perspective, the translation of experiments is an unsolved problem, as neither variables nor samples are matched between the domains (Figure 5.1d).

We have introduced a model for matching groups of variables between the two domains based on their similarity in responses to relevant experimental factors [6]. Further, we separated domain-specific responses from the responses shared by the domains [13, 14].

Model organism study for type 1 diabetes. Metabolic development in children progressing into Type 1 Diabetes (T1D) is not well understood. As members of an interdisciplinary consortium, we have studied the development of T1D through a model organism [15]. By comparing the results with a similar follow-up study of human children, we found out that before the onset of the disease, female NOD mice exhibit the same lipidomic pattern as pre-diabetic human children. The results suggest alternative metabolic-related pathways as therapeutic targets to prevent the disease. These biological findings were made possible by methods for learning a data-driven model that maps human and mouse lipidomes.

We have proposed translational modeling methods for detecting dependencies between heterogeneous data sets as well as estimating and predicting effects of relevant experimental factors across domains. The methods have been designed to work with high-dimensional biological real-world data.

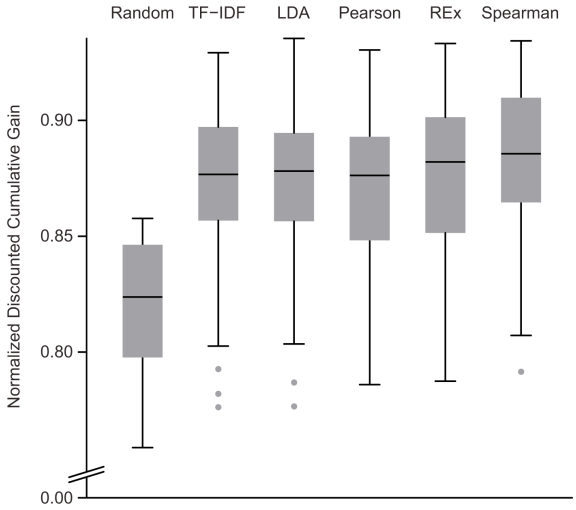

Figure 5.2: Figure taken from Caldas *et al.* [1]. Data-driven retrieval performance based on Normalized Discounted Cumulative Gain (NDCG); a measure of effectiveness of a search engine. The box plot summarize the distribution NDCG results for 219 interpretable query comparisons. "LDA" corresponds to our earlier method while REx corresponds to [1].

5.3 Data-driven comparison and retrieval of gene expression experiments

Considerable effort has been spent on collecting gene expression measurements into huge public repositories. This has opened up the door to large-scale comparisons and meta-analysis of data from different experiments. We have developed probabilistic methods that assist in these analysis tasks. In addition, we have introduced the concept of model-based retrieval of relevant biological experiments.

Content-based retrieval of relevant experiments. In previous work, we developed the first prototype of a content-based retrieval engine for biological experiments (REx: Retrieval of Relevant Experiments). To complement keyword search functionalities provided by most repositories for retrieval of similarly annotated studies, we developed probabilistic machine learning methods that relate gene expression studies through their actual measurement data, along with visualization tools that allow exploring and interpreting the results. The "model of biology" underlying our retrieval method is both data- and knowledge-driven: we use enrichment analysis for known functional gene sets (pathways) to obtain a representation of expression data that is comparable across measurement platforms. In [1], we extended the REx work to handle arbitrary experimental designs and to use a more accurate approach for modeling the activity of gene sets. In addition, the new REx model takes into account correlations in the activity of gene expression patterns. We also proposed a novel performance evaluation approach that is based on the Experimental Factor Ontology (EFO) of the ArrayExpress database and thereby much more scalable than manual relevance classification.

In a thorough comparison with alternative methods, REx performs competitively (see Figure 5.2). The advantage of our method lies in the interpretability of search results in terms of differential expression patterns. A previously unknown connection between

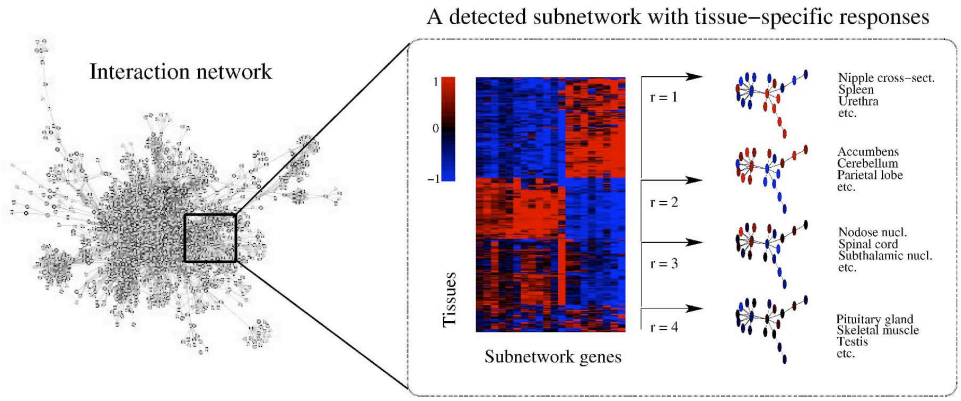

Figure 5.3: Figure taken from [10]. Organism-wide analysis of transcriptional responses in a human pathway interaction network reveals physiologically coherent activation patterns and tissue-specific regulation. One of the subnetworks and its tissue-specific responses, as detected by the NetResponse algorithm is shown. The expression of each gene is visualized with respect to its mean level of expression across all samples.

differential expression of the *SIM2s* gene and malignant pleural mesothelioma (MPM) suggested by our method in one of the case studies was experimentally verified in a new set of mesothelioma samples. Our work shows that the relatively unexplored paradigm of data-driven information retrieval in transcriptomics data offers the possibility of obtaining novel biological findings based on existing data, and holds the potential to ultimately accelerate biomedical research. A further extension of REx based on targeted regulatory models of gene expression has been submitted for publication.

Network-guided transcriptional response patterns. Different biological conditions (and tissues) can share the same cellular processes, which can be characterized by coordinated up- and down-regulation patterns in a specific set of genes, building so-called transcriptional signatures. Pathways and functional gene sets stored in public databases are typically not provided with information on the biological context of activation and generally too broad to define condition-dependent transcriptional signatures. We have developed an algorithm to detect gene sets that partition the biological conditions into groups where each group is characterized by a coherent activation pattern, modeled by a specific underlying signature (see Figure 5.3) [10]. The patterns are learned directly from gene expression data; to guide the analysis towards biologically interpretable signatures, the method exploits a network of known gene interactions to incrementally build larger candidate gene sets.

Hierarchical biclustering. Biclustering is the computational task of simultaneously clustering objects and inferring which features of the objects contribute to the grouping. Biclustering approaches are very popular in gene expression analysis, assisting in simultaneously uncovering relationships among biological samples and among genes. Our approach [2, 3] has two main contributions to the biclustering world: First, it applies the Bayesian framework to rigorously account for noise and uncertainty. Second, it learns a hierarchical tree structure for the samples, assigning characteristic genes to the nodes in the hierarchy. The model additionally yields natural information retrieval relevance

measures that can be used for relating samples to a query, making it eligible for the REx applications described above. The method outperformed four state-of-the-art biclustering procedures on a large miRNA data set.

5.4 Detection of dependencies between heterogeneous biological data types

Living cells are extremely complex systems, and hence integration of information from multiple sources is needed for accurate identification of underlying biological processes. 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 sample, but on different variables. The task is to detect aspects that are shared between different sources as well as source-specific components.

Bayesian Group Factor Analysis for understanding drug action mechanisms. We decomposed the dependencies between drug structures and their biological responses in multiple diseases, using a novel method called *Group Factor Analysis (GFA)* [17]. Unlike standard QSAR methods, which relate drug properties and univariate responses, we find relationships between a set of structural descriptors of drugs and their genome-wide responses. GFA is a novel extension to factor analysis that models dependencies between sets of variables ("views") instead of variables, representing them as group-wise sparse factors (see Figure 5.4). Unlike existing methods, GFA finds sparse factors shared by subsets of views (most interesting) along with those shared by all and those specific to one view.

In [7], we present details of the decomposed drug response relationships. With GFA, we are able to find factors that capture variation between chemical descriptors and biological responses in one, two, or all three diseases. The factors form hypotheses about drug response patterns, allowing us to relate specific chemical descriptors with targeted cellular responses. We find four main types of factors: (i) Factors shared by the chemical view and a subset (one or two) of the cell lines. These factors give hypotheses for drug responses specific to cancer type and are hence the most interesting ones. (ii) Factors shared by all cell lines and the chemical space, representing drug effects common to all three subtypes of cancer. (iii) Factors shared by all cell lines but not the chemical space. They are either drug effects not captured by the specific chemical descriptors used, or common biological response to the modulation of two or more different targets which can not be captured by any common chemical description. (iv) Factors specific to one view represent "biological noise". Our analysis shows that the discovered factors not only capture meaningful biological dependencies but are also more predictive of protein targets than similarly but individually analyzed chemical and biological response spaces.

Survival-associated biomarkers from multi-view functional genomics data. Genomic instability is a hallmark of cancer and high-throughput measurements of copy-number variation data have become commonplace in cancers. Given that copy-number alterations are noisy, one of the most successful approaches in increasing the reliability of putative driver genes involved in tumor progression and drug resistance is integration of copy number data with transcriptomics data. In [11], we demonstrate the benefits of using a systematic computational framework to include algorithms that enable indentification of context and clinically important patient groups. The results provide genes and genomic regions that have survival effect in Glioblastoma or a clinically defined subset, such as temozolomide-treated patients, and thus facilitate translation of large-scale biomedical data to knowledge.

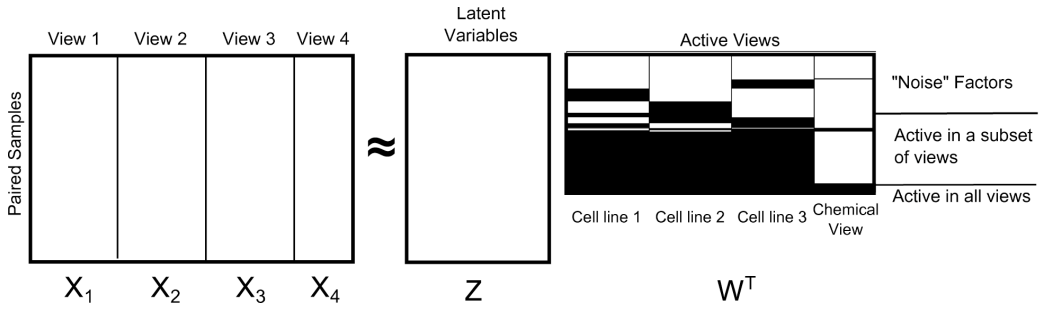

Figure 5.4: Illustration of the group factor analysis on three cell lines (diseases) and chemical view. The feature-wise concatenation of the data sets X_i is factorized as a product of the latent variables Z and factor loadings W. The factor loadings are group-wise sparse, so that each factor is active (gray shading) only in some subset of views (or all of them). The factors active in just one of the views model the structured noise, that is, variation independent of all other views, whereas the rest model the dependencies. The W shows the activity of the GFA factors in the 3 diseases (cell line 1: HL60, 2: MCF7, 3: PC3) and the drug descriptors (chemical view).

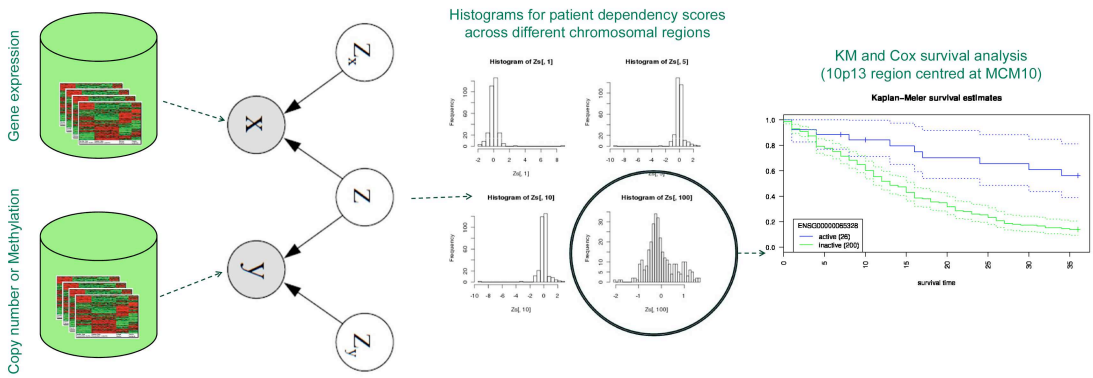

Figure 5.5: The analysis pipeline: A. Plate diagram for the canonical correlation analysis that captures the shared patterns Z from two data sources X and Z. B. Histograms of patients' contribution for four different genomic regions that have significantly high dependence scores; these histograms are used to form patient groups based on quantile clustering of the histogram. C. Sample Kaplan-Meier survival curve comparing the two patient groups for the genomic region centered at MCM10 gene; the patients with high dependence score have better survival than patients with low dependence scores, X-axis: months, Y-axis: percentage of GBM patients alive, dotted lines: 95% confidence intervals.

In [4], we present details of the approach used to identify potential genomic regions (or biomarkers) that effectively stratify patients in low and high survival groups. We first identify chromosomal regions that have high dependency between gene expression, methylation, and copy number changes, and then form patients groups from the regions and check whether the identified genomic aberrations have survival association (see Figure 5.5). The integration model is based on our earlier method for constrained canonical correlation analysis [9]. In addition, we incorporate suitable priors that model the positive correlation between gene expression and copy number data and the negative correlation between gene expression and methylation data. Furthermore, we incorporate sample-specific covariates in advanced survival analysis techniques. Results on Glioblastoma multiforme (GBM)

patient measurements identify known and novel genomic regions that may contribute to GBM progression and drug resistance.

References

[1] J. Caldas, N. Gehlenborg, E. Kettunen, A. Faisal, M. Ronty, A.G. Nicholson, S. Knuutila, A. Brazma and S. Kaski. Data-Driven Information Retrieval in Heterogeneous Collections of Transcriptomics Data Links SIM2s to Malignant Pleural Mesothelioma. *Bioinformatics*, 28(2):i246–i253, 2012.

[2] J. Caldas and S. Kaski. Hierarchical generative biclustering for microRNA expression analysis. *Journal of Computational Biology*, 18:251–261, 2011.

[3] J. Caldas and S. Kaski. Hierarchical generative biclustering for microRNA expression analysis. In *Research in Computational Molecular Biology, Proceedings of 14th Annual International Conference RECOMB 2010*, pages 65–79. Springer, 2010.

[4] A. Faisal, R. Louhimo, L. Lahti, S. Hautaniemi and S. Kaski. Biomarker discovery via dependency analysis of multi-view functional genomics data. In *NIPS 2011 workshop "From Statistical Genetics to Predictive Models in Personalized Medicine"*, 2011. Extendend abstract.

[5] I. Huopaniemi, T. Suvitaival, Janne Nikkilä, M. Orešič, and S. Kaski. Multivariate multi-way analysis of multi-source data. *Bioinformatics*, 26:i391–i398, 2010. (ISMB 2010)

[6] I. Huopaniemi, T. Suvitaival, M. Orešič, and S. Kaski. Graphical multi-way models. In J.L. Balcázar, F. Bonchi, A. Gionis and M. Sebag, editors, *Machine Learning and Knowledge Discovery in Databases. Proceedings of the European Conference on Machine Learning, ECML PKDD 2010*, volume I, pages 538–553. Springer, 2010.

[7] S.A. Khan, S. Virtanen, A. Klami, K. Wennerberg and S. Kaski. Decomposing Drug Response Patterns using Bayesian Group Factor Analysis. In *NIPS 2011 workshop "Machine Learning in Computational Biology"*, 2011. Abstract.

[8] L. Lahti, 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*, 8:217–225, 2011.

[9] L. Lahti and O.-P. Huovilainen. Dependency modeling toolkit. In *MLOSS workshop at ICML 2010*, 2010. Computer program.

[10] L. Lahti, J. Knuuttila, and S. Kaski. Global modeling of transcriptional responses in interaction networks. *Bioinformatics*, 26:2713–2720, 2010.

[11] R. Louhimo, V. Aittomäki, A. Faisal, M. Laakso, P. Chen, K. Ovaska, E. Valo, L. Lahti, V. Rogojin, S. Kaski, and S. Hautaniemi. Systematic use of computational methods allows stratifying treatment responders in glioblastoma multiforme. In *Proceedings of CAMDA 2011 conference, "Critical Assessment of Massive Data Analysis"*, 2011.

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

[13] T. Suvitaival, I. Huopaniemi, M. Orešič, and S. Kaski. Cross-species translation of multi-way biomarkers. In T. Honkela, W. Duch, M. Girolami, and S. Kaski, editors, *Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN), Part I*, volume 6791 of *Lecture Notes in Computer Science*, pages 209–216. Springer, 2011.

[14] T. Suvitaival, I. Huopaniemi, M. Orešič, and S. Kaski. Detecting similar high-dimensional responses to experimental factors between human and model organism. In *NIPS 2011 workshop "From Statistical Genetics to Predictive Models in Personalized Medicine"*, 2011. Extendend abstract.

[15] M. Sysi-Aho, A. Ermolov, P.V. Gopalacharyulu, A. Tripathi, T. Seppänen-Laakso, J. Maukonen, I. Mattila, S.T. Ruohonen, L. Vähätalo, L. Yetukuri, T. Härkönen, E. Lindfors, J. Nikkilä, J. Ilonen, O. Simell, M. Saarela, M. Knip, S. Kaski, E. Savontaus, and M. Orešič. Metabolic regulation in progression to autoimmune diabetes. *PLoS Computational Biology*, 7:e1002257, 2011.

[16] A. Tripathi, A. Klami, M. Orešič, and S. Kaski. Matching samples of multiple views. *Data Mining and Knowledge Discovery*, 23:300–321, 2011.

[17] S. Virtanen, A. Klami, S.A. Khan, and S. Kaski. Bayesian Group Factor Analysis. To appear in *Fifteenth International Conference on Artificial Intelligence and Statistics AISTATS* 2012, *pre-print at: arXiv:1110.3204 [stat.ML]*

[18] L. Yetukuri, I. Huopaniemi, A. Koivuniemi, M. Maranghi, A. Hiukka, H. Nygren, S. Kaski, M.-R. Taskinen, I. Vattulainen, M. Jauhiainen, and M. Orešič. High Density Lipoprotein structural changes and drug response in lipidomic profiles following the long-term fenofibrate therapy in the FIELD substudy. *PLoS ONE*, (8):e23589, 2011.