

---

# Detecting similar high-dimensional responses to experimental factors from human and model organism

---

**Tommi Suvitaival**  
Helsinki Institute for  
Information Technology HIIT,  
Department of  
Information and Computer Science,  
Aalto University  
tommi.suvitaival@aalto.fi

**Ilkka Huopaniemi\***  
Helsinki Institute for  
Information Technology HIIT,  
Department of  
Information and Computer Science,  
Aalto University  
ilkka.huopaniemi@tkk.fi

**Matej Orešič**  
VTT  
Technical Research Centre of Finland  
matej.oresic@vtt.fi

**Samuel Kaski**  
Helsinki Institute for  
Information Technology HIIT,  
Aalto University and  
University of Helsinki  
samuel.kaski@aalto.fi

## Abstract

We present a Bayesian model for analysing the effect of multiple experimental factors in two-species studies without the requirement of *a priori* known matching. From model studies of human diseases, conducted using \*omics technologies and various model organisms, the question emerges: is there something similar in the molecular responses of the different organisms under certain conditions, such as healthy vs. diseased? Our approach provides a generative model for the task of analysing multi-species data, naturally taking into account the additional information about the affecting factors such as gender, age, treatment, or disease status.

## 1 Introduction

Integration of 'omics' data types is in the center of research towards personalized medicine. Typically, the data types that are being integrated either describe (1) the same individual, e.g. the gene expression measurements and clinical information, or (2) similar individuals, e.g. gene expression and copy number aberration data from patients of the same cancer type.

We address the problem of data integration in a setting where the data sets come from two different species. Our task is to find similar response patterns in the two species under a similar disease, without making assumptions of the matchings of variables across the two data sets (see Figure 1a). In practice, this problem emerges when we want to compare the activities of genes or the concentrations of their end products between the two species: the genes and their products' functions are not necessarily conserved between the two species, and even if they are, the conservation is unknown.

The motivation for this line of work comes from translational medicine: if we are able to find similarities in the responses of human and a model organism, we want to utilize this information in

---

\*Currently at Mount Sinai School of Medicine.

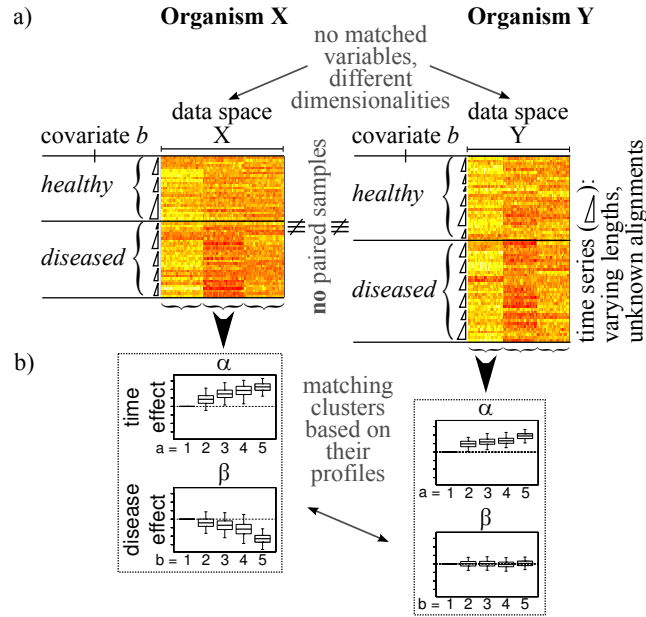


Figure 1: (a) Schematic of the cross-species analysis problem. Samples (rows) are organized into groups according to experimental factors (covariates) such as health status or age. Matching of both samples and variables (columns) between the two data sets is unknown. (b) A simplified illustration of the method we propose to solve the question of what is similar between the non-matched data sets. Response profiles of variable groups are compared within the model to match similar ones.

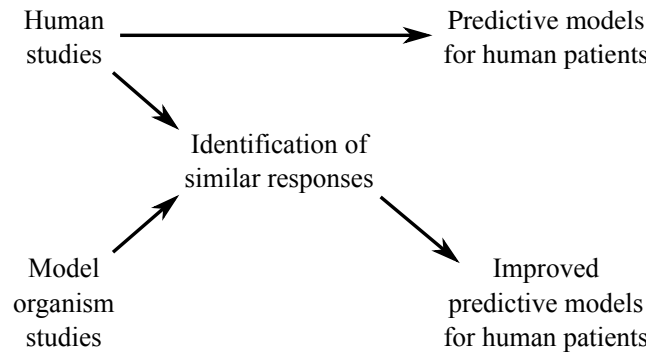


Figure 2: The idea of model organism studies improving the predictive statistical models for human diseases. We propose to do it in two steps: (1) by identifying similar responses to experimental factors in the studies of different species, and (2) utilizing the similarities to achieve more robust predictive models.

the design of further experiments aiming at individualized predictions of the disease (see Figure 2). Cross-species modeling is an emerging field within translational medicine, which has important applications in pharmaceutical research. During the recent years, few promising machine learning approaches have been presented for translating information from a model organism studies onto humans [1, 2, 3, 4].

Recently, lot of attention has been paid on how to take various experimental factors into account in medical data analysis [5]. Clinical information such as gender and age should be fed into the statistical models, as well as the possibly unknown stage of the disease. We use a recently-introduced graphical modeling framework [6] to simultaneously estimate the effects of known and partially unknown experimental factors. The approach is based on the assumption that the data contains groups of variables that respond to the experimental factors in a coherent fashion. In this way, the

model is able to reduce the analysis problem from the original high number of dimensions onto a small number of latent components. This type of a model has been earlier shown to be suitable for studies of metabolite concentrations in biological tissue, as correlations in the concentrations of related metabolite molecules are typically high [7].

We further develop the modeling framework into detecting similarities in the responses to experimental factors between two non-matched data sets (see Figure 1b). Most importantly, the proposed model clearly separates shared responses from data set-specific responses, thus allowing us to study the interesting shared part, while still properly explaining also the specific variation. The identification of shared responses between data sets is a step towards the translation of an experiment, where after learning the joint response structure, we could predict the response to a new experimental factor in one data set, based on an observed response in the other.

## 2 Model

The model that we present addresses the problem of identifying shared and specific responses between two non-matched data sets. We assume a similar experiment design in the two data sets in order to learn their shared responses.

The model aligns time series of varying lengths and simultaneously estimates the responses to the experimental factors, one of which is the state-aligned time. The hierarchical Bayesian model is capable of handling uncertainty across all levels, in contrast to a straightforward successive dimensionality reduction and analysis of variance (ANOVA). In the following sections, we present the building blocks of the graphical model, for which inference is carried via Gibbs sampling.

### 2.1 Dimensionality reduction and response to experimental factors

The model learns a small number of latent components from the higher-dimensional data set  $\mathbf{x}$ . The generative model for the observed sample  $\mathbf{x}_j$  (column  $j$  of the data matrix  $\mathbf{x}$ ) is

$$\mathbf{x}_j \sim \mathcal{N}(\boldsymbol{\mu} + \mathbf{V}\mathbf{x}_j^{\text{lat}}, \boldsymbol{\Lambda}) \quad (1)$$

$$\mathbf{x}_j^{\text{lat}} | (a_j, b_j) \sim \mathcal{N}(\boldsymbol{\alpha}_{a_j} + \boldsymbol{\beta}_{b_j} + (\boldsymbol{\alpha}\boldsymbol{\beta})_{a_j, b_j}, \mathbf{I}), \quad (2)$$

where  $\mathbf{x}_j$  is a  $p$ -dimensional sample vector from the  $n \times p$  data matrix,  $\boldsymbol{\mu}$  is a  $p$ -vector of variable means,  $\mathbf{V}$  is a  $p \times K$  sparse projection matrix,  $\mathbf{x}_j^{\text{lat}}$  the  $K$ -vector of latent components from the  $K \times n$  latent space matrix, and  $\boldsymbol{\Lambda}$  is a diagonal residual variance matrix with diagonal elements  $\sigma_i^2$ . Effects of the experimental factors are estimated in the  $K$ -dimensional latent space. In Equation 2, the model is written down for the case of two experimental factors with factor levels  $a_j$  and  $b_j$  selecting the main effects  $\boldsymbol{\alpha}_{a_j}$  and  $\boldsymbol{\beta}_{b_j}$ , and an interaction effect  $(\boldsymbol{\alpha}\boldsymbol{\beta})_{a_j, b_j}$ . In the notation, covariates  $a_j$  and  $b_j$  independently select a corresponding row from the main effect matrices  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$ , respectively, and  $(\boldsymbol{\alpha}\boldsymbol{\beta})_{a_j, b_j}$  is an interaction effect vector of the combination  $a_j, b_j$ . For each additional experimental factor, one main effect and all possible interaction effects are introduced to Equation 2.

### 2.2 Alignment of time series of varying lengths

Typically one of the experimental factors is age or time since the disease was acquired. We solve the problem of heterogeneous time series via built-in hidden Markov model (HMM)-type state projection. The learned state allocations  $s$  are used as an experimental factor and the corresponding HMM latent variable is interpreted as the effect of the factor for the aligned sample group [6].

As previously, the latent components  $\mathbf{x}_j^{\text{lat}}$  are assumed to be generated by responses to experimental factors as

$$\mathbf{x}_j^{\text{lat}} | (s_j, b_j) \sim \mathcal{N}(\boldsymbol{\alpha}_{s_j} + \boldsymbol{\beta}_{b_j} + (\boldsymbol{\alpha}\boldsymbol{\beta})_{s_j, b_j}, \mathbf{I}). \quad (3)$$

Now, in addition to the responses, also the level of one of the factors,  $s_j$ , is learned from the data (instead of being known from the experiment design as  $a_j$  is in Equation 1). Then  $\boldsymbol{\alpha}_{s_j}$  is the HMM-aligned time effect. We restrict the HMM to a linear chain structure, which is a reasonable decision for patient progression data of our experiment.

### 2.3 Shared and specific responses to experimental factors

The previously presented building blocks constitute the generative process of one data set  $\mathbf{x}$ . In this section, we explain how this framework can be utilized as a model for shared variation in two data sets and for the identification of shared and data set-specific responses. The model not only estimates responses to experimental factors for two data sets, but also estimates the probabilities of each response pattern being shared (vs. data set-specific).

The model makes a flexible assumption [6] that the observed data vectors in the two data sets  $\mathbf{x}$  and  $\mathbf{y}$  are generated as a response to the experimental factors through a transformation  $f^x$  and  $f^y$ , respectively:

$$\begin{aligned} \mathbf{x}_j | (s_j, b_j) &= \boldsymbol{\mu}^x + f^x \left( \boldsymbol{\alpha}_{s_j}^{\text{sh}} + \boldsymbol{\beta}_{b_j}^{\text{sh}} + (\boldsymbol{\alpha}\boldsymbol{\beta})_{s_j, b_j}^{\text{sh}} \right) + f^x \left( \boldsymbol{\alpha}_{s_j}^x + \boldsymbol{\beta}_{b_j}^x + (\boldsymbol{\alpha}\boldsymbol{\beta})_{s_j, b_j}^x \right) + \boldsymbol{\varepsilon}^x \\ \mathbf{y}_i | (s_i, b_i) &= \boldsymbol{\mu}^y + f^y \left( \boldsymbol{\alpha}_{s_i}^{\text{sh}} + \boldsymbol{\beta}_{b_i}^{\text{sh}} + (\boldsymbol{\alpha}\boldsymbol{\beta})_{s_i, b_i}^{\text{sh}} \right) + f^y \left( \boldsymbol{\alpha}_{s_i}^y + \boldsymbol{\beta}_{b_i}^y + (\boldsymbol{\alpha}\boldsymbol{\beta})_{s_i, b_i}^y \right) + \boldsymbol{\varepsilon}^y, \quad (4) \end{aligned}$$

where symbols with superscript <sup>sh</sup> represent responses to experimental factors shared by the two data sets, and symbols with superscripts <sup>x</sup> and <sup>y</sup> represent data set  $\mathbf{x}$  and  $\mathbf{y}$ -specific responses, respectively.

The variables of data sets  $\mathbf{x}$  and  $\mathbf{y}$  are not *a priori* matched, and therefore also the projections from the latent components  $\mathbf{x}^{\text{lat}}$  and  $\mathbf{y}^{\text{lat}}$  onto observed data  $\mathbf{x}$  and  $\mathbf{y}$ , respectively, are most likely different. Due to the lack of *a priori* knowledge about the matching between the observed variables between the data sets  $\mathbf{x}$  and  $\mathbf{y}$ , the model needs to learn the previously unknown projections  $f^x$  and  $f^y$  from the low-dimensional responses onto the observed variables. Additionally, it is unknown, whether the latent components  $\mathbf{x}^{\text{lat}}$  and  $\mathbf{y}^{\text{lat}}$  are responding similarly to the experimental factors or not (whether they match or not). This question is solved by a built-in matching algorithm for the latent components, explained in the next section.

The modeling question for two non-co-occurring data sets with a multi-way experiment design becomes the following: Does some dimension of  $\mathbf{x}^{\text{lat}}$  respond to the experimental factors  $\mathbf{s}$  and  $\mathbf{b}$  similarly as one of  $\mathbf{y}^{\text{lat}}$ ? If it does, one can represent this pattern with *shared* response to experimental factors  $\boldsymbol{\theta}^{\text{sh}} = \{\boldsymbol{\alpha}^{\text{sh}}, \boldsymbol{\beta}^{\text{sh}}, (\boldsymbol{\alpha}\boldsymbol{\beta})^{\text{sh}}\}$ . This means that a group of correlated variables in data set  $\mathbf{x}$  matches with a group of variables in data set  $\mathbf{y}$ , represented by a dimension of  $\mathbf{x}^{\text{lat}}$  and  $\mathbf{y}^{\text{lat}}$ , respectively. If found, such matched components can be considered as robust multi-species biomarkers related to the experimental factors. If there is no match, the response to the covariates is modeled by data set-specific responses  $\boldsymbol{\theta}^x = \{\boldsymbol{\alpha}^x, \boldsymbol{\beta}^x, (\boldsymbol{\alpha}\boldsymbol{\beta})^x\}$ , and similarly for data set  $\mathbf{y}$ . Most importantly, the model estimates the confidence of the shared responses.

### 2.4 Matching

We propose the following measure for quantifying the quality of the match between two factors from different data sets: whether the matching is better than an average matching (over other pairs). On a meta-level the measure is intuitively appealing in the spirit of permutation tests, and it can be formulated more exactly by specifying what we mean by “better.” We will use probabilistic modeling to measure the relative goodness below.

The matching problem of the components is a combinatorial problem, where possible configurations of pairs need to be evaluated, judging for each pair how similarly they respond to the experimental factors. We resort to an iterative algorithm that attempts to change the matching of one component at a time.

After selecting a candidate pair, we compare it to an average pair (randomly selected, having one same peer as the candidate pair), and accept forming a link between them by a Metropolis criterion that compares the likelihoods of the two matchings. A reverse operation is to attempt to break a link by comparing an existing link between two clusters to an average (random) pair. The goodness (likelihood) of the matched pair is evaluated by comparing likelihoods of the two shared response profiles. Components with no pairs are modeled by data set-specific responses. Averaging over sampling iterations, we can estimate the probability for matchings and the response patterns. High probability of a particular pair indicates a found matching. Low probability of any pair indicates that there might not be suitable match for the factor in the other data set.

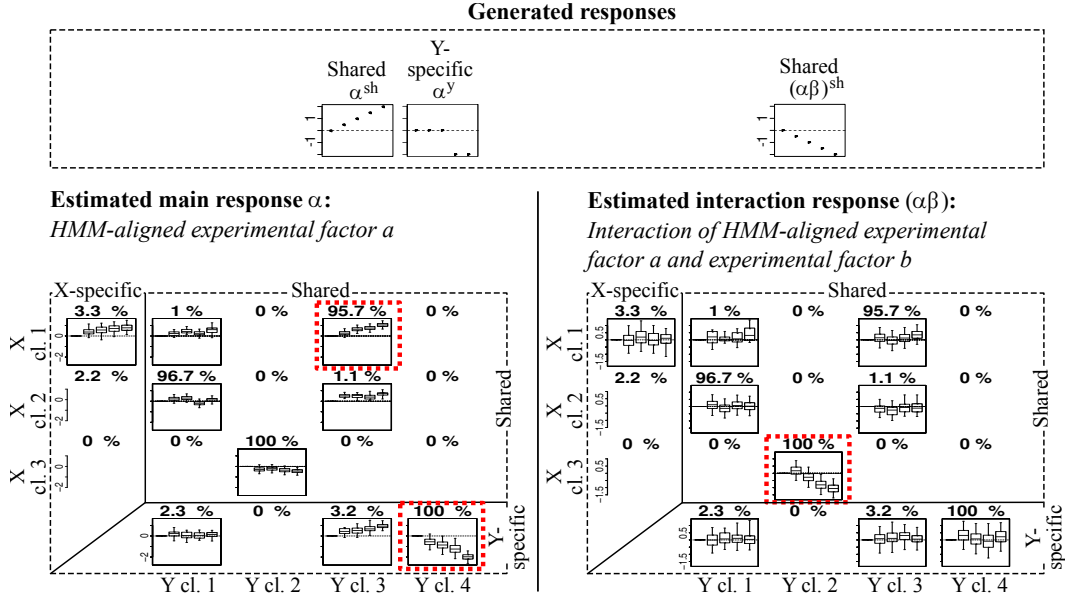


Figure 3: Matching results from simulated time-series data. Shown are the main responses of the HMM-aligned experimental factor  $a$  ( $\alpha$ ; left), and responses of the interaction of experimental factors  $a$  and  $b$  ( $\alpha\beta$ ); right). Topmost, the generated responses (the ground truth) are illustrated. In the lower parts, the table of estimated responses shows shared (top-right area) and data set-specific (left column and bottom row) responses  $\alpha$  (left table) and ( $\alpha\beta$ ) (right table). Rows and columns in the area of shared effects correspond to components in data sets  $x$  and  $y$ , respectively, and the table shows all possible combinations thereof. The found true match is highlighted by a dashed red box. The value on top of each plot shows the percentage of posterior samples, where the match was existent. The boxplots within each subplot represent posterior distributions of responses at different levels of the experimental factor. A distribution above or below zero with 95 % confidence (the whiskers) is considered significant.

### 3 Experiments

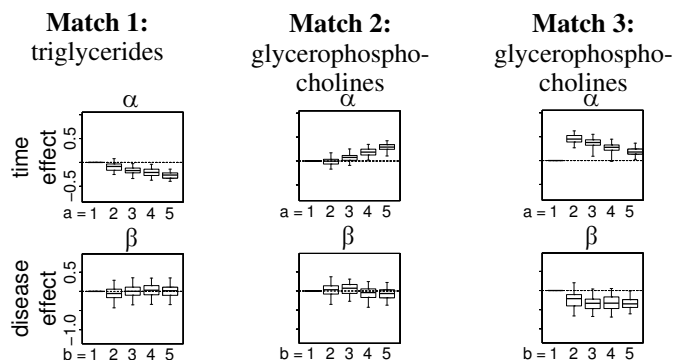
In this section, we demonstrate how the model works on high-dimensional toy data, and on biological data from human blood samples.

#### 3.1 Simulated Data

We generated from the model two data sets  $x$  and  $y$  with no co-occurring samples but only a similar two-way experiment design. There are 11 separate time series (“patients”) in both of the two data matrices, each series consisting of 5 to 15 time points. This results in 100 and 112 samples in total, and data matrices are 200- and 210-dimensional. The latent variables  $x_j^{\text{lat}}$  and  $y_j^{\text{lat}}$  are 3- and 4-dimensional, respectively. Two latent components in each data set were generated from a shared HMM chain with 5 states and one latent component in data set  $x$  was generated from a  $x$ -specific HMM chain.

We used the proposed model to simultaneously align the samples into matchable HMM states, learn the latent components as clusters of variables, search for the possible matches of the components between the two data sets, and model the responses to experimental factors acting on the found components. We *a priori* chose a model with 5 HMM states. During sampling, 150,000 burn-in samples and 150,000 posterior Gibbs samples were collected, of which every 50th sample was saved. The generated responses and inferred responses are shown in Figure 3. Our model found the previously generated components without mistakes on the included variables, and matched components between the data sets correctly.

a) Lipids to lipids



b) Lipids to metabolites

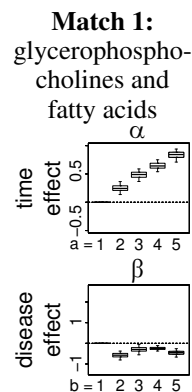


Figure 4: (a) The response profiles of three most confident matches of lipid groups. The three matches have clearly different response profiles in the two factors, time and the stage of the disease. Match 1 contains lipids from the family of triglycerides, and matches 2 and 3 constitute of different members of the glycerophosphocholine family. (b) The response profile of the most confident match between lipids and metabolites. The boxplots within each subplot represent the inferred posterior distributions of responses at different levels of the two experimental factors.

Table 1: The best-matched pair of a lipid and a metabolite cluster.

Lipids	Metabolites
GPCho(14:0/18:2)	X4.7.10.13.16.19.Docosahexaenoic.acid
GPCho(18:2/16:1)	X9.Octadecenoic.acid..Z.
GPCho(16:0/20:5)	Hexadecanoic.acid
	Phosphoric.acid

### 3.2 Biological data

We analysed biological data from a follow-up study of type 1 diabetes [8], where 53 lipid and 74 metabolite concentrations from blood samples were measured from two sets of human patients, respectively. In total, we had 1,153 and 417 samples from 124 and 37 patients, respectively.

We separated the normal development of young individuals from progression of the disease by labeling samples of patients, who acquired the disease, into four stages of progression of the disease using additional information about the antibody levels in blood. These stages were fixed as the levels of the experimental factor  $b_j$ , while the temporal alignment  $a_j$  of all patients was learned within the model by the HMM. We used a five-state HMM, and 6- and 15-dimensional latent variables to explain the correlated groups of lipids and metabolites in the data, respectively.

#### 3.2.1 Evaluating the match of lipid groups when ground truth available

First, we tested how the model finds matching, when the variables actually are co-occurring across the two data sets. Now we had a ground truth for the matching of individual variables between the two data sets but we did not give this information for the model. We created a pair of data sets with known matching by splitting the lipidomic data set into two groups of patients and used the two groups as data sets  $x$  and  $y$ .

As a result, we found out that the three strongest matches out of all possible combinations were correct. The response profiles of the three most confident matches are shown in Figure 4a.

### 3.2.2 Matching between lipids and metabolites

Next, we sought for matching groups between the lipidomic and metabolomic data sets. Some of the patients were the same in the two data sets, but we did not utilize this information to help the model.

The best match was a group of three glycerophosphocholine (GPCho) lipids to a group of four metabolites with probability of 19.7 % (see Table 1 for the molecule names and Figure 4b for the response profile). Three first of the metabolites in the list are fatty acids, which are building blocks for GPCho lipids. The found lipid and metabolite groups had a similar response pattern in time (increased concentration) and in the stages of the disease (decreased concentration).

## 4 Discussion

We presented a solution for identifying conserved responses between two species from experiments with several experimental factors and a time series structure. The model is applicable even in the extremely hard case of no *a priori* known matching between neither variables nor samples across the two data sets, but only a similar experiment design. The method estimates ANOVA-type responses to experimental factors for clusters of variables, and identifies and separates responses that are shared between the data sets and responses that are specific to one data set. The identified clusters of variables that respond to disease can be considered as robust biomarkers.

In this paper, we have shown the model to work on biological data sets, where the ground truth to the matching task is known. Next, we are looking forward to fusing data from human and a model organism, thus improving the predictions on the human disease.

### Acknowledgements

T.S., I.H. and S.K belong to the Finnish Centre of Excellence in Adaptive Informatics Research. The work was funded by Tekes MASI program and by Tekes Multibio project. I.H. is funded by the Graduate School of Computer Science and Engineering. S.K. is partially supported by EU FP7 NoE PASCAL2, ICT 216886.

## References

- [1] Samuel Kaski, Janne Nikkilä, Janne Sinkkonen, Leo Lahti, Juha E.A. Knuutila, and Christophe Roos. Associative clustering for exploring dependencies between functional genomics data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2:203–216, 2005.
- [2] Johannes Berg and Michael Lässig. Cross-species analysis of biological networks by Bayesian alignment. *Proceedings of the National Academy of Sciences*, 103(29):10967–10972, 2006.
- [3] Joseph Lucas, Carlos Carvalho, and Mike West. A Bayesian analysis strategy for cross-study translation of gene expression biomarkers. *Statistical Applications in Genetics and Molecular Biology*, 8(1):11, 2009.
- [4] Hai-Son Le and Ziv Bar-Joseph. Cross species expression analysis using a Dirichlet process mixture model with latent matchings. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1270–1278. 2010.
- [5] Jennifer Listgarten, Carl Kadie, and David Heckerman. Correction for hidden confounders in the genetic analysis of gene expression. *Proceedings of the National Academy of Sciences*, 2010.
- [6] Ilkka Huopaniemi, Tommi Suvitaival, Matej Orešič, and Samuel Kaski. Graphical multi-way models. In José Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag, editors, *Machine Learning and Knowledge Discovery in Databases (Proceedings of ECML PKDD 2010)*, Lecture Notes in Computer Science, pages 538–553. Springer, 2010.
- [7] Ilkka Huopaniemi, Tommi Suvitaival, Janne Nikkilä, Matej Orešič, and Samuel Kaski. Two-way analysis of high-dimensional collinear data. *Data Mining and Knowledge Discovery*, 19(2):261–276, 2009.
- [8] Matej 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.