Department of Information and Computer Science

Bayesian Multi-Way Models for Data Translation in Computational Biology

Tommi Suvitaival

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 171/2014

Bayesian Multi-Way Models for Data Translation in Computational Biology

Tommi Suvitaival

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall T2 of the school on 19 November 2014 at 12.

Aalto University
School of Science
Department of Information and Computer Science

Supervising professor
Prof. Samuel Kaski

Preliminary examiners
Dr. Laura Elo-Uhlgren, University of Turku, Finland
Dr. Lukas Käll, KTH Royal Institute of Technology, Sweden

Opponent
Asst. Prof. Anna Goldenberg, University of Toronto, Canada

Aalto University publication series
DOCTORAL DISSERTATIONS 171/2014

ISBN 978-952-60-5932-7 (printed)
ISBN 978-952-60-5933-4 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-5933-4

Unigrafia Oy
Helsinki 2014

Finland

Publication orders (printed book):
tommi.suvitaival@alumni.aalto.fi

Author
Tommi Suvitaival

Name of the doctoral dissertation
Bayesian Multi-Way Models for Data Translation in Computational Biology

Publisher School of Science

Unit Department of Information and Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 171/2014

Field of research Information and Computer Science

Manuscript submitted 9 September 2014 **Date of the defence** 19 November 2014

Permission to publish granted (date) 27 October 2014 **Language** English

☐ **Monograph** ☒ **Article dissertation (summary + original articles)**

Abstract

The inference of differences between samples is a fundamental problem in computational biology and many other sciences. Hypothesis about a complex system can be studied via a controlled experiment. The design of the controlled experiment sets the conditions, or covariates, for the system in such a way that their effect on the system can be studied through independent measurements. When the number of measured variables is high and the variables are correlated, the assumptions of standard statistical methods are no longer valid. In this thesis, computational methods are presented to this problem and its follow-up problems.

A similar experiment done on different systems, such as multiple biological species, leads to multiple "views" of the experiment outcome, observed in different data spaces or domains. However, cross-domain experimentation brings uncertainty about the similarity of the systems and their outcomes. Thus, a new question emerges: which of the covariate effects generalize across the domains? In this thesis, novel computational methods are presented for the integration of data views, in order to detect weaker covariate effects and to generalize covariate effects to views with unobserved data.

Five main contributions to the inference of covariate effects are presented: (1) When the data are high-dimensional and collinear, the problem of false discovery is curbed by assuming a cluster structure on the observed variables and by handling the uncertainty with Bayesian methods. (2) Prior information about the measurement process can be used to further improve the inference of covariate effects for metabolomic experiments by modeling the multiple layers of uncertainty in the mass spectral data. (3-4) When the data come from multiple measurement sources on the same subjects - that is, from data views with co-occurring samples - it is unknown, whether the covariate effects generalize across the views and whether the outcome of a new intervention can be generalized to a view with no observed data on that intervention. These problems are shown to be possible to solve by assuming a shared generative process for the multiple data views. (5) When the data come from different domains with no co-occurring samples, the inference of between-domain dependencies is not possible in the same way as with co-occurring samples. It is shown that even in this situation, it is possible to identify covariate effects that generalize across the domains, when the experimental design at least weakly binds the domains together. Then, effects that generalize are identified by assuming a shared generative process for the covariate effects.

Keywords ANOVA modeling, Bayesian modeling, computational biology, cross-species modeling, metabolomics, multi-view modeling, toxicogenomics

ISBN (printed) 978-952-60-5932-7 **ISBN (pdf)** 978-952-60-5933-4

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2014

Pages 170 **urn** http://urn.fi/URN:ISBN:978-952-60-5933-4

A? Aalto-yliopisto

Tiivistelmä

Aalto-yliopisto, PL 11000, 00076 Aalto www.aalto.fi

Tekijä
Tommi Suvitaival

Väitöskirjan nimi
Bayesilaisia monisuuntaisia malleja biologisten tietoaineistojen translaatio-ongelmaan

Julkaisija Perustieteiden korkeakoulu

Yksikkö Tietojenkäsittelytieteen laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 171/2014

Tutkimusala Tietojenkäsittelytiede

Käsikirjoituksen pvm 09.09.2014 **Väitöspäivä** 19.11.2014

Julkaisuluvan myöntämispäivä 27.10.2014 **Kieli** Englanti

☐ **Monografia** ☒ **Yhdistelmäväitöskirja (yhteenveto-osa + erillisartikkelit)**

Tiivistelmä

Eroavaisuuksien löytäminen näytteiden välillä on perustavanlaatuinen ongelma niin laskennallisessa biologiassa kuin muissakin tieteissä. Hypoteesia monimutkaisen järjestelmän toiminnasta voidaan tutkia tekemällä koe. Kokeen olosuhteet kontrolloidaan siten, että koeasetelman määrittelemien kovariaattien vaikutus tutkittavaan systeemiin voidaan todeta riippumattomien mittausten avulla. Jos mitattuja muuttujia on paljon ja niillä on keskinäisiä riippuvuuksia, perinteisten tilastollisten mallien olettamukset eivät päde. Tässä väitöskirjassa esitetään laskennallisia menetelmiä tähän ongelmaan ja sen jatko-ongelmiin.

Kun samanlainen koe tehdään useille samankaltaisille järjestelmille, kuten eri biologisille lajeille, saadaan "näkymiä" kokeen tuloksesta eri mittausavaruuksissa. Järjestelmien eroavaisuuksista seuraa kuitenkin epävarmuus tulosten yhteneväisyydestä ja kysymys siitä, mitkä kovariaattien vaikutukset yleistyvät tutkittaville järjestelmille? Tässä väitöskirjassa esitetään uusia laskennallisia menetelmiä havaintoaineistojen yhdistämiseen useista näkymistä, heikkojen kovariaattivaikutusten löytämiseen sekä vaikutusten yleistämiseen näkymiin, joista ei ole saatavilla vastaavia havaintoja.

Väitöskirja sisältää viisi pääkontribuutiota kovariaattien vaikutusten löytämiseen: (1) Kun havainnot ovat korkeaulotteisia ja niissä on muuttujien välisiä riippuvuuksia, väärien löydösten riskiä voidaan lieventää mallintamalla ilmiötä bayesilaisittain ja olettamalla, että muuttujat muodostavat ryhmiä. (2) Mittausmenetelmää koskevan prioritiedon tuominen malliin tarkentaa kovariaattien vaikutusten oppimista monitasoista mittauskohinaa sisältävistä metabolomiikkamittauksista. (3-4) Kun havainnot muodostuvat useasta mittausnäkymästä samoille mittauskohteille, on selvitettävä yleistyvätkö kovariaattien vaikutukset usealle näkymälle ja voidaanko uuden kokeen tulos yleistää näkymään, josta ei ole havaintoja uuden kokeen osalta. Nämä kysymykset ratkaistaan olettamalla, että näkymien havainnot ovat muodostuneet yhteisen generatiivisen prosessin kautta. (5) Kun havainnot muodostuvat useasta mittausnäkymästä mutta mittauksen kohteet eivät ole näkymien välillä samat, näkymien välisten riippuvuuksien löytäminen ei ole mahdollista samalla tavalla kuin silloin kun kohteet ovat samat. Väitöskirjassa osoitetaan, että tässäkin tapauksessa on mahdollista löytää näkymien välisiä riippuvuuksia ja niitä voidaan löytää tutkimalla näkymien yhteisten kovariaattien vaikutuksia.

Avainsanat ANOVA-mallitus, bayesilainen mallitus, laskennallinen biologia, lajienvälinen mallitus, metabolomiikka, usean näkymän mallitus, toksikogenomiikka

ISBN (painettu) 978-952-60-5932-7 **ISBN (pdf)** 978-952-60-5933-4

ISSN-L 1799-4934 **ISSN (painettu)** 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2014

Sivumäärä 170 **urn** http://urn.fi/URN:ISBN:978-952-60-5933-4

Preface

The research work for this thesis was done at Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University. The work was funded by Academy of Finland (The Finnish Centre of Excellence in Computation Inference Research, COIN; The Adaptive Informatics Research Centre, AIRC; and the project for Computational Modeling of the Biological Effects of Chemicals, ChemBio) and Tekes (MASI program and the Multibio project). Additionally, associated travel and a research visit were funded by the Finnish Doctoral Programme in Computational Sciences FICS and the Finnish Foundation for Technology Promotion.

The work was supervised by Prof. Samuel Kaski, to whom I am deeply grateful for all the advice, discussions and expertise that made this work possible, and for the opportunity to work in the group. I was privileged to learn so many things from a true researcher at the bleeding edge of machine learning and computational biology!

I equally want to thank the co-authors of the publications presented in the thesis for making this come true: Drs Ilkka Huopaniemi, Janne Nikkilä, Matej Orešič, Juuso Parkkinen, Simon Rogers and Seppo Virtanen. In addition to the research work, I am especially thankful to Ilkka for instructing me during the early stages of my doctoral studies, and to Simon for hosting my research visit to the School of Computing Science at University of Glasgow. I also thank the pre-examiners of this thesis—Drs Laura Elo-Uhlgren and Lukas Käll—for providing excellent comments on the thesis manuscript at the final stage towards the doctoral defense.

During the doctoral studies, I got to know many great colleagues, whose presence, insights and opinions significantly widened my perspective: Ali, Antti, Arto, Cho, Eemeli, Elina, Gayle, Hani, Ilari,

Jaakko, Jussi, Kalle, Kerstin, Konstantinos, Kristian, Leo, Maija, Manuel, Maria, Mehmet, Melih, Nicolau, Paula, Pejman, Pekka, Ricardo, Rito, Sami, Sohan, Suleiman, Tommi and Tuomas at Aalto, as well as Anna, Daniel, Edwin, Faiz, George, Joe, Rebecca, Rod, Shimin and Hugues in Glasgow, just to name "a few."

Finally, I am deeply grateful to my parents Eija and Raimo for supporting my journey towards the doctoral degree!

Espoo, October 30, 2014,

Tommi Suvitaival

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Ilkka Huopaniemi, Tommi Suvitaival, Janne Nikkilä, Matej Orešič, and Samuel Kaski. Two-way analysis of high-dimensional collinear data. *Data Mining and Knowledge Discovery*, Volume 19, Issue 2, Pages 261–276, 2009.

II Tommi Suvitaival, Simon Rogers, and Samuel Kaski. Stronger findings from mass spectral data through multi-peak modeling. *BMC Bioinformatics*, Volume 15, Article 208, 11 pages, 2014.

III Tommi Suvitaival, Simon Rogers, and Samuel Kaski. Stronger findings for metabolomics through Bayesian modeling of multiple peaks and compound correlations. *Bioinformatics*, Volume 30, Issue 17, Pages i461–i467, 2014.

IV Ilkka Huopaniemi, Tommi Suvitaival, Janne Nikkilä, Matej Orešič, and Samuel Kaski. Multivariate multi-way analysis of multi-source data. *Bioinformatics*, Volume 26, Issue 12, Pages i391–i398, 2010.

V Tommi Suvitaival, Juuso A. Parkkinen, Seppo Virtanen, and Samuel Kaski. Cross-organism toxicogenomics with group factor analysis. *Systems Biomedicine*, Volume 2, eLocation ID e29291, 9 pages, 2014.

VI Ilkka Huopaniemi, Tommi Suvitaival, Matej Orešič, and Samuel Kaski. Graphical multi-way models. In *Jose Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag, editors, Machine Learning and Knowledge Discovery in Databases — European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML PKDD 2010*, Barcelona, Spain, September 20–24, 2010, Proceedings, Part I, Pages 538–553, Springer, Berlin/Heidelberg, Germany, 2010.

VII Tommi Suvitaival, Ilkka Huopaniemi, Matej Orešič, and Samuel Kaski. Cross-species translation of multi-way biomarkers. In *Timo Honkela, Włodzisław Duch, Mark Girolami and Samuel Kaski, editors, Artificial Neural Networks and Machine Learning — ICANN 2011, 21st International Conference on Artificial Neural Networks*, Espoo, Finland, June 14–17, 2011, Proceedings, Part I, Pages 209–216, Springer, Berlin/Heidelberg, Germany, 2011.

Author's Contribution

Publication I: "Two-way analysis of high-dimensional collinear data"

All publications listed in the dissertation are a result of team work and each work was planned together with all the authors listed in the publication. Unless otherwise stated, the methods presented in the publications were designed together.

For Publication I, the author had the main responsibility for the implementation of the method and the experiments, and for the data analysis of the results: the author implemented the model for the covariate effects, tested the model, prepared the data, created the framework for the experiments, ran the experiments, had a main role at analyzing the results, created the figures for the publication and participated in the preparation of the manuscript.

Publication II: "Stronger findings from mass spectral data through multi-peak modeling"

The author had the main responsibility for the implementation of the method, for the design and implementation of the experiments, and for the preparation of the manuscript.

Publication III: "Stronger findings for metabolomics through Bayesian modeling of multiple peaks and compound correlations"

Similarly as with Publication II, the author had the main responsibility for the implementation of the method, for the design and implementation of the experiments, and for the preparation of the manuscript.

Publication IV: "Multivariate multi-way analysis of multi-source data"

The author had the main responsibility for the implementation of the method and the experiments, and for the data analysis of the results: the author implemented the model for the covariate effects, created the framework for the experiments, had a main role at analyzing the results, created the figures for the publication, and participated in the preparation of the manuscript.

Publication V: "Cross-organism toxicogenomics with group factor analysis"

The work was planned together with the other authors. The author had the main responsibility for solving the data translation problem, for the implementation of the experiments, for the data analysis of the results, and for the preparation of the manuscript: the author prepared the data for the model, enabling the data translation, ran the experiments, identified the associations between the molecular-level and organ-level observations based on the model, identified the enriched gene ontology terms and formulated and implemented the model-based retrieval of similar experiments.

Publication VI: "Graphical multi-way models"

The author had the main responsibility for the implementation of the method and the experiments, and for the data analysis of the results: the author implemented the models, created the framework for the experiments, ran the experiments, had a main role at analyzing the results, created the figures for the publication, and participated in the preparation of the manuscript.

Publication VII: "Cross-species translation of multi-way biomarkers"

The author had the main responsibility for the implementation of the method and the experiments, for the data analysis of results, and for the preparation of the manuscript: the author implemented the model, created the framework for the experiments, ran the experiments, had

a main role at analyzing the results, created the figures for the publication, and had a main role at the preparation of the manuscript.

List of Abbreviations and Symbols

Abbreviations

ANOVA	analysis of variance
ARD	automatic relevance determination
CCA	canonical correlation analysis
DNA	deoxyribonucleic acid
GFA	group factor analysis
GI	growth inhibition
HMM	hidden Markov model
LD_{50}	median lethal dose
m/z	mass-to-charge ratio
MANOVA	multivariate analysis of variance
MCMC	Markov chain Monte Carlo
PCA	principal component analysis
RNA	ribonucleic acid
RT	retention time

Symbols

x	scalar variable
x_i	scalar item i from the vector $\mathbf{x}$
$x_{j,i}$	scalar item from the row j and column i of the matrix $\mathbf{X}$
$\mathbf{x}$	vector variable
$\mathbf{x}_{-j}$	vector with all the items except the item j
$\mathbf{x}_{j,\cdot}$ or $\mathbf{x}_{\cdot,i}$	a vector consisting of the row j or the column i from the matrix $\mathbf{X}$, respectively
$\mathbf{X}$	matrix variable
$\mathbf{I}$	identity matrix
$\mathbf{X}^{\mathrm{T}}$	transpose of matrix $\mathbf{X}$
$\mathbf{X}^{-1}$	inverse of matrix $\mathbf{X}$
δ_0	Dirac delta function
$\delta_{a,b}$	Kronecker delta function
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	multivariate Gaussian probability density function with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$
$p(x)$	probability density function of the variable x
$p(\mathbf{x})$	multivariate probability density function of the vector variable $\mathbf{x}$
$\mathbb{R}_+$	positive real number
$\mathbb{R}^{P \times N}$	real-valued matrix with dimensions P-by-N
K	number of clusters, components or factors
N	number of samples
P	number of variables

1. Introduction

1.1 Motivation and Background

High-throughput measurement technologies, developed during the past decades, enable the quantification of molecular level changes in biological organisms. Biological processes at the molecular level are crucial for understanding diseases, since the most of the diseases result from perturbations in these processes. Also medical drugs perturb biological processes, ideally in an inverse way to the cause of the disease, thus restoring the organism to the normal state.

Since new drug compounds cannot be directly tested on humans due to their unknown and potentially dangerous effects, the drugs first go through rigorous testing on model organisms to assess both their therapeutic potential as well as their unwanted side-effects. Still after the testing phases on model organisms, it is unknown, which of the observed effects appear, when the drug is administered to a human. There is a need for computational methods that identify drug effects that generalize between the model organisms and humans, and to *translate* the expected outcome of a new experiment from model organisms to humans.

The data translation methods presented in this thesis enable the identification of similar responses between the organisms even though the observed organisms are different by their biological systems. The methods are based on the assumption that the experiments on different organisms share a similar experiment design that acts as a common ground for the identification of similar responses.

Modeling of differences between groups of samples from one organism is the starting point for the thesis. Even this is a non-trivial problem,

since the small number of samples compared to the number of observed variables renders traditional statistical methods ineffective for the reliable inference of the differences. This problem, known as the "small n, large p" problem, is prevalent in experiments on the molecular biology of an organism, where thousands of different molecules are measured on a handful of test subjects.

1.2 Contributions of the Thesis

The contributions of this thesis are distributed to seven peer-reviewed publications. The scientific contributions in each of the publications, referred by the Roman numerals, are summarized in the list below:

I. A hierarchical Bayesian model was introduced for inferring multi-way covariate effects from noisy and collinear data with a small number of samples but a high number of observed variables.

II. A Bayesian model was introduced to improve the inference of covariate effects from mass spectral measurements of the metabolome. The proposed method is a generative model both for the measurement process, integrating multiple spectral peaks from one chemical compound, and for the experiment design, inferring the effects of the experimental covariates.

III. The inference of weak covariate effects on metabolomic data was further improved by introducing a hierarchical Bayesian model that at two levels clusters the observed variables: peaks into latent compounds and, further, latent compounds into groups of compounds that respond coherently to the experimental covariates.

IV. A Bayesian multi-way, multi-view model was introduced to inferring covariate effects that generalize to multiple data views with co-occurring samples.

V. A multi-view biclustering model was introduced for identifying conserved molecular responses, when the interventions are experimented on multiple organisms. The generalization of the response from molecular measurements on model organisms to

humans at the population level was formulated as a task for retrieving similar interventions with known effects.

VI. A dynamical Bayesian model was introduced for inferring covariate effects, when the covariate is the time point of a time series and the points and the lengths of the series vary. The time series were aligned via dynamical modeling to enable the identification of consistent covariate effects. Further, it was shown that by modeling the matching experiment design, conserved covariate effects can be identified even between data sets with non-co-occurring samples.

VII. A dynamical Bayesian multi-way model was introduced for inferring multi-way covariate effects from non-co-occurring data sets by modeling the experiment design shared between the data sets. The model was shown to identify shared covariate effects between the data sets as well as covariate effects specific to a data set.

1.3 Organization of the Thesis

The background for the problem addressed in this thesis and for the Bayesian methodology used are presented in Chapters 2 and 3, respectively. The overview of the problem and contributions of this thesis are reviewed in the chapters that follow these introductory chapters.

The problem of the identification of covariate effects from high-dimensional and small sample-size data is presented in Chapter 4 and the Bayesian multi-way model (Publication I) is presented as a solution to the problem. Hierarchical multi-way models designed specifically for the inference of covariate effects from noisy mass spectral measurements of the metabolome (Publications II and III) are presented in Chapter 5 along with a review of the state-of-the-art measurement technology for metabolomics.

Bayesian multi-view methods for solving the problem of data translation between co-occurring views (Publications IV and V) are presented in Chapter 6. Solutions for the even more complex data translation problem, when the data sets do not share co-occurring samples (Publications VI and VII), are presented in Chapter 7.

2. Molecular Biology & Measurement Technologies

2.1 Introduction

The response of a biological organism to an environmental factor, such as an intruding chemical compound, is a complex cascade of events, whose details are still widely unknown. If these events can be understood properly, new treatments to diseases can be developed. Molecular measurement technologies have been invented to gain understanding of these events, and now the molecular response to an experimental factor can be quantified at multiple levels of the cascade.

Current technologies are still far from the entire quantification of the molecular activity of a human or even a single cell. Further, finding associations between environmental factors and changes in molecular activity is the more challenging the more complex the organisms under study is. However, the influence of environmental factors on the molecular balance can already be studied under controlled settings with model organisms (Joyce and Palsson, 2006): tissue extracts, cell lines, or even entire organisms such as single-cell yeasts, multi-cell worms or even small mammals such as rats.

The simultaneous quantification of thousands of molecular types sets demands for the analysis methods, since the number of tested and measured subjects is smaller than the number of quantified molecular types. Further, the integration of measurements from multiple levels of molecular biology, and even from multiple organisms, is an unsolved computational challenge. These two problems are addressed in this thesis.

Four main classes of biomolecules: deoxyribonucleic acid (DNA), ribonucleic acid (RNA), proteins and metabolites, all are subject to

regulation and changes resulting from environmental factors. The role of each of these main building blocks of molecular biology is reviewed in the following sections.

2.2 Genes

Genes are the biological cells' instructions for the construction of proteins. Since proteins are the main acting components in biological processes, genes essentially describe and regulate the operation of the cell.

In the cells of eukaryotic organisms such as yeast, rat and human, the genetic code is carried by deoxyribonucleic acid (DNA; Watson and Crick, 1953). DNA is a double-helical structure that consists of a sequence of *base* molecules. There are four different bases, which constitute the alphabet of the genetic code, and two of the bases—adenine and cytosine—are complementary to the other two—thymine and guanine, respectively. The complementary bases in two parallel base sequences bind together to form the double helix.

Three consecutive bases in the DNA sequence form a *codon*, which is a word in the genetic code. The sequence of codons describes the order in which a protein is constructed from amino acid molecules (Crick, 1968). With few exceptions, each unique codon corresponds to a specific amino acid.

DNA has a self-replication mechanism, which ensures that the daughter cells receive the same genetic information in the cell division (Meselson and Stahl, 1958). In the division, each of the two parts of the dividing cell receives a copy of the genetic code with the help of the DNA polymerase enzyme.

The entire genetic sequence written in the DNA, termed the genome, has been sequenced for human (Venter et al., 2001), among many other organisms. Origins of inheritable diseases and risk factors to diseases can be identified via sequencing and statistical analysis of inter-subject differences in the genetic code (Buetow et al., 2001; Burton et al., 2007).

Even though the genome has been sequenced, the identity and function of genes is still widely unknown. Gene ontologies (Ashburner et al., 2000)—that is, semantic databases on known or hypothesized gene function—have been constructed to accumulate knowledge from experiments, where genes' association to biological functions and conditions have been studied. The gene ontology provides a rough

mapping between the genetic sequence and the functions of the resulting proteins.

2.3 Transcripts

Genes are *expressed* through protein synthesis. Both diseases and medical drugs, among many other factors, influence the expression of genes. To understand the molecular mechanisms of diseases and drugs, it is crucial to understand how they affect the expression.

DNA is *transcribed* into ribonucleic acid (RNA), which is the carrier of the genetic code from the nucleus of the cell to the ribosomes, where the translation into a protein takes place. Transcripts are the middle post on the path from DNA to protein, and transcription is subject to regulation, thus affecting the amount of protein produced in the cell. The expression of a gene can be indirectly quantified by measuring the amount of the corresponding transcript—the messenger-RNA molecule—in the cell. *Microarrays* have been the predominant technology for quantifying the RNA molecules, now increasingly replaced by deep-sequencing technologies (Mortazavi et al., 2008), also known as RNAseq or "next-generation sequencing."

In the complementary DNA microarray technology (Schena et al., 1995; Brown and Botstein, 1999; Duggan et al., 1999), RNA molecules present in the sample become matched to template sequences from the genome, which are positioned on a chip. The amount of each RNA molecule can then be estimated via the amount of RNA attached to each type of a template. Due to mismatching of the sequences, different affinities of the bases and other reasons, the expression data from microarrays is noisy. As a benefit, the microarrays are relatively inexpensive. RNAseq technology can be used to sequence virtually all the RNA molecules present in the sample. However, it is still challenging to infer the quantity of the longer RNA molecules based on the short sequences acquired from the device.

Large amounts of gene expression data from microarray measurements are now available in public repositories, such as the ArrayExpress (Brazma et al., 2003). The accessibility to a wide spectrum of experimental samples makes it possible for data-driven approaches to identify commonalities between diseases and treatments across studies.

2.4 Proteins

By participating in the operation of the cell at the molecular level, proteins are the main acting agents in the cell and, thus, their abundance directly influences the operation of the cell and the organism. Proteins are built from amino acids as a sequence of amino acid molecules. There are 20 different amino acids, which results in a multitude of possible protein structures.

Proteins are eventually constructed by translating the sequence of base triplets in the RNA to a sequence of amino acids. A functioning protein results, when individual amino acid molecules are linked together in the order defined by the gene and the RNA sequence.

The protein synthesis is regulated by multiple factors, such as other proteins (Jacob and Monod, 1961; Vogel and Marcotte, 2012). The proteins can be quantified based on their amino acid content (Link et al., 1999; Washburn et al., 2001). However, the gene's activity is typically quantified at the transcript-level, following from the relative ease of string matching, instead of the mass spectrometry-based quantification of the proteins (Section 5.1.1).

2.5 Metabolites

It is not possible to quantify the expression of genes in terms of the amount of transcripts or proteins in organs or tissues within the body in a non-invasive manner. This is problematic for the study of diseases and their treatments, since the changes need to be studied on those cells which are affected by the condition. For cell extracts and cell lines, the conditions can be studied under a controlled experiment in the laboratory. However, potential indicators of the disease—or, *biomarkers*—need to be measurable in a straightforward and minimally invasive way from the patient.

Because of their association with biological processes and their minimally invasive measurement potential from the blood serum, metabolites are interesting as descriptors of the biological condition (Mamas et al., 2011). Metabolites are traces and end products of biological processes in the organism (Fiehn, 2002)—processes which are ultimately described and regulated by genes and mediated by proteins encoded by the genes. As the end point of this cascade,

metabolites are descriptive of what has actually recently happened in the cells of the organism, and how the conditions such as the disease or the treatment have affected the balance of the organism.

The set of metabolite molecules is large and still partially unknown. Metabolites are heterogeneous by their molecular size and concentration level. These factors set challenges to the quantification of metabolite compounds, which is mainly done via chromatography-coupled methods of mass spectrometry. There are many challenges in the analysis of mass spectral data, such as the annotation of the compounds in the spectrum. In this thesis, new methods specifically designed for the analysis of metabolomic mass spectral data are presented in Chapter 5, where also the measurement technology is reviewed in more detail.

Lipids are a subgroup of metabolites with an active role in the cells (Shevchenko and Simons, 2010). Lipids are important for the energy metabolism of the organism, and thus relevant for the molecular balance of the cells. Lipids are the main building blocks of the cell membranes, and lipid structures also act as transporter vehicles in the blood. Due to these important roles, changes in the concentrations of the lipids in the blood or biological tissue can be informative of changes in the metabolism of the cell, potentially in connection to diseases. Lipids can be quantified with the same mass spectral methods as other metabolites.

2.6 Model Organisms

With the new measurement technologies, the response of the cell to environmental factors can be quantified at multiple stages of the cascade of the molecular response. The integration of these multiple *views* of the phenomenon is a challenging problem that can be addressed with computational models presented in this thesis.

Despite the new molecular measurement technologies, tissues deep within the body cannot be screened at the present without invasive sample-taking, if possible at all. Thus, model organisms grown in a laboratory environment are used for understanding human diseases and treatments for diseases.

A cell line grown in a culture on a Petri dish—that is, *in vitro*—is the simplest model for a biological system, since it is a collection of homogeneous cells without specialization to tissues or organs. Many

biological phenomena, such as the development of cancer or the effects of a drug, can be studied on cells grown *in vitro*. An experiment with a cell line can be controlled to a high degree, since the experimenter can ensure that the drug is administered evenly to the entire population of the cells, and the experimental conditions such as the temperature can be kept stable. However, all the effects of the drug that result from changes in the interaction and operation of cells at the tissue or organ level cannot be observed based on an experiment *in vitro*. These higher-level effects can be studied with laboratory animals grown *in vivo*. Since the animal has to be sacrificed for the inspection of internal organs, the study of the temporal development of the disease or the temporal effects of the drug is more limited for an experiment *in vivo* than *in vitro*.

Chemical compounds that are developed for therapeutic purposes may have unexpected toxic effects, which need to be thoroughly assessed before entering the test phase with human subjects. Effects of new drug compounds are, thus, tested on model organisms *in vivo* as well as *in vitro* before being approved for human tests (Waters and Fostel, 2004). The measurement of changes in the growth of a cell culture after administering the chemical compound is the most straightforward way of assessing the toxic effects of the compound. By observing the cell culture, the level of toxicity can be experimentally quantified in terms of the growth inhibition (GI) and the median lethal dose (LD_{50}) measures, which describe the amount of the compound that, when administered to the cells, results in the inhibition of the growth or the death of 50 % of the cells, respectively (Kent, 1998). When the compound is administered to model organisms grown *in vivo*, organ-level effects, especially the damage on the liver, can be assessed after the death of the animal through histopathological methods. To gain a deeper understanding of the biological processes perturbed as a consequence of the treatment, the test organisms can be studied with modern molecular measurement technologies to assess the *toxicogenomic* effects of the compound.

The effects observed in model organisms do not necessarily generalize to humans, and it is not known, which of the effects do generalize. Experimentation on multiple model organisms *in vivo* and *in vitro* can give a broader picture about the potential effects. Further, effects that are conserved across organisms, potentially generalizing to humans as well, can be identified among the effects. The question of finding effects that generalize across multiple organisms is addressed in this thesis

in Publication V, which is discussed in more detail in Section 6.3.

2.7 Conclusion

Genome, transcriptome, proteome and metabolome are interconnected by protein synthesis, metabolic processes and by regulation in these processes. None of these molecular groups alone is sufficient to describe the biological diversity and the phenotype.

The biomolecules can be measured via modern technologies but the integration of data between molecular levels and between experiments is a challenge. In this thesis, new methods are presented to address both of these problems.

3. Bayesian Latent Variable Models

3.1 Introduction

Bayesian latent variable models provide a flexible but formal way of describing assumptions about how the observed data has been produced, and what is the level of uncertainty under these assumptions. In general terms, the data $\mathbf{X}$ are explained by a statistical model through the model parameters $\boldsymbol{\theta}$. The parametrization of the model defines the model family, within which the values of the parameters can be inferred given the data.

The Bayesian approach to statistics allows one to define a prior belief on the values of the parameters. Through the prior distribution $p(\boldsymbol{\theta})$, the model can be guided towards the values of the parameters, which are realistic given prior information about the generative process of the data. The Bayesian prior is a natural way of incorporating knowledge from earlier experiments to the model. For instance, when clustering metabolites based on noisy observations of their concentrations, information about the participation of the metabolites in biological pathways may be useful prior knowledge.

In the maximum likelihood-based approach for the estimation of the model parameters, the likelihood, $p(\mathbf{X}|\boldsymbol{\theta})$, of the data is maximized to acquire the parameter values that most likely generated the data. Such an approach has been observed to be susceptible to over-fitting when the number of observations is low and the level of noise in the observations is high (see, *e.g.*, Bishop, 2006).

In the Bayesian approach, the relationship between data and parameters is inverted by focusing on the probability of the parameter values given the observed data, $p(\boldsymbol{\theta}|\mathbf{X})$, which is termed the *posterior*

probability. The inversion can be done through the Bayes' theorem (Bayes and Price, 1763),

$$p(\boldsymbol{\theta}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathbf{X})}, \tag{3.1}$$

where the probability of the data, $p(\mathbf{X})$, is constant for a given data set and depends on the model family under consideration. When comparing models within a model family, the probability of the data remains constant and the posterior probability,

$$p(\boldsymbol{\theta}|\mathbf{X}) \propto p(\mathbf{X}|\boldsymbol{\theta})\, p(\boldsymbol{\theta}), \tag{3.2}$$

can be computed as proportional to the product of the prior and the likelihood. The uncertainty in the values of the model parameters can be expressed through the Bayes' theorem and the parameters are no longer assumed to have an exact value. To underline this change of perspective, the model parameters that are assumed to be behind the data generative process are called *latent variables*.

The understanding of complex systems is often based on a limited set of observations about the behavior of the system. For instance, the molecular-level quantification of the response of a biological organism to an experimental condition is constrained by the resources for the experiment, limiting the number of individual organisms, or *biological replicates*, on which the experiment and the measurement can be done. When the sample size is small and the observations are noisy, structural assumptions about the generative process have to be made to facilitate the learning of model parameters. The Bayesian approach to probability provides a means for learning structured models, where model parameters have a hierarchy.

3.2 Hierarchical Models

The model can be learned already from a small set of noisy observations, when learning is guided with assumptions on the expected structure of the model and with assumptions on the distributions of parameters in the model. First, distributional assumptions can be naturally formulated on the parameters (Section 3.1). Second, structural assumptions can be incorporated via hierarchical modeling, or graphical modeling, which enables the *a priori* specification of dependencies between latent variables of the model. Through these dependencies, dependencies

between observed variables of the data can be identified. Third, distributional assumptions on the priors of the parameters are a practical way of incorporating additional prior knowledge about the generative process of the data.

When the latent variables are assembled into a hierarchical structure, low-level latent variables channel information from the observations to higher-level latent variables (Jordan, 2004). For instance, data consisting of observations on the concentrations of biomolecules that participate in biological processes, which then are a part of larger processes, can be modeled as a two-level cluster model. The *a priori* specified two-level hierarchical cluster structure facilitates the learning of the molecule concentrations, since the concentrations of molecules participating in the same process are likely to be mutually correlated.

Due to mutual dependencies between latent variables, the posterior probability of the model and the data typically cannot be calculated exactly. However, the posterior probability may still be inferred approximately (Section 3.7). Uncertainty in the estimates is handled via distributions of the prior and the posterior, which is important especially when the data are noisy and the number of samples is small.

When the prior distribution is *conjugate* to the likelihood distribution, the posterior distribution can be written in a closed form, making efficient posterior inference possible (Gelman et al., 2003). When the prior distributions of the latent variables are conjugate, the posterior distribution of even a hierarchical structure can be inferred via a sequential local update scheme (Section 3.7, where one variable is updated given the nodes within its *Markov blanket*. The blanket only consists of the node's daughter nodes, parent and co-parent nodes (see Bishop, 2006).

3.3 Gaussian Mixture Model

A clustering algorithm finds structure in the data without supervision. It splits the data $\mathbf{X} \in \mathbb{R}^{P \times N}$ with N samples and P variables into K clusters of samples, with cluster memberships of the samples indicated by the vector $\mathbf{v} \in \{1, \ldots, K\}^N$. When applied to multi-patient data on molecular-level changes caused by a disease, a clustering algorithm can identify subtypes of the disease without expert information about the diagnosis.

The Bayesian mixture model (see, *e.g.*, Bishop, 2006) is a density estimation method that can be used for clustering. The model is one of the fundamental building blocks for more complex hierarchical models. At its core, the model consists of a prior over the mixing distribution, $p(\mathbf{v})$, and a prior over the cluster parameters, $\boldsymbol{\theta} \sim G_0$, defined in terms of the base distribution G_0. When the base distribution is a normal distribution, the model is termed the *Gaussian mixture model*, and the sample i in cluster k is generated from a normal distribution,

$$\mathbf{x}_{\cdot,i} | (v_i = k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{3.3}$$

with cluster-specific parameters $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$: the mean $\boldsymbol{\mu}_k \in \mathbb{R}^P$ and the covariance $\boldsymbol{\Sigma}_k \in \mathbb{R}_+^{P \times P}$, which is typically further simplified to a diagonal form.

3.4 Hidden Markov Model

The Gaussian mixture model (Section 3.3) is a useful tool for clustering samples that are observed under a controlled experiment. However, medical studies conducted on patients cannot be controlled to the same extent as experiments on model organisms. One of the main challenges of the analysis of patient data is the temporal heterogeneity of the samples: patients visit the clinic at different, potentially irregular, times and their medical condition develops at an individual rate. Thus, the time series have to be *aligned* between patients to study the development of the condition.

The *hidden Markov model* (HMM; see Rabiner and Juang, 1986) is a dynamical state-space model that solves the sample alignment problem by setting samples in the time series to temporal latent states that generalize across the multiple time series. By inferring such states, the model can identify, for instance, stages in the development of the disease from heterogeneous time series data. Unlike the standard mixture model, the HMM assumes a temporal dependency between the K states through the *transition matrix*, $\mathbf{A} \in [0, 1]^{K \times K}$. The entry $a_{i,j}$ in the transition matrix determines the probability of assigning the next sample in the series to the state j, given the previous sample was assigned to the state i. Analogous to the standard mixture model, each of the K states is defined by the state parameters $\boldsymbol{\theta}_k$.

To avoid over-fitting to the data with a small set of noisy samples,

the flexibility of the HMM can be restricted by imposing prior information about the possible state transformations. For instance, for the development of the disease, a linear and forward-advancing transition structure is a reasonable assumption to simplify the model, leading to the inference of sequential states of the disease development. In addition to providing a framework for incorporating prior knowledge on the state transition probabilities, the Bayesian formulation (Gauvain and Lee, 1992) of the HMM makes the model accessible for use as a part of more complex latent variable models.

3.5 Dirichlet Process

Bayesian methods are powerful when the prior adequately captures the assumptions about the generative process of the data. However, when the assumptions about the model complexity are incorrect, the outcome is unreliable. The *non-parametric* approach to Bayesian modeling provides a way to overcome the problem of potentially wrong model complexity through the assumption of an infinite parameter space. By assuming that the observed data are generated from a subset of an infinite set of parameters, a non-parametric model automatically determines the complexity present in the observed the data, without resorting to explicit model comparison (Hjort et al., 2010). Non-parametric models have been proposed for all major statistical problems, such as regression, clustering and factorization.

In the context of clustering, infinite parameter space implies that there is an unlimited number of clusters from which the data may arise. The automatic determination of model complexity is useful for models of complex biological phenomena: when it is not known how many subtypes or developmental stages of the disease there are, a non-parametric model can automatically determine the level of specificity at which the phenomenon can be described, given the limited number of noisy observations.

A non-parametric cluster model can be constructed from the Dirichlet process, which is a probability distribution on the space of probability measures. The process induces finite-dimensional Dirichlet distributions on the data. The process can be understood through the *Chinese restaurant process* formulation: a new item may be assigned to one of the existing clusters, or a new cluster may be created for the item.

The probability of assigning the item i to an existing cluster $k = 1, \ldots, K$,

$$p\left(v_i = k | \mathbf{v}_{-i}\right) = \frac{1}{N - 1 + \alpha_{\text{DP}}} \sum_{n \neq i} \delta_{v_n, k}, \tag{3.4}$$

is proportional to the size of the cluster, $\sum_{n \neq i} \delta_{v_n, k}$, where the Kronecker delta function $\delta_{v_n, k}$ receives the value 1 when item n is in cluster k, and 0 otherwise. The probability of creating a new cluster,

$$p\left(v_i = K + 1 | \mathbf{v}_{-i}\right) = \frac{\alpha_{\text{DP}}}{N - 1 + \alpha_{\text{DP}}}, \tag{3.5}$$

is controlled by the concentration parameter α_{DP}, which can be interpreted as the number of pseudo-items outside the K clusters. Despite the sequential assignment procedure of the Chinese restaurant process, the process is exchangeable (Aldous, 1985): the posterior distribution is independent of the order in which the items are introduced.

Following from the Bayesian formulation (Ferguson, 1973), the Dirichlet process can be plugged in as the mixing distribution, $p\left(\mathbf{v}\right)$, of a standard mixture model, thus constructing the *infinite mixture model* (Escobar, 1994; Rasmussen, 2000).

3.6 Sparse Models

A mixture model learned on "small n, large p" data has considerably more parameters than the number of samples in the data. In such a situation, even Bayesian models are prone to over-fitting, unless the flexibility of the model is constrained. Structural assumptions can then support the learning of a model that is simple enough to be learned from the small number of samples but still descriptive of the generative process of the data.

Non-parametric formulation of the model is one way of finding the appropriate complexity to describe the observed data. For high-dimensional observations, however, this may not be sufficient. By introducing sparsity to the model, the effective number of parameters in the model can be constrained by setting redundant parameters to zero. The increased interpretability is a great benefit that results from sparse modeling: for instance, for the mixture model of disease subtypes, the introduction of sparsity to the mean parameter of the mixture model leads to a biclustering model, where one mixture component explains a subset of samples and variables, potentially identifying pathway-level

changes that are specific to a subtype but modeling the variation in the remaining variables as noise.

When sparsity is induced by a mechanism that treats all the observed items equally, the model learns the sparsity structure purely based on the data. Such a model can be constructed using a Bayesian prior (Tipping, 2001) or a penalty on the likelihood (Tibshirani, 1996) that treats all the items equally.

When applied to molecular-level biological data, a sparse model is not guaranteed to reveal structure that is in alignment with biological pathways which carry a biological interpretation. To improve the detection of weak signals that are visible in all or most of the molecules within a pathway, the *a priori* known relationships of the molecules can be incorporated to the model through the *group sparsity* assumption (Yuan and Lin, 2006). In this way, semantically meaningful models can be constructed by inducing sparsity on groups of variables instead of individual variables. A Bayesian model can be encouraged to follow *group sparse* structure by moving the sparsity prior higher up in the hierarchy of the latent variables.

The *automatic relevance determination prior* and the *spike-and-slab prior* were used in the works presented in this thesis, and they are introduced next. The Laplace prior and the Jeffrey's prior are other widely-used Bayesian priors, which produce a comparable outcome in terms of sparsity of the model. The ℓ_1-regularizer (Tibshirani, 1996) and its variants are the most widely-used approaches for models learned via maximum likelihood estimation.

3.6.1 Automatic Relevance Determination Prior

The *automatic relevance determination* (ARD) prior is one of the Bayesian approaches for achieving sparsity. The ARD prior assumes that the coefficients or weights of the model,

$$\mathbf{w} \sim \mathcal{N}\left(\mathbf{0}, (\boldsymbol{\alpha}\mathbf{I})^{-1}\right), \tag{3.6}$$

are independent and Gaussian-distributed with a zero mean and an item-specific gamma-distributed variance,

$$\alpha_j \sim \text{Gamma}\left(a_0, b_0\right), \tag{3.7}$$

with shape parameters a_0 and b_0, for each item j of the vector $\mathbf{w}$. When the same variance parameter is shared by a group of items, the ARD prior induces group sparsity (Virtanen et al., 2012).

The ease of inference is one of the benefits of the ARD prior. Since the items are assumed to be Gaussian-distributed and the prior distribution for their variance is conjugate to the Gaussian distribution, standard and efficient tools can be used for the inference of the model.

3.6.2 Spike-and-Slab Prior

The *spike-and-slab* prior (Mitchell and Beauchamp, 1988) is another widely-applied mechanism for inducing sparsity in Bayesian models. The prior,

$$w_j \sim (1 - p_0)\,\mathcal{N}\left(0, \sigma^2\right) + p_0\delta_0, \tag{3.8}$$

is a mixture of two components: first, the "slab," which typically is a Gaussian distribution, and second, the "spike," which is a point mass of probability density located at the origin, defined as the Dirac delta function δ_0. The width of the Gaussian distribution is defined by the variance σ^2. The parameter p_0 defines the prior ratio of the items following the "spike" versus the "slab."

3.7 Model Inference via Gibbs sampling

Since a hierarchical model is typically not within the reach of exact inference, approximation methods are used for the inference. Due to dependencies between latent variables, also the hierarchical models presented in this thesis are inferred approximately. For all the models presented in this thesis, *Gibbs sampling* was chosen as the inference method, thanks to its convenience of formulation. Two other widely-used methods for approximate inference of the posterior distribution are the variational Bayesian approximation (see, *e.g.*, Wainwright and Jordan, 2008) and the expectation propagation (Minka, 2001). A point estimate of the posterior distribution can be acquired via the expectation-maximization algorithm (Dempster et al., 1977), which is also used as the foundations of the variational Bayesian approximation.

Gibbs sampling (see Casella and George, 1992) is a method for approximate inference, where one node of the hierarchical model is updated at a time, given all the other nodes of the model. The conditional update is simplified by the conditional independence between the updated node and all the other nodes beyond its Markov blanket, meaning that the updated node's marginal distribution is dependent only

on its daughter, parent and co-parent nodes in the graphical model.

Further, if the prior distribution specified for the updated node is conjugate to the prior distributions of its parent and daughter nodes in the hierarchical model, the conditional distribution of the node can be written in a closed form and, most importantly, is of the same distributional family as the prior distribution, leading to an efficient sampling from a convenient distribution. For instance, the Gaussian prior distribution for a latent variable and a Gamma prior distribution for its inverse variance in the ARD prior (Section 3.6.1) produce a Gaussian posterior distribution for the latent variable.

If the sampler has converged after the initialization, it generates samples from the true posterior distribution of the model. Due to its sequential nature, consecutive samples from the algorithm are correlated. The correlation is reduced by *thinning* the sequence, that is, by down-sampling a subset of the samples from the sequence at constant intervals (see Gelman et al., 2003). The acquired Gibbs samples then can be considered as samples drawn from the posterior distribution of the model. The marginal distribution of a latent variable in the model can be studied through the histogram of its Gibbs samples.

3.8 Conclusion

The Bayesian approach to modeling provides a means for making assumptions about the generative process of the data. Structured assumptions and the probabilistic treatment of the data are especially helpful, when the observations are noisy and the number of samples is limited.

The dependency structure of a Bayesian latent variable model can be made to reflect the prior assumptions about the generative process of the data. Prior distributions for the latent variables guide the model, when the data are noisy and the availability of observations is poor. When the structure and distributional assumptions correctly reflect the data generative process, the model can learn structure even from noisy data.

In this thesis, Bayesian models are used to identify responses in high-dimensional molecular-level observations of biological organisms, and to integrate data from multiple measurement platforms and experiments.

4. Inference of Differences Between Groups of Samples

4.1 Introduction

Inference of differences between groups of samples is in the core of the data translation approaches presented in this thesis. Through a matching experiment design, the computational method can detect similarities between the otherwise non-matching data domains. However, the inference of covariate effects from even a single biological high-throughput data set is a non-trivial problem due to the high dimensionality of the observations and the small sample size. In this chapter, a Bayesian multi-way model for the analysis of high-dimensional experimental data (Publication I) is introduced.

At the simplest, there are two groups of samples, which are observed under two different conditions, labeled as 1 and 2. When a biological organism is in question, examples of the condition include a disease, or an intervention such as a treatment with a drug.

4.1.1 Designed Experiment

The design of the experiment (see Montgomery, 2001) determines the controlled conditions under which observations are made. The condition present in the sample is expressed as a covariate, which is a categorical variable attached to the actual observation. The covariate is sometimes termed the *independent variable*, while the actual observed variable is termed the *dependent variable*.

To make differences between sample groups more interpretable, one of the conditions is a control condition, the *normal* or the base-level group, to which the other conditions are compared. In a medical experiment, the control group may be the healthy, or the non-treated group.

A proper design of the experiment ensures that the effect of confounding factors is minimal. Then the experimenter can study the effect of the designed intervention without interference from other factors. When only one aspect of the conditions is controlled, there is one covariate attached to the observations and the experiment has a one-way design. With multiple aspects of the conditions controlled simultaneously, there is a corresponding number of co-occurring covariates and the experiment has a multi-way design. A multi-way medical experiment typically includes the disease status, a treatment with a drug, and the passed time since the treatment, all three as controlled conditions.

In a one-way experiment with two sample groups, the *t*-test (Student, 1908) is the standard statistical tool for analyzing the difference between the groups. Hotelling's T^2 test (Hotelling, 1931) generalizes the *t*-test to multivariate observations.

4.1.2 Analysis of Variance

When there are more than two sample groups in the experiment, the *t*-test can be used for analyzing pairwise differences between the groups. However, pairwise testing does not reveal interaction effects of the covariates, that is, how a sample group determined by a combination of the levels of multiple covariates differs from another sample group determined by another such combination. If the difference between each possible pair of combinations of the covariate levels is assessed with the *t*-test, the problem of multiple tests emerges and correction procedures have to be applied to the acquired *p*-values of the test.

Analysis of variance (ANOVA; Fisher, 1919) is a statistical model for observations with multiple covariates. It models both the effects of the covariates as well as the interaction effects of multiple covariates. The ANOVA method decomposes the variance in the data into variance within and between the groups. This decomposition enables the simple calculation of a test on the null hypothesis of no difference in the group means via the *F*-test.

When there are two covariates, $\mathbf{a} \in \{1, \ldots, A\}^N$ and $\mathbf{b} \in \{1, \ldots, B\}^N$, the ANOVA model for the dependent variable, $\mathbf{y} \in \mathbb{R}^N$, observed in sample i is

$$y_i = \alpha_{a_i} + \beta_{b_i} + (\alpha\beta)_{a_i,b_i} + e_i, \tag{4.1}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^A$ and $\boldsymbol{\beta} \in \mathbb{R}^B$ are the effects of the covariates a and b, respectively, and the $(\boldsymbol{\alpha\beta}) \in \mathbb{R}^{A \times B}$ are their interaction effects. The observation y_i, thus, is a linear combination of these effects with a Gaussian residual e_i.

Both ANOVA and the *t*-test are univariate statistical tests. Multivariate analysis of variance (MANOVA; see Mardia et al., 1979), generalizes the idea of ANOVA to multivariate observations, assuming that the observed variables are independent and Gaussian-distributed. For molecular-level biological observations, the independence criterion typically does not hold, since the biological molecules are interdependent through the pathway structure.

Principal component analysis (PCA)-based projection methods for MANOVA have been presented (Langsrud, 2002) to address the problem arising from a violated independence assumption by first projecting the observed data into a low-dimensional space, where the variables are orthogonal. On the other hand, PCA has been also used for projecting the covariate effects of the ANOVA model (Equation 4.1) into a lower-dimensional space determined by the principal components, for the improved interpretation of differences between multiple sample groups (Smilde et al., 2005). These methods, however, compromise the variable-level interpretability of the differences.

4.1.3 Enrichment Analysis

When studying a complex system such as a biological organism, the inferred statistical associations between the covariates and observed data are only a starting point for understanding the phenomenon. Enrichment analysis (Huang et al., 2009) provides the link between the quantified response in the variables and semantic information about the variables, giving clues about the mechanisms behind the response.

Enrichment analysis is based on the semantic categorization of the molecular-level units, such as the gene ontologies (Ashburner et al., 2000), accumulated over the decades of research on gene function. Gene ontologies are aiming at describing the gene function, chromosomal location, or regulation through simple semantic annotations.

The enrichment analysis methods identify semantic categories that deviate from the expected, for instance, gene sets where the genes are differentially expressed more often than the entire observed transcriptome on average (Subramanian et al., 2005). In a typical

approach (Subramanian et al., 2005) to study the enrichment of gene sets, the transcripts are ranked based on their differential expression, and the Kolmogorov-Smirnov test is computed on the enrichment of each of the semantic categories among the items at the top of the ranked list. The test is non-parametric and operates on the list of transcripts, thus, making no distributional assumptions about the expression data.

4.1.4 Regression Models

The search of statistical dependencies between covariates and observations can be considered as a regression problem, enabling the variety of regression models to be applied to the problem.

A regression model (see, *e.g.*, Seber and Lee, 2003) explains the dependent variable y given the independent variables $\mathbf{x}$. When the observations are arranged into the vector $\mathbf{y} \in \mathbb{R}^N$ of the dependent variable, and the matrix $\mathbf{X} \in \mathbb{R}^{N \times P}$ of the independent variables, the linear dependency,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{4.2}$$

between the dependent variable and the independent variables is modeled through the regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^P$.

The standard regression model assumes both continuous dependent and independent variables and independent Gaussian noise as the residual $\mathbf{e} \in \mathbb{R}^N$. Extensions for categorical data can find associations between the continuous dependent variables and the covariates of a designed experiment.

Mixed-effects model

A mixed-effects regression model (Laird and Ware, 1982) is a regression model,

$$\mathbf{y}^{(i)} = \mathbf{X}^{(i)}\boldsymbol{\alpha} + \mathbf{Z}^{(i)}\mathbf{b}_{\cdot,i} + \mathbf{e}^{(i)}, \tag{4.3}$$

where multiple observations of the dependent variable $\mathbf{y}^{(i)} \in \mathbb{R}^{N_i}$ for the subject, $i = 1, \ldots, S$, with N_i observations, are explained through random and fixed effects: the vector $\boldsymbol{\alpha} \in \mathbb{R}^A$ and the ith column vector of the matrix $\mathbf{B} \in \mathbb{R}^{K \times S}$, respectively. The operator "$\cdot$" in the variable $\mathbf{b}_{\cdot,i}$ implies that the entire ith column, representing the covariate effects specific to the subject i, is included from the matrix $\mathbf{B}$. The known design matrices $\mathbf{X}^{(i)} \in \mathbb{R}^{N_i \times A}$ and $\mathbf{Z}^{(i)} \in \mathbb{R}^{N_i \times K}$ determine the two types of covariates with A and K levels, respectively, for the subject i. The fixed effects, $\boldsymbol{\alpha}$, are analogous to the coefficients, $\boldsymbol{\beta}$, in the standard linear

regression model (Equation 4.2) and are shared by all subjects, or regression tasks, $i = 1, \ldots, S$. Typically, the fixed effects, or a part thereof, are the sought-for statistical result of the study.

However, the mixed-effects model is different from the standard regression model, since it also includes random effects, $\mathbf{B}$, that are task-specific and modeling responses that do not generalize across subjects. The random effects are typically considered as structured noise from the perspective of the designed experiment. As in the standard regression model, the residual, $\mathbf{e}^{(i)} \in \mathbb{R}^{N_i}$, is Gaussian noise which is independent of the known covariates.

Restricted maximum likelihood (REML) estimation method has been proposed for the inference of the mixed-effects model. The point-estimation method may lead to problems when the data are noisy or the experiment design is complex with multiple covariates.

Sparse factor regression model

The sparse factor regression model (Carvalho et al., 2008) captures low-dimensional relationships between subgroups of samples. The model,

$$\mathbf{x}_{\cdot,i} = \mathbf{B}\mathbf{h}_{\cdot,i} + \mathbf{W}\boldsymbol{\lambda}_{\cdot,i} + \mathbf{e}_{\cdot,i}, \tag{4.4}$$

decomposes the observed data into three parts: responses to the known covariates, structured variation explained by latent factors, and sample-specific variation, which is regarded as noise. High-dimensional observations $\mathbf{X} \in \mathbb{R}^{P \times N}$, for instance, the expression of P genes for N subjects, are partially explained by the covariates $\mathbf{H} \in \mathbb{R}^{A \times N}$. The regression coefficients $\mathbf{B}^{\mathrm{T}} \in \mathbb{R}^{A \times P}$ work in the same way as the fixed effects, $\boldsymbol{\alpha}$, in the mixed-effects model (Equation 4.3), modeling the influence of the known covariates on the observed data.

When more detailed group structure among the samples, for instance subtypes of the disease, is not described by the covariates, the remaining heterogeneity among the samples can be modeled with the K latent factors, $\boldsymbol{\Lambda} \in \mathbb{R}^{K \times N}$. The factors exhibit a shared activation among a subset of samples—modeling, for instance, changes in a specific biological pathway, when the changes result from an unknown mutation that has occurred in a subset of the samples. The substructure can be inferred via the Dirichlet process prior (Section 3.5). With sparsity in the factor loadings, $\mathbf{W} \in \mathbb{R}^{P \times K}$, the model identifies bicluster structure in the data that is not explained by the known covariates.

Regression models can be used for finding statistical dependencies

between the observed variables and the covariates. However, the standard linear regression model does not account for interaction effects of two or more covariates, but the interactions can be included in the model by adding redundant variables that describe combinations of the covariates. In the ANOVA model, the decomposition into main effects and interaction effects comes naturally (Equation 4.1) but an independent model for each variable becomes unreliable when the sample size is small and the number of variables is high. By making structural assumptions about the generative process of the data, the covariate effects can be learned also in the "small n, large p" regime. A Bayesian multi-way model for such a setting is presented next.

4.2 Bayesian Multi-Way Model

Clustering is a simple starting point to modeling correlated groups of variables without the need for additional information about their similarity (Publication I). In a factor model for the observed variables,

$$\mathbf{x}_{\cdot,i} = \mathbf{V}\mathbf{x}^{\text{lat}}_{\cdot,i} + \mathbf{e}_{\cdot,i}, \tag{4.5}$$

the P observed variables are assigned into clusters by the clustering matrix $\mathbf{V} \in \{0,1\}^{P \times K}$ and the K variable clusters for the sample i are represented by the latent variable $\mathbf{x}^{\text{lat}}_{\cdot,i} \in \mathbb{R}^K$.

By assuming that the members of a cluster respond coherently to the experimental covariates, the effects of the covariates can be inferred on the K-dimensional latent representation $\mathbf{x}^{\text{lat}}_{\cdot,i}$ instead of the full dimensionality P of the observed data. The generative model,

$$\mathbf{x}^{\text{lat}}_{\cdot,i} \sim \mathcal{N}\left(\boldsymbol{\alpha}_{\cdot,a_i} + \boldsymbol{\beta}_{\cdot,b_i} + (\boldsymbol{\alpha}\boldsymbol{\beta})_{\cdot,a_i,b_i}, \mathbf{I}\right), \tag{4.6}$$

is analogous to the ANOVA model (Section 4.1.2) but the effects are inferred for each of the K clusters and are independent and Gaussian-distributed,

$$\begin{aligned} \boldsymbol{\alpha}_{k,c_a} &\sim \mathcal{N}(0,1), \\ \boldsymbol{\beta}_{k,c_b} &\sim \mathcal{N}(0,1), \\ (\boldsymbol{\alpha}\boldsymbol{\beta})_{k,c_a,c_b} &\sim \mathcal{N}(0,1), \end{aligned} \tag{4.7}$$

for all the clusters, $k = 1, \ldots, K$, and the levels of the covariates, $c_a = 2, \ldots, A$ and $c_b = 2, \ldots, B$, except for the base-levels, $c_a = 1$ and $c_b = 1$, for which all the effects are set to zero.

4.3 Conclusion

The influence of a covariate on a system can be studied through a controlled experiment, where other factors that potentially influence the system are kept unchanged. The ANOVA model, which is the basic tool for the analysis of data from a controlled experiment, decomposes the observed data into effects of the known covariates.

In the “small n, large p" regime, statistical models are prone to over-fitting. These problems can be avoided, while still allowing the model to learn weak patterns from the data, by introducing structural assumptions about the generative process of the data, and in this way guiding the model. For high-dimensional data from molecular measurements of biological organisms, the clustering assumption of collinear variables is a meaningful way of restricting model complexity while still acquiring interpretable covariate effects in the same way as in the standard ANOVA model.

5. Multi-Peak Models for Metabolomics

5.1 Introduction

One of the main problems in understanding the metabolome of a biological organism is the quantification of the levels of metabolites and the inference of differences in these levels between sample groups. For the expression of genes, such a measurement is possible by matching messenger-RNA sequences to known templates from the genes via microarray technologies (Schena et al., 1995; Brown and Botstein, 1999; Duggan et al., 1999), or by identifying the abundance of messenger-RNA sequences directly by sequencing using the RNAseq technologies (Mortazavi et al., 2008), giving a link between the sequence and its abundance.

For metabolites, identification needs to be done via the spectral decomposition of the sample, typically by using a *chromatography-coupled mass spectrometer*. Each chemical compound in the sample produces a unique set of peaks to the mass spectrum. The identity and the abundance of the compound can be inferred from the locations and the heights of the peaks, respectively. Traditionally the inference has been based on the strongest peak of the compound, termed the "main peak." However, the peaks are a noisy representation of the sample and the inference of covariate effects is unreliable due to the "small n, large p" problem.

Understanding the complex measurement process—that is, the true generative process of the data—is essential to constructing powerful models for noisy metabolomic data. In this chapter, a new approach is presented for the inference of covariate effects. The new model integrates data from multiple peaks that can be associated with a compound.

5.1.1 Chromatography-coupled mass spectrometry

The quantification of metabolite concentrations from the blood serum gives a minimally invasive proxy to identifying potential perturbations in the molecular balance of a complex biological organism. Since many metabolites are end or side products of the metabolism of the cells, changes in the regulatory processes of the metabolism are reflected in metabolite concentrations. Following from the great complexity of the cell metabolism, the set of chemical compounds that are classified as metabolites is large and still partially unknown. This diversity sets high standards for the measurement technology, which in an unbiased way needs to quantify a wide range of molecules with varying size, polarity and other chemical properties.

Chromatography-coupled mass spectrometry is a measurement technology that enables the simultaneous quantification of a large number of chemical compounds in a sample (Dunn et al., 2011). Thus, it is the most widely used tool for the quantification of metabolites and other small molecules in a biological sample (Wilson et al., 2005). However, since the method is based on a spectral decomposition of the sample, multiple pre-processing steps are required before the data can be analyzed for changes in the compound concentrations. The mechanism behind the decomposition, thus, is important for the analysis of the concentrations.

The sample in liquid or gas form first enters the chromatograph, where the compounds are separated by the time it takes for them to pass through the capillary of the chromatograph (Snyder et al., 2010). The pass-through time, termed the *retention time* (RT), is dependent on the chemical properties of the molecule, such as the polarity. The retention time from the chromatograph is the first dimension of separation between the compounds. However, the retention time separation is not perfect and it is not accurate enough to generalize across experiments and devices to enable the annotation of the compounds.

To acquire another dimension of separation for the compounds in the sample, the partially-separated compounds are measured with the mass spectrometer. After exiting the chromatograph, the sample is ionized using electro-spray, shot through a magnetic field and eventually detected with sensors. The mass spectrometer separates the ions by their

mass-to-charge ratio (m/z), since the ion is deviated from its original trajectory by the magnetic field and the deviation is inversely proportional to the mass-to-charge ratio of the compound. From the signal detected by the sensors in different positions, a mass-to-charge spectrum can be constructed for each retention time point. This way, compounds with the same retention time are separated indirectly by their mass.

However, there is a further complication that follows from the ionization process (de Hoffmann, 2005): First, a compound can be ionized in multiple ways, resulting in multiple observations of the compound, termed *adduct* peaks, each with a unique mass-to-charge ratio at the same retention time point. Second, atomic isotopes also result in multiple observations of the compound, termed *isotope* peaks. Also the isotope peaks of a compound appear at the same retention time point.

As an output from the coupled chromatograph and mass spectrometer, the biological sample is decomposed into a two-dimensional intensity spectrum, where the first dimension is the retention time and the second dimension is the mass-to-charge ratio. Each chemical compound in the sample produces an unknown set of intensity peaks to the spectrum and the peaks may overlap with peaks from other compounds. Prior knowledge from experiments with pure compounds is required for the identification of the chemical source of the peak, termed the *annotation* of the peaks, and further pre-processing is required for the identification of changes in the concentrations of the compounds between experimental samples.

5.1.2 Pre-Processing of the Spectral Data

The output data from the chromatograph and the mass spectrometer lack the *identification* of the peaks in the continuous spectrum, the *alignment* of the identified peaks between the experimental samples, the *summarization* of the continuous intensity peaks as scalar values that describe the concentration of the compounds in the samples, and the *annotation* of peaks to the compounds they are produced by. Further, the intensity values need to be *normalized* to remove any systematic bias related to the position of the peak or the time of the measurement of the sample. All these tasks are challenging and the error made in any of the tasks adds up to the uncertainty in the data. This results in noise

at multiple levels of the generative process of the data.

With the current tools the identification, alignment, summarization, annotation and normalization steps are done sequentially. Algorithms typically used for completing these steps have been collected into tool packages that provide the entire pipeline. Examples of widely-used pipeline packages are the MZmine (Pluskal et al., 2010) and the XCMS (Smith et al., 2006).

5.1.3 Analysis of Compound Concentrations

After the measurement and pre-processing, it is still unknown how the concentrations of the chemical compounds vary between the experimental samples, and most importantly, how the experimental covariates affect the concentrations.

Standard statistical methods for the inference of differences between groups of samples, such as the ANOVA model (Chapter 4), reveal covariate effects in the data, and are widely used also for the analysis of metabolomic data. In addition, multivariate regression or classification methods, such as partial least squares (Wold et al., 2001), are used to identify multivariate statistical associations between the observed metabolome and the experimental covariates.

Since the data are noisy, resulting from biological variation among the experiment subjects, noisy measurement technology and uncertainty in the pre-processing steps, standard statistical methods may fail at the task of identifying the covariate effects. By accounting for the collinearity of the compound concentrations in the data, a probabilistic approach to modeling groups of compounds infers interpretable covariate effects even from "small n, large p" data (Publication I). Moreover, by modeling the generative process of the intensity data from the chromatography-coupled mass spectrometer, the inference of covariate effects can be further improved.

5.2 Model for Multiple Peaks from One Compound

The single-peak analysis discards data from the adduct and isotopic peaks. The ionization process produces between-sample noise in the peak heights and all the peaks are affected by the noisy process. However, the natural distribution of the atomic isotopes is known and

constant across samples. Since there is no bias in the ionization on the isotope peaks, the covariate effects are argued to be preserved well in the isotope peaks. Integrated modeling of the multiple peaks from the same chemical compound could then improve the inference of compound-wise covariate effects from the data.

The PeakANOVA model (Publication II) is an approach for clustering peaks into latent compounds and for inferring covariate effects in the data using all available peaks. Since the mass spectral peaks from one compound appear at an identical retention time, it turns out to be possible to cluster them based on the similarity in their peak shapes in the retention time dimension.

To account for the unknown number of compounds in the experimental sample and for the unknown and varying number of spectral peaks associated with a compound, a Dirichlet process prior (Section 3.5) is assumed for the clustering model. The clusters in the inferred model then correspond to latent compounds and the peaks from one compound are assumed to respond to the covariates in an identical way. This way, the covariate effects can be inferred based on multiple peaks instead of a single peak per sample and compound.

5.3 Model for Correlated Compounds with Multiple Peaks

Even with multiple peaks integrated, the "small n, large p" problem remains, since the sample size is limited compared to the number of compounds present in a sample. To address this problem, the compound clusters inferred by the PeakANOVA model (Publication II) can be further clustered into latent groups of coherently-responding compounds.

Another level of model hierarchy is introduced in the two-level PeakANOVA model (Publication III). Compounds that respond to the experimental covariates in a coherent way are assumed to follow the same generative process. The second level of model hierarchy is shown to further improve the accuracy of inference of the covariate effects.

5.4 Conclusion

Metabolites are traces and end products of biological processes. The metabolite concentrations can be quantified from a blood sample, thus providing a window to changes in the biological processes of the organism.

However, current technologies for quantifying the metabolome add noise to the observations at multiple levels, making the inference of covariate effects a non-trivial problem. It was shown that learning from the data can be improved through probabilistic modeling that accounts for the specific generative process of the observations. Most importantly, multiple peaks from the mass spectrometer device can be integrated to make the inference of covariate effects more accurate.

6. Cross-Domain Data Translation with Co-Occurring Samples

6.1 Introduction

The existence of dependencies between the data domains is a fundamental requirement for the data translation between them. In this thesis, it is shown that two main types of dependencies can be discovered: First, methods for identifying dependencies that follow from co-occurring samples are presented in this chapter. Second, methods for identifying dependencies that follow only from a shared experiment design, which is considered a more complex problem to solve, are presented in Chapter 7.

Unsupervised multi-view learning methods identify statistical dependencies between data sets with co-occurring samples but with observations from different data spaces, or domains. These methods, such as *canonical correlation analysis* (CCA; Hotelling, 1936), find general dependencies between the data views. CCA has been successfully applied to problems in computational biology, for instance, to finding dependencies between the genome and the transcriptome (Lahti et al., 2009). However, further processing of the result is needed for gaining understanding of the relationship between the dependencies and the experimental covariates.

The CCA model identifies linear multivariate dependencies between two data sets, $\mathbf{X} \in \mathbb{R}^{P^x \times N}$ and $\mathbf{Y} \in \mathbb{R}^{P^y \times N}$, with N co-occurring samples. CCA finds a linear combination of the original variables in each of the two views. The linear combinations are selected via the generalized singular value decomposition in such a way that the correlation between the two linear combinations is maximized. Due to its great flexibility, the model breaks down in the "small n, large p" regime, unless the model

is kernelized (Hardoon et al., 2004), or regularized with a penalty on the likelihood or with Bayesian priors (Klami et al., 2013).

Generative formulation of CCA assumes that the observations for sample i in the two data sets $\mathbf{X}$ and $\mathbf{Y}$ are generated by a shared normally-distributed latent variable,

$$\mathbf{z}_{\cdot,i} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{6.1}$$

through linear projections,

$$\begin{aligned} \mathbf{x}_{\cdot,i} &= \mathbf{W}^{\mathbf{x}}\mathbf{z}_{\cdot,i} + \mathbf{e}^{\mathbf{x}}_{\cdot,i}, \\ \mathbf{y}_{\cdot,i} &= \mathbf{W}^{\mathbf{y}}\mathbf{z}_{\cdot,i} + \mathbf{e}^{\mathbf{y}}_{\cdot,i}, \end{aligned} \tag{6.2}$$

where the projection matrices, $\mathbf{W}^{\mathbf{x}} \in \mathbb{R}^{P^{\mathbf{x}} \times K}$ and $\mathbf{W}^{\mathbf{y}} \in \mathbb{R}^{P^{\mathbf{y}} \times K}$, project the latent variable, $\mathbf{Z} \in \mathbb{R}^{K \times N}$, from the K-dimensional representation onto the $P^{\mathbf{x}}$ and $P^{\mathbf{y}}$ dimensions of the observed data, respectively. The residuals, $\mathbf{E}^{\mathbf{x}} \in \mathbb{R}^{P^{\mathbf{x}} \times N}$ and $\mathbf{E}^{\mathbf{y}} \in \mathbb{R}^{P^{\mathbf{y}} \times N}$, describe view-specific signals and noise.

6.2 Multi-Way Model for Multiple Data Sources

CCA is a model for linear dependencies between co-occurring data sets but it does not explicitly model covariate effects, which typically are the most descriptive statistics for summarizing a designed experiment. On the other hand, standard ANOVA-type methods that model covariate effects do not reveal what there is in common between multiple data views. However, inference of covariate effects that generalize across multiple views of the data is central to understanding changes in the complex cascade of molecular responses from the genome to the metabolome as well as to understanding relationships of the responses between different tissues of the organism.

As demonstrated in Publications I–III, an ANOVA-type population prior can be incorporated as a part of a more complex generative model, allowing the high-dimensional data to be described in terms of the covariates. Combining the idea of covariate effects as the generative source of clusters of variables (Section 4.2) with the assumption of linear dependencies between data views (Section 6.1) results in a multi-way and multi-view model (Publication IV), which infers covariate effects that generalize across the views. The model decomposes the data into three parts: variation that is explained by the covariates of the experiment

either (1) generalizing across the views or (2) specific to one view, or finally, (3) variation that is not associated with the covariates.

The model is a step towards the integration of data from the multiple levels of the molecular response of a biological organism. It can be used to integrate the cascade of responses from the genome to the metabolome as well as synchronous responses from multiple tissues or organs.

6.3 Group Factor Analysis for Cross-Organism Toxicogenomics

With modern molecular measurement technologies, the effects of a drug can be studied at the molecular level with model organisms before entering the test phase on humans (Section 2.6). Due to the potentially dangerous effects of the drug and the invasive sample taking from the internal organs, the same experiment is not possible with human subjects. When the drug is experimented with multiple model organisms, the measurements provide multiple views of the drug's effects. Even though experiments with model organisms reveal many therapeutic and toxic effects of the drug, all effects do not appear on all organisms, and it is not known which of the observed effects generalize to humans (Boverhof and Zacharewski, 2006). Effects that generalize across multiple model organisms have the potential of being conserved in humans as well. However, the identification of conserved effects from multi-view data is not a trivial problem.

Group factor analysis (GFA; Virtanen et al., 2012) generalizes the Bayesian CCA model (Section 6.1) to more than two views, enabling the discovery of statistical dependencies between multiple data views. When applied to multi-organism drug response data, GFA can identify responses that are conserved across organisms.

In the spirit of a standard factor model, the factor loadings of the GFA model describe the associations between the factors and the observed variables. However, with M co-occurring data views, $\mathbf{X}^{(m)} \in \mathbb{R}^{P_m \times N}$, $m = 1, \ldots, M$, there are also M distinct *loadings* matrices, $\mathbf{W}^{(m)} \in \mathbb{R}^{P_m \times K}$, that define the linear relationship,

$$\mathbf{x}_{\cdot,i}^{(m)} = \mathbf{W}^{(m)} \mathbf{z}_{\cdot,i} + \mathbf{e}_{\cdot,i}^{(m)}, \tag{6.3}$$

from the global *factors*, $\mathbf{Z} \in \mathbb{R}^{K \times N}$, to the P_m variables in each of the data views. The activity of the factors in the data views is determined by the group sparsity prior (Section 3.6.1), which allows a factor to be active

either in all the views, in a subset of the views, or in only one of the views. When applied to multi-organism data on molecular drug responses, factors active in multiple organisms then describe responses that are conserved across the organisms. However, all effects of a drug do not generalize across all organisms, and are explained by the remaining factors.

Both CCA and GFA assume co-occurring samples. The assumption establishes a common ground between the data views and, thus, allows the discovery of between-dataset dependencies. Co-occurrence is not always directly available, for instance, when the data views are measurements from different types of organisms. In Publication V, it was shown that even then the co-occurrence can be constructed by summarizing one sample group as a single sample that can be matched to the other data views through the shared covariates. For instance, biological replicates from one type of an organism can be summarized as a single sample when they have received the same drug treatment. This sample then is matchable to the summary sample of another organism that has been experimented with the same drug. An approach to constructing the common ground based on the covariates without sample summarization is discussed in Chapter 7.

A chemical compound may disrupt the operation of a biological regulatory process by binding to a protein that is participating in the process (Iorio et al., 2010). Such a disruption leads to changes in the biological pathway, which can be observed as a change in the expression of genes in the pathway. Drugs disrupting the same pathway may affect the expression of the same set of genes, leading to a bicluster structure in the effects of the drugs on the expression of genes.

Neither the standard factor model nor GFA account for the bicluster structure. However, bicluster structure can be achieved for a factor model by introducing element-wise sparsity priors on factors and factor loadings (Hochreiter et al., 2010), allowing the factors and factor loadings to operate on a subset of samples and variables, respectively.

The biclustering model is powerful for describing effect structure with multiple drugs affecting pathways consisting of multiple genes. Still, the standard biclustering models learn the structure within one data view, leaving open whether the effects appear in other organisms. With element-wise sparsity structure introduced to factors and factor loadings,

GFA turns out to generalize the additive biclustering model to multiple data views (Publication V). This generalization then reveals drug effects that are conserved across multiple organisms. Such effects have potential of appearing in humans as well.

When a new drug compound is tested on a model organism, it is unknown to which extent the effects generalize to humans. Matching data on the effects of the drug on humans is available only from experiments *in vitro* because of the potentially dangerous effects of the drug on the individual. When the dependencies between the data views are too weak for prediction, robust cross-view factors can still be used for retrieving similar drug interventions from a database. In Publication V, it was shown that factors that generalize across model organisms on the molecular level are powerful for retrieving drug compounds that have a similar toxic or therapeutic effect on humans at the population level. This way, the retrieval of similar experiments is useful for making a hypothesis about the effects of a new experiment.

6.4 Conclusion

Modern molecular measurement technologies enable the subjects of a biological experiment be to observed at multiple views. Computational methods are needed for identifying statistical dependencies between the data views. When the data come from a controlled experiment, the analysis typically focuses on identifying effects of the experimental covariates. In this thesis, it was shown that covariate effects that generalize across multiple data views can be learned by constructing a generative model that assumes shared covariate effects.

Further, it was shown that a model assuming co-occurring samples is applicable even to data from multiple organisms, when the organisms share the same experiment design. When each sample corresponds to the response to a unique treatment, a CCA-type model can be used to identify response patterns that generalize across multiple organisms.

In order to apply CCA-type methods on multi-organism data, the co-occurrence of the samples needs to be constructed, for instance through the summarization of a sample group as a single sample. However, sample summarization unavoidably leads to loss of information. Methods that do not require summarization for multi-organism data, are presented in the next chapter.

7. Cross-Domain Data Translation without Co-Occurring Samples

7.1 Introduction

Experiments on different organisms are not made on the same subjects, that is, the samples are not naturally co-occurring. The co-occurrence between the data sets from different types of organisms can still be constructed, if the experiments have been designed to match through matching covariates. However, then the replicates in each sample group have to be summarized as a single sample. Without the co-occurrence of samples, CCA-based methods are inapplicable to the problem of finding dependencies between the data sets. In this chapter, it is shown that even when the samples are not co-occurring, it is possible to find statistical dependencies between the data sets, if they share a similar experiment design.

Methods that summarize an experiment via descriptive semantic categories have been used for finding similar experiments both between biological conditions (Schmid et al., 2012) and treatments (Lamb et al., 2006), as well as between biological organisms (Lu et al., 2009). This type of expression meta-analysis methods are useful for matching similar experiments done on different organisms (Wise et al., 2012). However, they do not provide a means for computationally translating the outcome of a new experiment between the organisms.

If at least some of the variables are assumed matched between the data sets, known variable pairs can be used to match larger clusters of variables between the data sets (Mi et al., 2010) and latent factors learned from the data of one organism can be used as a starting point for the analysis of the data from another organism (Lucas et al., 2009). Methods assuming a variable-level matching are vulnerable to errors in

the matching, for instance, when the matching of genes is based solely on sequence similarity. On the other hand, the matching of variables can be inferred from expression data by using sequence similarity as a prior for the matching (Le and Bar-Joseph, 2010).

Another approach to finding the common ground between non-co-occurring data sets is to infer the matching of samples between the data sets (Gholami and Fellenberg, 2010; Tripathi et al., 2011; Klami, 2012). With the inferred matching, CCA-type methods become applicable for finding dependencies between the data sets. However, for experiments from multiple biological species, a one-to-one matching between the samples may not be reasonable, unless the samples are first summarized among each sample group (Publication V).

7.2 Model for Dynamical Responses Across Domains

In Publication VI, it is shown that even without co-occurring samples, the common ground between the data sets can be established through the shared experiment design. In addition to the problem of non-co-occurring samples, in many cases there is no one-to-one matching between the observed variables of the data sets and, further, it is not known, which of the variables share the same covariate effect. For instance, when the data sets are measurements of the metabolome in two types of organisms, it is not known whether a metabolite has the same biological function in both the organisms and whether the effects of the experimental covariates on the metabolite are the same across the organisms. In Publication VI, these two problems are solved by building on the idea of inferring the covariate effects on groups of variables (Section 4.2; Publication I): it is assumed that for a group of variables there is a matching group in the other data set.

In spite of an experiment design with matching experimental covariates, the different life span and metabolism of different organisms set a challenge to the data translation: even if the effect of an experimental intervention is similar between the organisms, the temporal delay and temporal span of the effect may be variant. In Publication VI, it is shown that by aligning the time points of the two data sets dynamically to follow a trajectory with shared temporal covariate effects but with different dynamics, the cross-species multi-way model can identify temporal development shared between

the non-co-occurring data sets.

The matching of the time covariate between the data sets is done by assuming a hidden Markov model (HMM; Section 3.4) structure on the time effect. The HMM is incorporated as a part of the generative model of the observations, enabling the simultaneous alignment of the time points to latent states and the inference of covariate effects associated with these states.

7.3 Model for Shared and Domain-Specific Responses

All responses to the covariates are not necessarily shared between the organisms. Since the dynamical model of shared multi-way effects (Publication VI; Section 7.2) always assumes a matching pair between the data domains, it may be prone to false findings when the covariate effect is present only in one of the domains but no better-matching pair is available.

In Publication VII, a split-merge step for matching the clusters between the data sets was proposed for identifying covariate effect patterns that generalize across the data sets and for separating them from patterns specific to one data set. Through a Metropolis sampling step, the model determines whether the covariate effect is more likely to be shared than a random, "average," covariate effect from the model.

7.4 Conclusion

In Publications VI and VII, it was shown that even without co-occurring samples or variables, graphical modeling techniques can be used to find the common ground between data sets by modeling covariate effects that are shared by them. Time series of different lengths can be aligned with dynamical modeling, integrated to the graphical multi-way model.

The next challenge in cross-species modeling is the computational translation of a new experimental intervention from a model organism to humans, when no observations are available on the intervention on humans *a priori*. The new and existing experiments can then be modeled together to predict the expected outcome of the unobserved new experiment on humans in terms of similar existing experiments (Publication V; Socher et al., 2013).

8. Discussion

Data translation of experimental outcomes between organisms is one of the most important unsolved problems in computational biology, since it enables the prediction of the outcome of a new treatment in humans based on a controlled experiment on model organisms. Multi-way models, presented in this thesis, identify effects of experimental covariates that generalize across organisms.

In Publication I, a Bayesian multi-way model was presented for inferring covariate effects from high-dimensional data when the variables are collinear. The main focus of the work was at inferring the magnitude of the covariate effect. As with other approaches, assessing the statistical significance of the effect is not a trivial task. Thus, avenues for further research include a sparse model for the covariate effects to decrease the occurrence of false positive findings. Additionally, the incorporation of prior information about the similarity of the variables may improve the model when the observed variables are correlated, for instance, as a result of pathway structure in the data from a biological organism.

In Publications II and III, the Bayesian multi-way model was extended to account for the special nature of mass spectral metabolomic data. By first clustering peaks based on their shape similarity, it was shown that the use of multiple isotope and adduct peaks from a single molecular source improves the inference of covariate effects on metabolite compounds. Models for mass spectral data can still be developed further in many ways by structuring the model to describe more details of the generative process of the mass spectrometer device. For instance, some weak peaks may be too noisy for the reliable inference of covariate effects. The strength and the type of the peak could be taken into account when assessing the reliability of the peak. This could be

built into the model through a prior on the peak-specific variance parameter. On the other hand, the known positions of isotope peaks and their relative heights could be incorporated to the model through a prior to improve both the clustering of same-source peaks as well as the inference of covariate effects.

In Publication IV, a model was presented for the inference of covariate effects that generalize to multiple views. The model was shown to learn whether the effect is shared by the data views or whether it is specific to one view. However, the inference of multiple components representing different subsets of observed variables, in the same way as in Publications I–III, remains an open modeling problem. Further, for some applications it may be necessary to formulate a more interpretable connection between observed variables and covariate effects.

In Publication V, a multi-view model was applied to finding drug responses that are conserved across organisms. Since a direct prediction of the drug response based on model organisms was not feasible due to the small amount of available data compared to the complexity of the problem, a cross-organism retrieval approach was proposed for generalizing the drug response from model organisms to humans. When more data becomes available, the model may become useful for the direct prediction of the response as well.

In Publications VI and VII, a model was introduced to the data translation problem in the situation, where the data sets do not have co-occurring samples. Unlike in Publication V with summarized samples, standard multi-view models are not applicable to the problem with no co-occurring samples. However, it was shown that it is possible to identify statistical dependencies even in this situation, if the experiment design is similar between the data sets. Groups of variables were matched between the data sets based on the similarity in their covariate effects.

It is assumed in the model of Publications VI and VII that all covariate effects are similar between the matched groups of variables, which is a limitation in some applications. In future work, groups of variables could be matched based on the covariate of interest while allowing other less relevant covariates to have different effects on different organisms. The prediction of the effect of a new covariate in one organism based on an experiment on another organism—that is, the actual cross-species data translation—remains a challenge.

The matching of individual constituents of the metabolic pathways between organisms is a future challenge for computational modeling. Such a matched network could enable the translation of the complex cascade of responses and its mechanism from a model organism to humans.

With increasingly large databases of intervention experiments on numerous model organisms openly available, data-driven machine learning methods will become increasingly useful for both understanding diseases and for drug development. Another growing area of research in computational biology—that is already happening (Costello et al., in press)—is *personalized medicine* with the aim of making computer-based decisions on disease diagnosis, prognosis and treatment (Chin et al., 2011). Data translation methods are necessary tools for personalized medicine as well, since molecular measurement devices may be different between different hospitals. Further, the underlying databases may consist of both historical patient data and experimental data from model organism studies.

Bibliography

David J. Aldous. Exchangeability and related topics. In P.-L. Hennequin, editor, *École d'Été de Probabilités de Saint-Flour XIII — 1983*, volume 1117 of *Lecture Notes in Mathematics*, pages 1–198. Springer, Berlin/Heidelberg, Germany, 1985. URL http://dx.doi.org/10.1007/BFb0099421.

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Harris Midori A. Eppig, Janan T., David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000. URL http://dx.doi.org/10.1038/75556.

Mr. Bayes and Mr. Price. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions*, 53:370–418, 1763. URL http://rstl.royalsocietypublishing.org/content/53/370.short.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, New York, 2006. ISBN 0-387-31073-8.

Darrell R. Boverhof and Timothy R. Zacharewski. Toxicogenomics in risk assessment: Applications and needs. *Toxicological Sciences*, 89(2):352–360, 2006. URL http://dx.doi.org/10.1093/toxsci/kfj018.

Alvis Brazma, Helen Parkinson, Ugis Sarkans, Mohammadreza Shojatalab, Jaak Vilo, Niran Abeygunawardena, Ele Holloway, Misha Kapushesky, Patrick Kemmeren, Gonzalo Garcia Lara, Ahmet Oezcimen, Philippe Rocca-Serra, and Susanna-Assunta Sansone. Arrayexpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, 31(1):68–71, 2003. URL http://nar.oxfordjournals.org/content/31/1/68.abstract.

Patrick O. Brown and David Botstein. Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, 21:33–37, 1999. URL http://www.nature.com/ng/journal/v21/n1s/full/ng0199supp_33.html.

Kenneth H. Buetow, Michael Edmonson, Richard MacDonald, Robert Clifford, Ping Yip, Jenny Kelley, Daniel P. Little, Robert Strausberg, Hubert Koester, Charles R. Cantor, and Andreas Braun. High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Proceedings of*

the National Academy of Sciences, 98(2):581–584, 2001. URL http://www.pnas.org/content/98/2/581.abstract.

Paul R. Burton et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007. URL http://dx.doi.org/10.1038/nature05911.

Carlos M. Carvalho, Jeffrey Chang, Joseph E. Lucas, Joseph R. Nevins, Quanli Wang, and Mike West. High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484), 2008. URL http://dx.doi.org/10.1198%2F016214508000000869.

George Casella and Edward I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992. URL http://www.tandfonline.com/doi/abs/10.1080/00031305.1992.10475878.

Lynda Chin, Jannik N. Andersen, and P. Andrew Futreal. Cancer genomics: from discovery science to personalized medicine. *Nature Medicine*, 17(3):297–303, 2011. URL http://dx.doi.org/10.1038/nm.2323.

James C. Costello et al. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology*, in press. URL http://dx.doi.org/10.1038/nbt.2877.

Francis H.C. Crick. The origin of the genetic code. *Journal of Molecular Biology*, 38(3):367–379, 1968. URL http://dx.doi.org/10.1016/0022-2836(68)90392-6.

Edmond de Hoffmann. Mass Spectrometry. In A. Seidel, editor, *Kirk-Othmer Encyclopedia of Chemical Technology*. John Wiley & Sons, 2005. URL http://dx.doi.org/10.1002/0471238961.1301191913151518.a01.pub2.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977. URL http://www.jstor.org/stable/2984875.

David J. Duggan, Michael Bittner, Yidong Chen, Paul Meltzer, and Jeffrey M. Trent. Expression profiling using cDNA microarrays. *Nature Genetics*, 21:10–14, 1999. URL http://www.nature.com/ng/journal/v21/n1s/full/ng0199supp_10.html.

Warwick B. Dunn et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature Protocols*, 6(7):1060–1083, 2011. URL http://dx.doi.org/10.1038/nprot.2011.335.

Michael D. Escobar. Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association*, 89(425):268–277, 1994. URL http://www.jstor.org/stable/2291223.

Thomas S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230, 1973. URL http://www.jstor.org/stable/2958008.

Oliver Fiehn. Metabolomics—the link between genotypes and phenotypes. *Plant Molecular Biology*, 48(1–2):155–171, 2002. URL http://dx.doi.org/10.1023/A%3A1013713905833.

Ronald A. Fisher. XV.—The correlation between relatives on the supposition of Mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 52(2):399–433, 1919. URL http://dx.doi.org/10.1017/S0080456800012163.

Jean-Luc Gauvain and Chin-Hui Lee. Bayesian learning for hidden Markov model with Gaussian mixture state observation densities. *Speech Communication*, 11(2-–3):205–213, 1992. URL http://dx.doi.org/10.1016/0167-6393(92)90015-Y.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Texts in Statistical Science. Chapman & Hall/CRC, Boca Raton, Florida, 2nd edition, 2003. ISBN 1-58488-388-X.

Amin M. Gholami and Kurt Fellenberg. Cross-species common regulatory network inference without requirement for prior gene affiliation. *Bioinformatics*, 26(8):1082–1090, 2010. URL http://bioinformatics.oxfordjournals.org/content/26/8/1082.abstract.

David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004. URL http://dx.doi.org/10.1162/0899766042321814.

Nils L. Hjort, Chris Holmes, Peter Müller, and Stephen G. Walker, editors. *Bayesian nonparametrics*. Number 28 in Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, United Kingdom, 2010. ISBN 978-0-521-51346-3.

Sepp Hochreiter, Ulrich Bodenhofer, Martin Heusel, Andreas Mayr, Andreas Mitterecker, Adetayo Kasim, Tatsiana Khamiakova, Suzy Van Sanden, Dan Lin, Willem Talloen, Luc Bijnens, Hinrich W.H. Göhlmann, Ziv Shkedy, and Djork-Arné Clevert. FABIA: factor analysis for bicluster acquisition. *Bioinformatics*, 26(12):1520–1527, 2010. URL http://bioinformatics.oxfordjournals.org/content/26/12/1520.abstract.

Harold Hotelling. The generalization of Student's ratio. *The Annals of Mathematical Statistics*, 2(3):360–378, 1931. URL http://projecteuclid.org/euclid.aoms/1177732979.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936. URL http://www.jstor.org/stable/2333955.

Da Wei Huang, Brad T. Sherman, and Richard A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, 2009. URL http://nar.oxfordjournals.org/content/37/1/1.abstract.

Francesco Iorio, Roberta Bosotti, Emanuela Scacheri, Vincenzo Belcastro, Pratibha Mithbaokar, Rosa Ferriero, Loredana Murino, Roberto Tagliaferri, Nicola Brunetti-Pierri, Antonella Isacchi, and Diego di Bernardo. Discovery of drug mode of action and drug repositioning from transcriptional responses.

Proceedings of the National Academy of Sciences, 107(33):14621–14626, 2010. URL http://www.pnas.org/content/107/33/14621.abstract.

François Jacob and Jacques Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3(3):318–356, 1961. URL http://dx.doi.org/10.1016/S0022-2836(61)80072-7.

Michael I. Jordan. Graphical models. *Statistical Science*, 19(1):140–155, 2004. URL http://projecteuclid.org/euclid.ss/1089808279.

Andrew R. Joyce and Bernhard Ø. Palsson. The model organism as a system: integrating 'omics' data sets. *Nature Reviews Molecular Cell Biology*, 7(3): 198–210, 2006. URL http://dx.doi.org/10.1038/nrm1857.

Chris Kent. *Basics of Toxicology*. Preserving the Legacy. John Wiley & Sons, New York, New York, 1998. ISBN 0-471-29982-0.

Arto Klami. Variational Bayesian matching. In S.C.H. Hoi and W. Buntine, editors, *Proceedings of Asian Conference on Machine Learning*, volume 25 of *JMLR C&WP*, pages 205–220. JMLR, 2012. URL http://jmlr.org/proceedings/papers/v25/klami12/klami12.pdf.

Arto Klami, Seppo Virtanen, and Samuel Kaski. Bayesian canonical correlation analysis. *Journal of Machine Learning Research*, 14:965–1003, 2013. URL http://jmlr.org/papers/v14/klami13a.html.

Leo Lahti, Samuel Myllykangas, Sakari Knuutila, and Samuel Kaski. Dependency detection with similarity constraints. In *Proceedings of MLSP 2009, IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE, 2009. URL http://dx.doi.org/10.1109/MLSP.2009.5306192.

Nan M. Laird and James H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982. URL http://www.jstor.org/stable/2529876.

Justin Lamb, Emily D. Crawford, David Peck, Joshua W. Modell, Irene C. Blat, Matthew J. Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N. Ross, Michael Reich, Haley Hieronymus, Guo Wei, Scott A. Armstrong, Stephen J. Haggarty, Paul A. Clemons, Ru Wei, Steven A. Carr, Eric S. Lander, and Todd R. Golub. The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–1935, 2006. URL http://www.sciencemag.org/content/313/5795/1929.abstract.

Øyvind Langsrud. 50–50 multivariate analysis of variance for collinear responses. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 51(3):305–317, 2002. URL http://dx.doi.org/10.1111/1467-9884.00320.

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. Curran Associates, Inc., 2010. URL http://papers.nips.cc/paper/4153-cross-species-expression-analysis\-using-a-dirichlet-process-mixture-model-with-latent-matchings.

Andrew J. Link, Jimmy Eng, David M. Schieltz, Edwin Carmack, Gregory J. Mize, David R. Morris, Barbara M. Garvik, and John R. Yates. Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnology*, 17(7): 676–682, 1999. URL http://dx.doi.org/10.1038/10890.

Yong Lu, Peter Huggins, and Ziv Bar-Joseph. Cross species analysis of microarray expression data. *Bioinformatics*, 25(12):1476–1483, 2009. URL http://bioinformatics.oxfordjournals.org/content/25/12/1476.abstract.

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):1–26, 2009. URL http://dx.doi.org/10.2202/1544-6115.1436.

Mamas Mamas, Warwick B. Dunn, Ludwig Neyses, and Royston Goodacre. The role of metabolites and metabolomics in clinically applicable biomarkers of disease. *Archives of Toxicology*, 85(1):5–17, 2011. URL http://dx.doi.org/10.1007/s00204-010-0609-6.

Kantilal V. Mardia, J.T. Kent, and John M. Bibby. *Multivariate Analysis*. Probability and Mathematical Statistics. Academic Press, London, United Kingdom, 1st edition, 1979. ISBN 0-12-471250-9.

Matthew Meselson and Franklin W. Stahl. The replication of DNA in Escherichia coli. *Proceedings of the National Academy of Sciences*, 44(7):671–682, 1958. URL http://www.pnas.org/content/44/7/671.short.

Zhibao Mi, Kui Shen, Nan Song, Chunrong Cheng, Chi Song, Naftali Kaminski, and George C. Tseng. Module-based prediction approach for robust inter-study predictions in microarray data. *Bioinformatics*, 26(20):2586–2593, 2010. URL http://bioinformatics.oxfordjournals.org/content/26/20/2586.abstract.

Thomas P. Minka. Expectation propagation for approximate Bayesian inference. In J. Breese and D. Koller, editors, *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc., 2001. URL http://dl.acm.org/citation.cfm?id=2074022.2074067.

Tom J. Mitchell and John J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032, 1988. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.1988.10478694.

Douglas G. Montgomery. *Design and Analysis of Experiments*. John Wiley & Sons, New York, New York, 5th edition, 2001. ISBN 0-471-31649-0.

Ali Mortazavi, Brian A. Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7):621–628, 2008. URL http://dx.doi.org/10.1038/nmeth.1226.

Tomáš Pluskal, Sandra Castillo, Alejandro Villar-Briones, and Matej Orešič. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*, 11(1): 395, 2010. URL http://www.biomedcentral.com/1471-2105/11/395.

Lawrence Rabiner and Biing-Hwang Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1):4–16, 1986. URL http://dx.doi.org/10.1109/MASSP.1986.1165342.

Carl E. Rasmussen. The infinite Gaussian mixture model. In S.A. Solla, T.K. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 554–560. MIT Press, 2000. URL http://papers.nips.cc/paper/1745-the-infinite-gaussian-mixture-model.pdf.

Mark Schena, Dari Shalon, Ronald W. Davis, and Patrick O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, 1995. URL http://www.sciencemag.org/content/270/5235/467.abstract.

Patrick R. Schmid, Nathan P. Palmer, Isaac S. Kohane, and Bonnie Berger. Making sense out of massive data by going beyond differential expression. *Proceedings of the National Academy of Sciences*, 109(15):5594–5599, 2012. URL http://www.pnas.org/content/109/15/5594.abstract.

George A.F. Seber and Alan J. Lee. *Linear Regression Analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, New Jersey, 2nd edition, 2003. ISBN 0-471-41540-5.

Andrej Shevchenko and Kai Simons. Lipidomics: coming to grips with lipid diversity. *Nature Reviews Molecular Cell Biology*, 11(8):593–598, 2010. URL http://dx.doi.org/10.1038/nrm2934.

Age K. Smilde, Jeroen J. Jansen, Huub C.J. Hoefsloot, Robert-Jan A.N. Lamers, Jan van der Greef, and Marieke E. Timmerman. ANOVA-simultaneous component analysis (ASCA): A new tool for analyzing designed metabolomics data. *Bioinformatics*, 21(13):3043–3048, 2005. URL http://bioinformatics.oxfordjournals.org/content/21/13/3043.abstract.

Colin A. Smith, Elizabeth J. Want, Grace O'Maille, Ruben Abagyan, and Gary Siuzdak. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Analytical Chemistry*, 78(3):779–787, 2006. URL http://pubs.acs.org/doi/abs/10.1021/ac051437y.

Lloyd R. Snyder, Joseph J. Kirkland, and John W. Dolan. *Introduction to Modern Liquid Chromatography*. John Wiley & Sons, Hoboken, New Jersey, 3rd edition, 2010. ISBN 978-0-470-16754-0.

Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 935–943. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.

Student. The probable error of a mean. *Biometrika*, 6(1):1–25, 1908. URL http://biomet.oxfordjournals.org/content/6/1/1.short.

Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set

enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102 (43):15545–15550, 2005. URL http://www.pnas.org/content/102/43/15545.abstract.

Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. URL http://www.jstor.org/stable/2346178.

Michael E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001. URL http://dx.doi.org/10.1162/15324430152748236.

Abhishek Tripathi, Arto Klami, Matej Orešič, and Samuel Kaski. Matching samples of multiple views. *Data Mining and Knowledge Discovery*, 23(2):300–321, 2011. URL http://dx.doi.org/10.1007/s10618-010-0205-7.

J. Craig Venter et al. The sequence of the human genome. *Science*, 291(5507): 1304–1351, 2001. URL http://www.sciencemag.org/content/291/5507/1304.abstract.

Seppo Virtanen, Arto Klami, Suleiman A. Khan, and Samuel Kaski. Bayesian group factor analysis. In N. Lawrence and M. Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *JMLR W&CP*, pages 1269–1277. JMLR, 2012. URL http://jmlr.org/proceedings/papers/v22/virtanen12/virtanen12.pdf.

Christine Vogel and Edward M. Marcotte. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics*, 13(4):227–232, 2012. URL http://dx.doi.org/10.1038/nrg3185.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008. URL http://dx.doi.org/10.1561/2200000001.

Michael P. Washburn, Dirk Wolters, and John R. Yates. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology*, 19(3):242–247, 2001. URL http://dx.doi.org/10.1038/85686.

Michael D. Waters and Jennifer M. Fostel. Toxicogenomics and systems toxicology: aims and prospects. *Nature Reviews Genetics*, 5(12):936–948, 2004. URL http://dx.doi.org/10.1038/nrg1493.

James D. Watson and Francis H.C. Crick. Genetical implications of the structure of deoxyribonucleic acid. *Nature*, 171(4361):964–967, 1953. URL http://www.nature.com/nature/journal/v171/n4361/abs/171964b0.html.

Ian D. Wilson, Jeremy K. Nicholson, Jose Castro-Perez, Jennifer H. Granger, Kelly A. Johnson, Brian W. Smith, and Robert S. Plumb. High resolution "ultra performance" liquid chromatography coupled to oa-TOF mass spectrometry as a tool for differential metabolic pathway profiling in functional genomic studies. *Journal of Proteome Research*, 4(2):591–598, 2005. URL http://dx.doi.org/10.1021/pr049769r.

Aaron Wise, Zoltán N. Oltvai, and Ziv Bar-Joseph. Matching experiments across species using expression values and textual information. *Bioinformatics*, 28(12):i258–i264, 2012. URL http://bioinformatics.oxfordjournals.org/content/28/12/i258.abstract.

Svante Wold, Michael Sjöström, and Lennart Eriksson. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2): 109–130, 2001. URL http://dx.doi.org/10.1016/S0169-7439(01)00155-1.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006. URL http://dx.doi.org/10.1111/j.1467-9868.2005.00532.x.

DISSERTATIONS IN INFORMATION AND COMPUTER SCIENCE

Aalto-DD105/2014 Parkkinen, Juuso
Probabilistic components of molecular interactions and drug responses. 2014.

Aalto-DD108/2014 Faisal, Ali
Retrieval of Gene Expression Measurements with Probabilistic Models. 2014.

Aalto-DD110/2014 Virtanen, Seppo
Bayesian latent variable models for learning dependencies between multiple data sources. 2014.

Aalto-DD120/2014 Bergström-Lehtovirta, Joanna
The Effects of Mobility on Mobile Input. 2014.

Aalto-DD127/2014 Zhang, He
Advances in Nonnegative Matrix Decomposition with Application to Cluster Analysis. 2014.

Aalto-DD138/2014 Sovilj, Dušan
Learning Methods for Variable Selection and Time Series Prediction. 2014.

Aalto-DD144/2014 Eirola, Emil
Machine learning methods for incomplete data and variable selection. 2014.

Aalto-DD149/2014 Äijö, Tarmo
Computational Methods for Analysis of Dynamic Transcriptome and Its Regulation Through Chromatin Remodeling and Intracellular Signaling. 2014.

Aalto-DD156/2014 Sjöberg, Mats
From pixels to semantics: visual concept detection and its applications. 2014.

Aalto-DD157/2014 Adhikari, Prem Raj
Probabilistic Modelling of Multiresolution Biological Data. 2014.

Aalto-DD 171/2014

Inference of differences between samples is a fundamental problem in computational biology. Molecular measurements of biological organisms produce high-dimensional data but the number of test subjects in the experiments is limited. In this thesis, computational methods are presented for finding differences between high-dimensional observations and for extensions of this problem.

Since the effects and side-effects of new drug treatments are unknown and potentially dangerous, model organisms are used to study human diseases and their treatments. The computational translation of the outcome of an experiment from the model organism to humans is a problem, which is addressed in this thesis. Presented data translation methods identify responses to experimental treatments that are conserved across organisms.

ISBN 978-952-60-5932-7 (printed)
ISBN 978-952-60-5933-4 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Information and Computer Science
www.aalto.fi

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL DISSERTATIONS