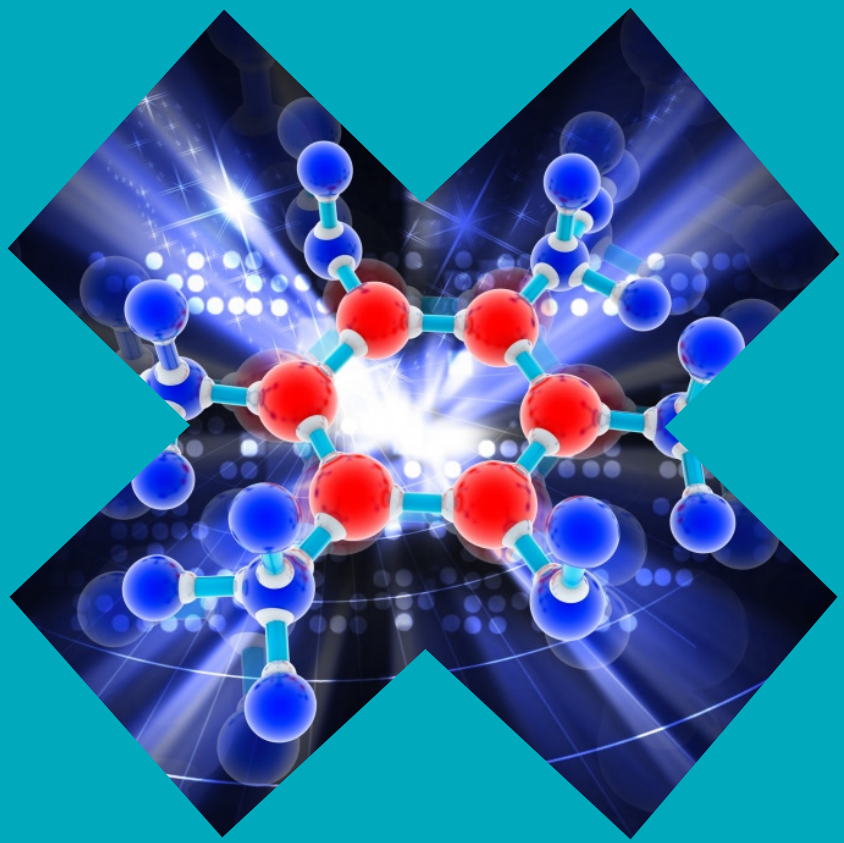
Department of Information and Computer Science

Probabilistic components of molecular interactions and drug responses

Juuso Parkkinen



Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 105/2014

Probabilistic components of molecular interactions and drug responses

Juuso Parkkinen

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 29 August 2014 at 12.

Aalto University
School of Science
Department of Information and Computer Science

Supervising professor
Prof. Samuel Kaski

Preliminary examiners
Prof. Matti Nykter, University of Tampere, Finland
Prof. Motoki Shiga, Gifu University, Japan

Opponent
Prof. Yoshihiro Yamanishi, Kyushu University, Japan

Aalto University publication series
DOCTORAL DISSERTATIONS 105/2014

ISBN 978-952-60-5773-6
ISBN 978-952-60-5774-3 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-5774-3

Images: rajcreationzs via FreeDigitalPhotos.net

Unigrafia Oy
Helsinki 2014

Finland

Publication orders (printed book):
juuso.parkkinen@iki.fi

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Juuso Parkkinen

Name of the doctoral dissertation
Probabilistic components of molecular interactions and drug responses

Publisher School of Science

Unit Department of Information and Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 105/2014

Field of research Bioinformatics

Manuscript submitted 16 April 2014 **Date of the defence** 29 August 2014

Permission to publish granted (date) 4 June 2014 **Language** English

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

Abstract

A fundamental question in medicine is how cancer and other complex diseases operate on the molecular level. Identifying the detailed mechanisms and interactions of how diseases progress and respond to drug treatments is essential for developing effective therapies. High-throughput molecular profiling technologies have provided vast amounts of measurement data of these phenomena. However, making sense of these masses of data is far from straightforward and requires advanced computational analysis methods.

Probabilistic component models have been proven an effective tool in analysing and integrating high-dimensional and noisy molecular profiling data sources, such as gene expression. Such models can identify coherent components from the data, and interpreting these components provides insights about the underlying biological processes, such as disease progression and drug responses. In this thesis, probabilistic component models are applied and extended to identify and analyse molecular interaction and drug response patterns.

Identifying functionally coherent gene modules from high-throughput measurements is a central task in many biomedical applications. In this thesis, an earlier component model for network data is extended for capturing functional modules from combinations of gene expression and protein interaction data. The identified modules provide hypotheses for novel molecular pathways and protein functions.

High-throughput drug treatment measurements have made possible the detailed analysis of molecular drug responses and toxicity. In this thesis, probabilistic component models are applied to detect coherent drug response patterns from gene expression data. These patterns provide detailed insights to drug mechanisms of action and are highly applicable in cancer therapy development. Moreover, by associating the identified drug response components to toxicological outcomes, the first comprehensive view of molecular toxicogenomic responses is constructed with high performance in drug toxicity prediction.

Keywords Bayesian statistics, gene expression, machine learning, molecular medicine, probabilistic component models, toxicology

ISBN (printed) 978-952-60-5773-6 **ISBN (pdf)** 978-952-60-5774-3

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:978-952-60-5774-3

A? Aalto-yliopisto

Tiivistelmä

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

Tekijä
Juuso Parkkinen

Väitöskirjan nimi
Solutason lääkevasteiden ja molekyylivuorovaikutusten tilastollisia komponentteja

Julkaisija Perustieteiden korkeakoulu

Yksikkö Tietojenkäsittelytieteen laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 105/2014

Tutkimusala Bioinformatiikka

Käsikirjoituksen pvm 16.04.2014 **Väitöspäivä** 29.08.2014

Julkaisuluvan myöntämispäivä 04.06.2014 **Kieli** Englanti

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

Tiivistelmä

Syövän ja muiden monimutkaisten tautien molekyylitason mekanismien selvittäminen on keskeinen ongelma lääketieteessä. Tautien yksityiskohtaisten leviämismekanismien ja lääkevasteiden tunnistaminen on tärkeää tehokkaiden hoitomenetelmien kehityksessä. Näistä ilmiöistä on kerätty suuria molekyylitason aineistoja uusien mittausmenetelmien avulla. Näiden aineistojen tehokas hyödyntäminen vaatii kehittyneitä laskennallisia analyysimenetelmiä.

Tilastolliset komponenttimallit ovat osoittautuneet tehokkaaksi työkaluksi suuriulotteisten ja kohinaisten mittausaineistojen, kuten geenieksprressiodatan, analysoinnissa ja yhdistämisessä. Tällaisilla malleilla datasta voidaan tunnistaa komponentteja, joita tulkitsemalla saadaan uuttaa tietoa biologisista prosesseista, kuten tautien etenemisestä ja lääkevasteista. Tässä väitöskirjatyössä tilastollisia komponenttimalleja sovelletaan molekyylien vuorovaikutusten ja lääkevasteiden tunnistamiseen ja analysointiin.

Toiminnallisten geenimoduulien tunnistaminen suurista mittausaineistoista on keskeinen tehtävä monissa biolääketieteen sovelluksissa. Tässä väitöskirjatyössä aiempaa vuorovaikutuskomponenttimallia laajennetaan geenimoduulien tunnistamiseen geenieksprressio- ja proteiinien vuorovaikutusaineistoja yhdistämällä. Tunnistetut moduulit auttavat yksittäisten proteiinien ja niiden vuorovaikutusketjujen toiminnan selvittämisessä.

Perimänlaajuiset lääkealtistusmittaukset ovat mahdollistaneet lääkevasteiden ja toksisuuden yksityiskohtaisen tutkimuksen. Tässä väitöskirjatyössä komponenttimalleilla haetaan säännönmukaisia lääkevasteita geenieksprressioaineistoista. Tunnistetut vasteet antavat uutta tietoa lääkkeiden vaikutuksista ja niitä voidaan käyttää uusien syöpähoitojen kehittämiseen. Hakemalla tilastollisia yhteyksiä geenieksprressiovasteiden ja toksisuusmittausten välillä saadaan lisäksi kokonaiskuva lääkkeiden toksisuuteen liittyvistä molekyylitason vasteista ja mahdollistetaan lääkkeiden toksisuuden tehokas ennustaminen.

Avainsanat Bayesiläinen tilastotiede, geenieksprressio, koneoppiminen, molekyylilääketiede, tilastolliset komponenttimallit, toksikologia

ISBN (painettu) 978-952-60-5773-6 **ISBN (pdf)** 978-952-60-5774-3

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:978-952-60-5774-3

Preface

This thesis describes my doctoral studies at the Department of Information and Computer Science in Aalto University School of Science. I have had the privilege of being part of two Finnish Centres of Excellence: Computational Inference Research (COIN) and Adaptive Informatics Research Centre (AIRC), and also the Helsinki Institute for Information Technology HIIT. My work has been partially supported by the Helsinki Doctoral Programme in Computer Science (HECSE), Academy of Finland project Computational Modeling of the Biological Effects of Chemicals, and Pattern Analysis, Statistical Modelling and Computational Learning Network of Excellence (PASCAL 2 EU Network of Excellence).

I would like to thank my instructor and supervisor, Prof. Samuel Kaski, for introducing me to the world of machine learning and bioinformatics. His continuous support and guidance has taught me a lot about research and science.

I am thankful for the pre-examiners of my thesis, Prof. Matti Nykter and Prof. Motoki Shiga, for their expert comments.

I would like to thank my co-authors Prof. Roland Grafström, and Drs. Krister Wennerberg, Egon Willighagen, Pekka Kohonen and Rebecca Ceder for leading me to the fascinating world of toxicology and exploring the possibilities of computational toxicology with me. I would also like to express my gratitude for all the wonderful people at the Institute for Molecular Medicine Finland FIMM, including Prof. Olli Kallioniemi, Dr. Tero Aittokallio and Tea Pemovska, for discussions and insights about molecular medicine.

I am thankful to my co-authors Prof. Jaakko Peltonen, Kristian Nybo, Tommi Suvitaival and Seppo Virtanen, and also all the other current and former MI group members, especially Suleiman Ali Khan, Jussi Gillberg, Eemeli Leppäaho, Ali Faisal, and Drs. Leo Lahti, Arto Klami, Janne

Sinkkonen, Elisabeth Georgii and José Caldas, for numerous inspiring discussion about machine learning, bioinformatics and life in general. Special thanks go to the fellow students whom with I enjoyed countless coffee breaks playing soccer.

This thesis would not have been possible without the wonderful research environment at the ICS and the former CIS departments. I thank Prof. Pekka Orponen, Prof. Erkki Oja and Prof. Olli Simula for leading the departments. I also thank Tarja Pihamaa, Leila Koivisto, Minna Kauppila, Miki Sirola and Markku Ranta for providing all necessary facilities and answering countless practical questions.

I am grateful to all friends who have been there when I needed them. Special thanks go to Mika for being always ready for a beer and Kalle for providing peer support for simultaneously finishing a thesis and building a family. I am grateful to the Open Knowledge Finland community and especially the fellow Louhos dudes, Leo, Joona and Markus, for showing me the value of openness in all things. I also thank Espoo Ultimate Club for offering a chance to relieve the mental burden of the thesis.

Finally, I would like to thank my family. I thank my parents Helena and Jaakko for their continuous love and support in all stages of my life. I thank my sisters Johanna and Pauliina for providing an invaluable counterweight in the family. And most importantly, I thank my dear wife Sanna and my wonderful children for always being there to support me, and tolerating both the upsides and downsides of the thesis work.

Espoo, July 14, 2014,

Juuso Parkkinen

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 Juuso A. Parkkinen and Samuel Kaski. Searching for functional gene modules with interaction component models. *BMC Systems Biology*, **4**:4, 2010.

II Juuso A. Parkkinen, Pekka Kohonen, Egon L. Willighagen, Rebecca Ceder, Krister Wennerber, Samuel Kaski, Roland C. Grafström. Toxicogenomics-based probabilistic modelling enables prediction of dose-dependent toxicity. *Submitted to a journal*, 2014.

III Juuso A. Parkkinen and Samuel Kaski. Probabilistic drug connectivity mapping. *BMC Bioinformatics*, **15**:113, 2014.

IV Tommi Suvitaival, Juuso A. Parkkinen, Seppo Virtanen, Samuel Kaski. Cross-organism toxicogenomics with group factor analysis. *Systems Biomedicine*, 2:e29291, 2014.

V Juuso Parkkinen, Kristian Nybo, Jaakko Peltonen and Samuel Kaski. Graph Visualization With Latent Variable Models. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, Pages 94-101. ACM, New York, NY, USA, 2010.

Author's Contribution

Publication I: "Searching for functional gene modules with interaction component models"

Co-designed, derived and implemented the extension to the probabilistic interaction component model. Designed and ran all experiments and validation. Co-wrote the manuscript.

Publication II: "Toxicogenomics-based probabilistic modelling enables prediction of dose-dependent toxicity"

Co-designed the study. Co-designed the probabilistic model. Designed and ran all experiments and quantitative validation. Participated in interpreting the biological and toxicological results. Co-wrote the manuscript.

Publication III: "Probabilistic drug connectivity mapping"

Co-designed the implemented the application of group factor analysis to data retrieval. Designed and ran all experiments and validation. Co-wrote the manuscript.

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

Co-designed the retrieval approach to the data translation problem and the related experiments and validation. Co-wrote the introduction, results, and conclusions in the manuscript.

Publication V: "Graph Visualization With Latent Variable Models"

Co-designed and implemented the probabilistic model-based graph visualisation. Co-designed and ran all experiments and quantitative validation. Co-wrote the manuscript.

List of Abbreviations and Symbols

In this thesis boldface symbols are used to denote matrices and vectors. Capital symbols (**X**) denote matrices and lowercase symbols (**x**) denote vectors. Scalar variables are denoted by lowercase symbols.

$\mathbb{R}$	Real domain
$\mathbf{X}$	Data matrix
$\mathbf{x}$	Data vector
x	Scalar data point
$\mathbf{I}$	Identity matrix
$p(x)$	Probability or probability density of x
$p(x\|y)$	Conditional probability of x given y
$p(x, y)$	Joint probability of x and y
$x \propto y$	x is proportional to y
$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
$\mathcal{G}(a, b)$	Gamma distribution with shape a and scale b
ARD	Automatic relevance detection
ATC	Anatomic Therapeutic Chemical classification
AUC	Area under the curve
CCA	Canonical correlation analysis
CMap	Connectivity Map
DNA	Deoxyribonucleic acid
GFA	Group factor analysis
GI50	50 % growth inhibition
GO	Gene Ontology
GSEA	Gene set enrichment analysis
HMoF	Hidden modular random field
IC50	50 % lethal concentration
ICM	Interaction component model
LDA	Latent Dirichlet allocation

LINCS	Library of integrated network-based cellular signatures
MCMC	Markov chain Monte Carlo
mRNA	Messenger RNA
NeRV	Neighbour retrieval visualiser
PCA	Principal component analysis
PTGS	Predictive toxicogenomics space
QSAR	Quantitative structure-activity relationship
RNA	Ribonucleic acid
SSN-LDA	Simple social network LDA
TGI	Total growth inhibition
VB	Variational Bayes

1. Introduction

1.1 Motivation

The role of computational data analysis is increasing throughout the quantitative research fields. In molecular biology for example, the rapid development of measurement technologies has lead to an exponential growth in the amount of available genomic data in public repositories [8, 90]. This trend also extends to the society in general, as data are increasing at an unprecedented rate from numerous sources such as web pages, personal images and videos, government databases, and data from mobile and sensory devices [93].

Data analysis refers to the process of collecting, transforming and modelling data with the aim of discovering useful information, providing understanding, and supporting decision making. It is largely dependent on computational methods, as analysing large amounts of data manually is impossible. The rapidly growing amount of data imposes an increasing demand for the development of effective computational methods that can handle various demanding data analysis challenges.

Computational biology and medicine are among the most advanced fields in adopting advanced computational data analysis methods. The increasing use of high-throughput measurement technologies and computational modelling has led to the era of systems biology, revealing new levels of biological function and organisation [89]. Prediction of drug effects in humans will advance pharmaceutical research and clinical trials. Computational prognostics and diagnostics, combining clinical data with molecular profiling is causing fundamental changes in the practice of medicine [97]. Large amounts genomic data available in public databases provide good opportunities for developing computational methods that can also lead to

new discoveries [96].

Clear progress has been seen in the development of both measurement technologies and computational analysis tools for molecular medicine applications, but also many challenges remain. A fundamental requirement is that the computational methods must be able to filter and combine multiple disparate and noisy data sources, and moreover provide proper understanding of the underlying molecular mechanisms in addition to providing reliable predictions [25].

A promising data analysis genre for tackling these challenges is machine learning, combining elements from statistics and computer science. This thesis focuses on applying and extending Bayesian probabilistic models that offer a suitable toolbox for approaching the complex data analysis challenges in biomedical applications. In Bayesian modelling, probabilities are used to quantify uncertainties in the data and model parameters, allowing inference even from very noisy data sources [42].

Probabilistic component models enable inference about the underlying processes behind the noisy and high-dimensional observation data. For example, components identified from combinations of gene expression and protein interaction data can be interpreted as molecular pathways [103]. When applied to drug treatment measurements, component models can provide insights into the specific chemical structures inducing molecular drug responses [68]. Additionally, the component models can be used for inferring similarity structure within the data and further applied to advanced information retrieval and visualisation tasks [16].

The data analysis methods used and developed in this thesis are exploratory in the sense that the idea is to discover new information and formulate new hypotheses about the data, as opposed to confirmatory data analysis, where predefined hypotheses are tested statistically [126]. The methods are also unsupervised, meaning that the general task is to discover underlying patterns in the unlabelled data, in contrast to supervised methods, where the aim is to predict or classify the properties of the elements in the input data based on given labels [10].

1.2 Objectives and scope

In this thesis probabilistic component models are extended and applied for analysing noisy and high-dimensional molecular profiling data sources, especially gene expression. The contributions are in the cross-section

of methods development and applied research, where the challenges in molecular medicine motivate model development, which in turn allows solving novel application problems. Thus both methodological and biomedical research relevant for the thesis are presented in detail.

The general objective of the thesis is to study the applicability of probabilistic component models in various molecular medicine applications. The specific objectives of this thesis are formulated as the following research questions:

1. RQ1: Can component models improve the identification of functional gene modules?

2. RQ2: Can component models capture molecular drug response patterns that are informative of therapeutic and toxic effects?

3. RQ3: Can component models be used for visualising global structures in graphs?

The first main contribution of the thesis is to answer RQ1 by identifying and analysing functional gene modules from combinations of gene expression and protein interaction data. In Publication I an earlier probabilistic model is applied and extended to integrate the data sources for improved module detection.

The second main contribution is to answer RQ2 by applying probabilistic component models for drug response and toxicity analysis. In Publication II, drug treatment gene expression measurements are associated to the corresponding toxicological responses, providing novel insights into toxicity-related gene expression patterns with high predictive performance. In Publications III and IV, a novel model-based method is introduced to drug connectivity mapping, where the aim is to match drugs to other drugs or diseases based on their gene expression profile similarities.

The third contribution is to answer RQ3 by applying model-based retrieval to information visualisation applications. In Publication V, a probabilistic model is used to visualise graph data in a way that reveals interesting structural patterns.

Figure 1.1 gives a schematic overview of how the research questions and publications are related.

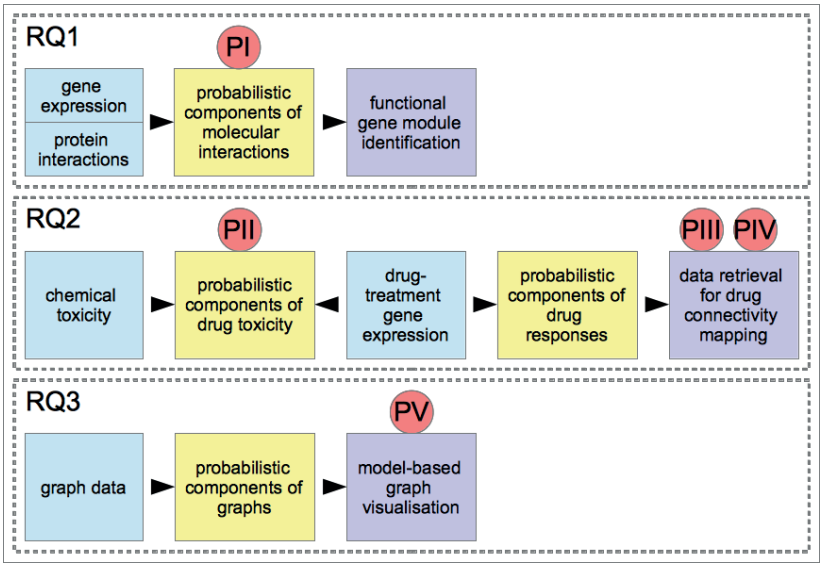


Figure 1.1. A schematic overview of the publications and research questions addressed in the thesis. Blue boxes represent data sets, yellow boxes represent probabilistic component models of the data, and purple boxes represent specific tasks. Publications are indicated with red circles and research questions with dashed line boxes.

1.3 Organization of the thesis

The thesis is organised as follows: Chapter 2 gives biomedical background for understanding the contributions in molecular medicine and gene expression analysis, and introduces the specific applications in more detail. Chapter 3 gives background for the probabilistic component models used and developed in the thesis. In Chapter 4, probabilistic models are applied to identify molecular interaction components. In Chapter 5, probabilistic models are applied to toxicogenomics and drug connectivity mapping. Chapter 6 introduces a probabilistic model for graph visualisation. Finally, the thesis is concluded in Chapter 7.

2. Molecular biology and medicine

In this chapter, the basics of molecular biology and medicine are introduced, including common computational analysis methods used in the field. Focus is especially on gene expression data analysis, which is one of the corner stones of modern molecular biology and medicine. Additionally, molecular interactions and drug responses are covered.

2.1 Molecular biology and gene expression

Molecular biology aims to understand how the different molecules in the cell function and interact. The basic cellular molecule types are *nucleic acids* and *proteins* [3]. Nucleic acids form long molecular chains called *DNA* and *RNA*, where the sequence of the four basic nucleic acids encodes heritable information. Specific parts of the DNA sequence are called *genes*, and they contain the information for building protein molecules [3]. The collection of all genes is called the *genome*, containing all heritable information in living organisms.

The so called *central dogma of molecular biology* [26] describes the chain of information from DNA to messenger RNA (mRNA) through *transcription*, and further from mRNA to proteins through *translation*, also called *protein synthesis*. The functional product of a gene is thus a protein molecule. This process where genetic information is transformed through protein synthesis is called *gene expression*, and it is the fundamental mechanism underlying cellular functions. The different cells in a complex organism differ notably in both their structure and function. This is due to changes in the gene expression between the cells, resulting in different proteins being synthesised. Gene expression can in principle be regulated in all the steps in the chain of information from DNA to proteins, but the most important is *transcriptional regulation*, controlling when and how

often a given gene is transcribed [3].

After the proteins have been synthesised, they carry out their functions through complicated networks, where proteins interact with each other and other molecules to mediate the cellular signals. The molecular cascades carrying specific tasks in the cells are called *pathways* [3]. Pathways have been identified for many important functions, such as cell signalling and cell cycle. Studying gene expression and molecular networks and pathways are central tasks in molecular biology.

2.1.1 Measuring gene expression

Measuring gene expression is a central part of molecular biology research, providing detailed information about the cellular signalling networks and molecular basis of diseases and drug mechanisms. One of the most prominent gene expression measurement technologies are *microarrays*, which allow parallel quantification of thousands of mRNA molecules [99]. Later, also *RNA-sequencing* technology has been applied to whole genome expression analysis, promising less noisy measurements at lower costs [84, 124, 137]. This thesis focuses on gene expression data based on microarray measurements.

Microarray analysis assumes that the measured intensities for each gene represent its relative expression level [92]. Biologically relevant patterns of expression are typically identified by comparing measured expression levels between different states. However, before appropriate comparisons can be made, the data must be transformed to account for low-quality measurements and noise. Microarray measurements contain a lot of noise from various sources [94]. Some amount of noise exists due to purely biological reasons, including the stochastic nature of the biochemical reactions, differences in internal states of the cells, and ongoing mutations. Additionally, the measurement technologies impose variation to the results.

2.1.2 Differential expression

A central way to analyse gene expression measurements is called *differential expression*, where measurements from the condition of interest are compared against a control. Differential expression is computed as the ratio, or fold-change, for each gene between the treatment and control samples, describing the gene activity under the specific condition. For ex-

ample, a typical way to analyse cancer mechanisms in gene expression is to compare measurement samples from cancerous samples from patients to healthy tissues.

To assess the significance of differential expression, various statistical tests are used, such as Student's t-test [28]. Due to the high number of genes typically tested, the problem of multiple hypothesis testing occurs. Without a proper correction, such testing is likely to produce many false positives, that is, genes falsely identified as differentially expressed. Standard methods for correcting for this are Bonferroni correction and false discovery rate [33]. Instead of individual genes, one can also test the differential expression for a set of genes. This is more robust against noise and more likely to detect subtle changes in gene expression. An established method is Gene Set Enrichment Analysis (GSEA) [118].

2.1.3 Gene expression clustering

Genes and proteins are typically organised into functional categories [30, 95]. Thus, a central task in molecular biology is to identify and analyse such functionally coherent *modules* [59]. The most common gene module detection approach is *clustering*, where the idea is to group the genes such that similar genes are in the same groups while dissimilar genes are in different groups. The rationale is that genes with similar expression profiles that end up clustered together are typically functionally similar [30]. This allows researchers to make hypotheses for the functions of unknown genes that appear to be similar to known genes. Clustering can based on absolute or differential expression.

A large number of methods have been developed for clustering genes, differing in how the similarity is defined and how the similarity is used for clustering. A commonly used similarity measure is Pearson correlation [35]. Popular clustering methods include hierarchical clustering, K-means, self-organising maps, graph-theoretic approaches, and model-based methods [30, 66, 105]. Despite extensive efforts, no clear one-size-fits-all solution for gene clustering has been developed [30]. Additional gene filtering and dimensionality reduction steps have proposed to enhance clustering performance [29]. The goodness of clustering is typically measured against external ground truth for the genes, such as functional annotations. The most popular such annotation is the Gene Ontology [6].

Biclustering is an alternative to clustering methods that operates based on the whole range of measured conditions. In biclustering, closely re-

lated to subspace clustering [66], subsets of genes exhibiting consistent patterns over a subset of the conditions are searched for, making the analysis local rather than global [120]. In addition to functional similarity of genes, biclusters can be used to make hypotheses about the conditions within the cluster that exhibit consistent gene expression. For example drugs that act consistently on the set of genes can have shared mechanisms of actions through these genes.

2.1.4 Molecular interactions

Molecular interaction data provide another view to the cellular functions, complementing gene expression measurements. The most important molecular interaction type are protein-protein interactions. Multiple high-throughput measurement technologies have been developed for measuring interactions between proteins, such as yeast two-hybrid systems and protein complex purification [134]. Measuring interactions is challenging, because the molecular interactions last for only very short periods of time and may be highly context-dependent. Thus the measured data sets are also highly noisy, containing both spurious interactions and missing, unobserved interactions [134].

Protein interaction data can be analysed to detect gene modules, with the idea that if the proteins that the genes code are interacting in the cell, the genes are again likely to share similar functions [49, 106] and participate in the same pathways [80, 103]. This can also be done by combining protein interactions with gene expression data. Such integrative analyses have the potential of providing more reliable results, as both gene expression and protein interaction data are known to be noisy [94, 134].

An often-used way to combining gene expression and interaction data is to transform both data types into distances, resulting in a combined network data set. The task is then transformed into network clustering, which is a common task in multiple fields. Several network clustering methods have been applied to the combined network clustering based on gene expression and protein interaction data by identifying tightly connected gene groups [51, 128, 130].

In this thesis, probabilistic models for combining gene expression and protein interaction data are reviewed in Section 3.3 and novel model extensions for the task are introduced in Chapter 4.

2.2 Molecular medicine and toxicology

Molecular medicine refers to the application of genetic or DNA-based knowledge to medical applications [125]. A central factor in the advancement of gene expression-based analysis methods for molecular medicine is that large amounts of gene expression experiments are uploaded to public databases, such as ArrayExpress [90] and Gene Expression Omnibus (GEO) [34]. Using such *compendiums* of gene expression, the discovery of novel gene functions has increased rapidly [60]. Moreover, databases with measurements collected from thousands of patients allow the association of genes to specific patient phenotypes, such as diseases. Several approaches have identified general gene expression modules as well as modules specific to tissues, disease or drug treatments [14, 21, 32, 45, 75, 100, 119].

The Connectivity Mapping (CMap) [78] database has pioneered the use of high-throughput drug treatment measurements for drug discovery and development. With a large collection of drug-induced gene expression alterations, connections can be searched between drugs and genes. The CMap data has been studied for details of drug-induced regulation of target proteins [64]. Furthermore, using expression profiling data from medical conditions, also diseases can be connected to the genes and drugs, creating huge potential for computational drug repositioning [63, 91].

Chemical toxicity analysis is another field that has benefited from the joint development of measurement technologies and computational methods. Ensuring the healthiness of novel drugs, chemicals, and other environmental chemicals is an essential part of pharmacology and toxicology. It is also costly, requiring extensive animal experimentation. There is thus a clear need for developing more efficient and accurate computational and cell line measurements-based screening procedures for identifying chemical hazards and prioritising chemicals for further testing on live patients [24, 53].

A recent focus in molecular medicine is towards *personalised medicine*, combining personal molecular profiling measurements and existing disease and drug treatment databases to find the best suitable therapies for individual patients [25]. Machine learning methods show great promise in providing predictive tools for personalised medicine [36]. One key challenge for advanced computational methods is thus to provide results that generalise from model organisms to humans.

2.2.1 Drug sensitivity analysis

Drug screening refers to evaluation of the sensitivity of different tissues to drug chemicals, typically done *in vitro*, that is, on cell lines. By measuring the growth of the cell lines over a range of concentrations, a dose-response-curve for each drug treatment is obtained. The curve can then be summarised with various drug sensitivity values, such as 50 % growth inhibition (GI50) and 50 % lethal concentration (IC50). Drug screening is also sometimes called pharmacological or toxicological profiling, depending on the application field.

Drug screening is often done in a high-throughput manner, measuring simultaneously a large set of chemicals and cell lines. The US National Cancer Institute human tumour cell line anticancer drug screen (NCI60) for example consists of 60 human cancer cell lines and sensitivity measurements for over 40 000 chemicals [111]. NCI60 offers chemical structure information for the drugs and gene expression profiles for the cell lines, and this information can be used to predict drug sensitivity with computational methods.

One of the most common computational drug sensitivity prediction methods is called quantitative structure-activity relationship (QSAR) assessment. In QSAR, computational methods are used to predict the drug sensitivity or toxicity values based on various *molecular descriptors*, representing the structure of the chemicals [79, 142]. QSAR has been widely successful, but it has some serious weaknesses: It is in many cases not able to detect the differences caused by tiny changes in the chemical structures, and it can not be used for tissue-specific drug sensitivity prediction. An alternative for tissue-specific drug sensitivity prediction is to use gene expression measurements [74, 138]. More recently, several large-scale drug screening efforts have profiled the genomes for hundreds of human cancer cell lines and associated genetic mutations with tissue-specific drug sensitivity [7, 40]. Some *in vitro* drug sensitivity predictions have also been successfully validated in clinical trials [116].

2.2.2 Toxicogenomics

QSAR has traditionally been the most common approach to predictive toxicology, but recently also many toxicogenomic approaches have been introduced. The aim in toxicogenomics is to find associations between gene expression and toxicological data from chemical perturbations. Such as-

sociations can then be used to understand the molecular mechanisms underlying toxicological outcomes and ultimately to predict toxicity [23, 54].

One big obstacle for the wide-spread applicability of toxicogenomics is the lack of suitable data sets with enough chemicals screened for proper statistical evaluation of the associations. Recently, a few projects have been initiated to tackle this problem, including ToxCast [122], DrugMatrix [39] and ToxBank [72].

The TG-GATEs database from the Japanese Toxicogenomics Project [127] offers a collection of genome-wide gene expression measurements for about 150 drug chemicals for liver cells from three organisms: human and rat *in vitro*, and living rats *in vivo*. Additionally, the database contains toxic outcome observations, including blood level measurements and observed liver injuries from the rats *in vivo*. Recently, several statistical classification methods were used for predicting these rat *in vivo* liver toxicity outcomes based on both QSAR and gene expression data. By selecting a subset of the genes a very good toxicity classification performance was achieved, whereas the performance using chemical descriptors alone was notably lower, suggesting that drug treatment gene expression data are more informative about toxicology than chemical descriptors. [82]

Another challenge for toxicogenomics is to separate the toxic effects from intended therapeutic effects and noise. For example, many traditional cancer drugs are in general highly toxic, killing both cancer and normal cells [117]. Toxic responses also vary across organisms, and thus using model organisms to predict toxicity in humans is indirect at best and requires sophisticated predictive models that can find signals that generalise across organisms. In this thesis, contributions to toxicogenomics are introduced in Chapter 5.

2.2.3 Drug connectivity mapping

The process of connecting drugs to other drugs or diseases based on gene expression profile similarity is known as *connectivity mapping*. The general idea is to identify a representative signature from the genome-wide differential expression profile that captures the essential phenotype-related modifications. Using a novel drug as a query and searching for other drugs with similar signatures from a large database, such as CMap, researchers have then been able to make hypotheses of novel mechanisms of actions for new drugs [62, 78]. Moreover, extracting similar signatures from patient samples one can search for inverse correlations between spe-

cific drugs and diseases, indicating potential novel therapies for existing drugs [109, 114].

Connectivity mapping has also been proposed for toxicity analysis, where the focus is on making hypotheses about the toxicity risks and toxicity-related mechanisms of the query drug, based on other drugs [115]. Recently, several efforts have applied the connectivity mapping principle to the TG-GATEs data set with promising results [15, 144]. These applications show that connectivity mapping can provide relevant information also when quantitative prediction of drug sensitivity, toxicity, or other properties is not possible due to lack of data.

Formulating connectivity mapping as an information retrieval problem, the key is how to define the relevance measure well. Current connectivity mapping approaches define the relevance by computing rank-based similarity statistics between the gene expression profiles [62, 78]. This approach can integrate data from multiple platforms and reduce the effect of batch effects. However, the current methods have simply aggregated data over multiple experimental factors, such as cell types, doses and time points. For personalised medicine applications, it would be important to identify also the cell type-specific drug responses. Already existing drug treatment transcriptional databases, such as the Connectivity Map [78] and TG-GATEs [127], provide measurements for multiple cell types, and the number will likely grow in the future. For example, the recently established Library of integrated network-based cellular signatures (LINCS, http:/ /www.lincsproject.org/), offers data for thousands of chemicals on tens of cell lines. Proper data integration methods will thus be needed for separating the cell line-specific effects from the general drug responses. In this thesis, contributions to drug connectivity mapping are introduced in Section 5.2.

3. Probabilistic component models for molecular biology

Many application problems in molecular biology and medicine involve analysis of high-dimensional and noisy data from multiple sources. The Bayesian data analysis framework provides suitable tools for tackling such modelling challenges. In this chapter, the methodological background of Bayesian data analysis is first covered, and the specific probabilistic models used and developed in this thesis are then introduced in more detail.

3.1 Bayesian data analysis

The aim of Bayesian data analysis is to make inferences from data using probability models for both observed quantities and unobserved quantities of interest. The core of Bayesian methods is the explicit use of probability to quantify uncertainty in the statistical data analysis [42], which makes Bayesian probabilistic models a convenient choice for addressing the biomedical applications in this thesis. As discussed in Chapter 2, gene expression and other common data sources are highly noisy, due to the complexity of the biological phenomena and measurement techniques. Bayesian probabilistic modelling of such data is beneficial, as it handles the noise as uncertainties in the data in terms of an underlying probabilistic model. As long as these modelling assumptions about these uncertainties are correct, the results are likely to be better than with methods that ignore the uncertainty.

The genomic data sources also typically have a large number of features and low number of samples, commonly referred to as the *'large p, small n'* problem. This can be approached by assuming that the data has some underlying lower dimensional structure, which can then be estimated with probabilistic inference. For example, a probabilistic model can be used to

decompose the genome-wide gene expression measurements into a set of coherent components, interpretable as biological processes.

Probabilistic models also provide principled tools for combining data from multiple sources. *Multi-view* models allow integration of multiple data sets with shared samples. More complicated multi-source settings can be approached with specifically tailored probabilistic models that also allow integration of different types of data.

3.1.1 Basic concepts

In Bayesian statistical data analysis, probability is the fundamental measure of uncertainty, used to make statements about the partial knowledge of the system [42]. In principle, everything unknown is described with a suitable probability distribution.

The basic building block of Bayesian analysis is the *joint probability distribution* $p(\theta, y)$ for the model *parameters* θ and the observed data y. This can be written as the product of the *prior distribution* $p(\theta)$ and the *likelihood function* $p(y|\theta)$: $p(\theta, y) = p(\theta)p(y|\theta)$. Bayes' rule then provides the conditional probability of θ given y, called the *posterior* distribution

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)} \ . \tag{3.1}$$

These simple expressions encapsulate the technical core of Bayesian inference: develop the model $p(\theta, y)$ and perform the necessary computations, known as *inference*, to summarise $p(\theta|y)$ in appropriate ways [42].

3.1.2 Inference

The process of finding the posterior $p(\theta|y)$ is called inference on parameters [42]. The most conventional inference procedure is drawing random samples from the posterior. In simple modelling tasks the posterior $p(\theta|y)$ of the parameters of interest can be computed in an analytic form, allowing direct draws from the distribution. In most practical applications, however, the exact computation of the posterior is intractable and it needs to be approximated.

Markov chain Monte Carlo (MCMC) is a commonly used Bayesian inference technique when sampling θ directly from $p(\theta|y)$ is not feasible [42]. MCMC is based on drawing values of θ from some approximate distributions, and then correcting those draws to better approximate the target posterior distribution $p(\theta|y)$. The samples are drawn sequentially, with distribution of the sampled draws depending on the last value drawn.

The idea is that the approximate distributions are improved at each step in the simulation, eventually converging to the target distribution.

Variational Bayesian (VB) methods are an alternative to sampling methods for making use of a posterior distribution that is computationally too intensive to sample from directly [10]. In VB, the posterior distribution $p(\theta|y)$ is approximated by a variational distribution $q(\theta)$: $p(\theta|y) \approx q(\theta)$, where q is chosen as a simpler distribution than the original posterior. The goal is then to make q as similar to p as possible, based on the Kullback-Leibler divergence.

3.1.3 Generative models

A common way to define a Bayesian probabilistic model is to use the *generative modelling* framework [10]. A generative model describes the process by which the observed data was generated, using probability distributions to describe the relationships between the observed and unobserved variables. Generative models can also be expressed in terms of *graphical models*, referring to diagrammatic representations of probability distributions. In a graphical model, each node represents a random variable, and the links express dependency relationships between these variables. The graph thus captures the generative process of the model.

An example of a graphical model is shown in figure 3.1, corresponding to the probability model

$$P(Y|X,\alpha,\beta) = \prod_{g}^{G} P(X_g|\alpha) \cdot P(Y_g|X_g,\beta) \ , \tag{3.2}$$

where Y are observed data, X are model parameters, and α and β are prior parameters.

3.2 Probabilistic latent variable models

In this section some common probabilistic models for analysing noisy and high-dimensional data, such as gene expression, are described. A central modelling assumption is that the high-dimensional input data is generated by a set of underlying factors, or components. As these components are not observed, they are called *latent* components [10]. The number of latent components is usually much lower than the dimensionality of the data, helping to tackle the 'large p, small n' problem. In this thesis the terms latent variable, component and factor are used interchangeably.

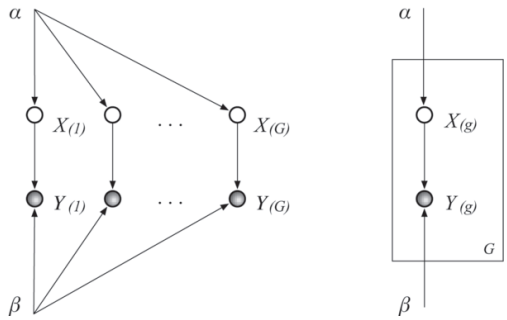


Figure 3.1. An example of a probabilistic graphical model, with two equivalent representations: **Left:** full model, **right:** a *plate diagram* representation of the model. Random variables are denoted as nodes and shading indicates an observed variable. Edges denote dependencies between variables. The box in the plate diagram represents replicates of random variables that are independent and identically distributed. Figure adapted from [1].

The latent components captured by a model can be used to create hypotheses about the underlying mechanisms that they are assumed to represent. For example, in gene expression analysis the components can represent functional gene modules and pathways. Based on drug treatment measurements, the components can be used as a hypothesis about the drug mechanisms of action. However, such interpretation can be difficult in practice, if each component is associated with a large number of conditions and genes. To solve this problem, different kinds of *sparsity* priors have been proposed [5, 20, 31, 140, 141]. The general idea is that sparse priors drive the probabilistic weights towards zero, leaving only a subset of weights active at a time. This makes the subsequent interpretation easier, as it can be based on a lower number of variables.

3.2.1 Factor analysis

Factor analysis (FA) is a standard unsupervised data analysis method for capturing and understanding linear relationships between variables [123]. A set of K factors are used to model dependencies between the features in a data matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$:

$$\mathbf{X} = \mathbf{Z}\mathbf{W}^T + \mathbf{E} \,, \tag{3.3}$$

where the columns of $\mathbf{Z}$ are the K latent factors, $\mathbf{W} \in \mathbb{R}^{D \times K}$ contains their loadings, and $\mathbf{E}$ is residual noise. Different factor analysis models can be defined by choosing specific priors and structure for $\mathbf{Z}$, $\mathbf{W}$ and $\mathbf{E}$.

A common factor analysis model is principal component analysis (PCA), where the aim is to identify a set of K orthogonal components that maxi-

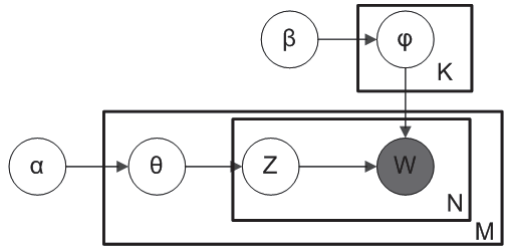


Figure 3.2. Plate diagram for the topic model. Symbols: multinomial topic distributions θ for M documents i with N_i words; multinomial word distribution φ for K topics z; parameters α and β for the Dirichlet priors for θ and φ. Figure adapted from http://commons.wikimedia.org/wiki/File:Smoothed_LDA.png.

mally capture the variance in the data X. The Bayesian PCA is obtained by setting the noise $\mathbf{E}$ in 3.3 to be equal over all variables [9].

3.2.2 Topic models

In many computational data analysis applications data does not originally exist in numerical form. For example, in text analysis the data can be a collection of text documents. The simplest way to analyse the documents is to think of each as a *bag of words*, representing them as non-negative count vectors over the set of words in the document collection. Such data can be modelled with Latent Dirichlet allocation [12], also known as the *topic model*, which assumes that each text document is a mixture of a small number of latent topics and that each word in the document is generated by one of the document's topics. Topic modelling can be viewed as probabilistic PCA of discrete data, and hence the name discrete PCA is also used [13].

A plate diagram representation of the topic model is shown in Figure 3.2, and the generative process goes as follows: The model is first initiated by generating for each $i \in 1, \ldots, M$ document a multinomial distribution θ_i over the topics Z from a K-dimensional Dirichlet distribution $Dir(\alpha)$. Likewise, for each $z \in 1, \ldots, K$ topics a multinomial distribution φ_z over the words W is generated from an N-dimensional Dirichlet distribution $Dir(\beta)$. The word data j for document i is then generated by drawing a latent component $z_{i,j}$ from the multinomial distribution θ_i, and then drawing a word $w_{i,j}$ from φ_z. This is repeated until all words for all documents are generated.

3.2.3 Multi-view models

In many practical applications multiple data sources are available providing complementary information about the system under study. Examples mentioned in Chapter 2 include combination of gene expression and protein interaction data, and drug treatment gene expression and toxicological profiles. In such cases it would be sensible to define a joint probabilistic model for all available data.

In the machine learning field, learning from multiple data sources with shared samples is called *multi-view* learning, where a view refers to a single data set. A classical multi-view method for studying dependencies between two data sources is *canonical correlation analysis* (CCA) [52, 57]. Given two data matrices $\mathbf{X}_1 \in \mathbb{R}^{D_1 \times N}$ and $\mathbf{X}_2 \in \mathbb{R}^{D_2 \times N}$ with N shared measurement samples, the task is to find linear projections $\mathbf{U} \in \mathbb{R}^{D_1 \times K}$ and $\mathbf{V} \in \mathbb{R}^{D_2 \times K}$ so that the correlation between $\mathbf{u}_k^T \mathbf{X}_1$ and $\mathbf{v}_k^T \mathbf{X}_2$ is maximised for the components k. The components are additionally forced to be uncorrelated. The solution can be found by analytically solving a set of eigenvalue problems.

Also Bayesian versions of CCA have been introduced [69, 70, 132, 136]. One effective solution is to formulate the problem as a factor analysis model (3.3) with a specific block structure in the factor loadings $\mathbf{W}$ corresponding to shared and view-specific components [70, 132]. The Bayesian CCA was recently extended to handle an arbitrary number of data views. The resulting method, called *group factor analysis* (GFA) [133], is a generalisation of the factor analysis -type modelling of dependencies between variables to dependencies between data sets. The central assumption is that variables are active or inactive in groups, matching to the data views.

For GFA, the $\mathbf{X}$ in (3.3) represents a collection $\mathbf{X}_1, ..., \mathbf{X}_M$ of M views, with shared samples and dimensionalities $D_1, ..., D_M$. The modelling task is to identify K factors that describe the structure and dependencies between the views $\mathbf{X}_m$. The likelihood for observed data $\mathbf{X}$ is thus

$$p(\mathbf{X}|\mathbf{Z}, \mathbf{W}, \boldsymbol{\tau}) = \prod_{m=1}^{M} \mathcal{N}(\mathbf{X}_m|\mathbf{Z}\mathbf{W}_m^T, \boldsymbol{\tau}_m^{-1}\mathbf{I}) \; . \tag{3.4}$$

The noise $\mathbf{E}$ in (3.3) is now set to diagonal $[\boldsymbol{\tau}_1^{-1}, ..., \boldsymbol{\tau}_M^{-1}]$ with each $\boldsymbol{\tau}_m^{-1}$ repeated D_m times. A Gamma prior is used for the inverse variances $\boldsymbol{\tau}_m$, and the factors $\mathbf{z}$ are assumed to be normally distributed with zero mean and unit covariance:

$$p(\boldsymbol{\tau}_m) \sim \mathcal{G}(a^\tau, b^\tau) \tag{3.5}$$

$$p(\mathbf{z}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \; . \tag{3.6}$$

In GFA, the weight matrix $\mathbf{W}$ is made group sparse with a automatic relevance determination (ARD) prior that is specific to each pair of data view m and component k:

$$p(\mathbf{W}|\alpha) = \prod_{k=1}^{K} \prod_{m=1}^{M} \prod_{d=1}^{D_m} \mathcal{N}(\mathbf{w}_{m,k}(d)|0, \alpha_{m,k}^{-1}) \tag{3.7}$$

$$p(\alpha_{m,k}) \sim \mathcal{G}(a^\alpha, b^\alpha) \; . \tag{3.8}$$

where k indexes factors, m views and d dimensions. The inverse variance of each $w_{m,k}$ is controlled by the parameter $\alpha_{m,k}$ with a Gamma prior. The hyperparameters a^τ, b^τ, a^α and b^α are set to very small values.

The ARD prior in GFA makes groups of variables, matching to the data views, inactive for specific factors by forcing their $\alpha_{m,k}^{-1}$ to zero. This group-wise sparsity results in factors that are active in only a subset of the data views. GFA thus effectively separates effects that are specific to certain views from shared effects. This group-sparsity assumption is the key difference of GFA compared to earlier factor analysis models. Inference for the GFA model is carried out with a variational approximation, using the R package CCAGFA available in CRAN [133]. The detailed equations are provided in the Appendix of Publication III.

In this thesis, GFA is used for modelling drug treatment experiments from multiple cell types in Publications III and IV, covered in Section 5.2.

3.3 Probabilistic models for gene expression data

In this section, probabilistic models for different data analysis problems in molecular biology are reviewed. Focus is on different types of models for gene expression modules and models that combine multiple data sources. Also model-based retrieval of gene expression experiments is reviewed.

3.3.1 Modelling gene clusters

Common clustering methods of gene expression data were discussed in Section 2.1.3. Also probabilistic latent variable models have been used extensively for gene expression data analysis, providing different kinds of clusterings for genes [2, 43, 71, 102, 110]. A Bayesian variant of a

standard biclustering method, called plaid model, was proposed [18] and shown to outperform hierarchical clustering in clustering gene expression profiles. A hierarchical version of biclustering was recently introduced that uses a nonparametric Bayesian prior for automatically inferring the numbers of biclusters at each level of the hierarchy [19]. One central benefit of the probabilistic component models is that each data point can belong to multiple components, whereas in standard clustering methods the data points are assigned to a single cluster.

Also the probabilistic topic models have been applied to analyse large collections of gene expression data [16]. In this application each gene expression experiment was considered as a document in the topic model terminology. The differential expression profiles were processed with Gene set enrichment analysis (GSEA) [118] to reduce the dimensionality and bring in prior knowledge in the form of gene sets. The GSEA output was then quantised and the resulting non-negative count vectors for each experiment were considered as the word data for the topic model. The method identified biologically relevant components from the ArrayExpress database [90]. Compared to standard clustering methods, the probabilistic model allows both genes and conditions to belong to multiple components. The probabilistic topic model was used as basis for retrieval of relevant experiments, as described in more detail in Section 3.3.3.

3.3.2 Modelling multiple data sources

Probabilistic models have been proposed for combining gene expression with various other data types [86, 87, 98, 101, 104, 107]. As discussed in Section 2.1.4, a highly interesting task is to combine gene expression and protein interaction data to search for functional gene modules, and many probabilistic models have been proposed to solve this task. A model by Segal *et al.* [103] combines a mixture of Gaussians model for the gene expression profiles with a Markov random field model for the protein interactions into a joint probabilistic model, which was then used to discover coherent functional gene groups and protein complexes. Shiga *et al.* [108] extended the Markov random field model into a *hidden modular random field* (HMoF), seeking clusters of genes with high network modularity and similar gene expression profiles. Lahti *et al.* [76] proposed a method for identifying local, connected regions in the protein interaction network with a coherent transcriptional response in a subset of experimental conditions, such as tissues. In this thesis, a novel probabilistic

model extension for identifying functional gene modules from protein interaction and gene expression data was introduced in Publication I and is described in more detail in Chapter 4.

The probabilistic multi-view models are applicable in cases where multiple measurement types are available from the same samples. For example, gene expression and DNA copy number from the same samples can be modelled with CCA [77]. Huopaniemi *et al.* [61] used a Bayesian CCA variant to combine metabolic profiles from two tissues. Also other CCA variants have been applied to such settings with multiple genomic data sources [81, 143, 145]. Combining the drug treatment gene expression data with other data sources is a promising direction for drug discovery and development. For example, The Connectivity Map gene expression data has been used in combination with chemical descriptors to infer specific relations between chemical structures and molecular mechanisms, first with non-probabilistic CCA [68] and later with GFA [133].

3.3.3 Model-based retrieval of gene expression experiments

Online databases of gene expression measurements are growing rapidly [8, 90]. It would be useful for many researchers to compare their experiments to those made by others, but the problem is how to find the most relevant ones from large databases. The traditional way of solving this information retrieval task is to use manual annotations, such as disease or drug treatment labels. However, such annotation-based retrieval methods depend on the goodness of the annotations, and will miss any novel similarities.

An alternative approach is content-based retrieval, where the measurement data is used directly to infer relevance in some way. This has the benefit of not being restricted to user-generated annotations and can detect novel similarities if the data supports them. For example, gene expression experiments can be searched by giving as input a list of genes of interest, with the aim of finding other experiments where the same genes are differentially expressed. Connectivity mapping [78], described in Section 2.2.3, is an example of a such content-based retrieval method, where gene signatures are used to match drug treatment gene expression measurements to other drugs or diseases. In related work, a disease-drug network was constructed by computing similarities between disease and drug experiments [58].

Retrieving of relevant gene expression measurements from public databases has also been approached in a model-driven fashion [16]. Using a

probabilistic model, the relevance can be defined based on relevant effects detected by the model. Model-based retrieval is also more commensurable across measurements from various platforms than using raw data values. In practice, the topic model was used to capture relevant components from a collection of gene expression measurements in the ArrayExpress database [90]. Relevance for the retrieval was then be computed based on the model, as

$$rel(q|r) = P(x_q|\Psi_r) = \prod_{x \in x_q} \sum_{t=1}^{T} \theta_{r,t}\varphi_{t,x} \ . \tag{3.9}$$

In other words, the relevance of the query q to sample r is the probability of r to generate the query data x_q. The probability is computed using point estimates of the topic model parameters $\Psi_r = \{\theta_{r,t}, \varphi_{t,x}\}$, where $\theta_{r,t}$ and $\varphi_{t,x}$ are the multinomial distributions over the topics and words, respectively.

Later, a fully Bayesian variant of the relevance measure was proposed [17], integrating over the posterior of the latent variables Ψ:

$$rel(q|r) = \int_{\Psi} P(x_q|\Psi_r)P(\Psi_r|X)d\Psi \ . \tag{3.10}$$

In this thesis, contributions to probabilistic model-based retrieval of relevant gene expression data are presented in Section 5.2 using group factor analysis as the model.

4. Probabilistic components of molecular interactions

In this chapter, contributions are presented on applying probabilistic models for functional gene module analysis. The problem is first introduced and the specific contributions in Publication I are then described, covering the application and extension of the interaction component model [112, 113] to the task of finding functionally coherent gene modules from combinations of protein interaction and gene expression data.

4.1 Integrating gene expression data with protein interactions

The task of finding groups of genes with similar functions is very important in molecular biology for predicting functions for unknown genes and for understanding molecular mechanisms of drugs and diseases. A promising approach is to combine gene expression and protein interaction data for finding functional modules with high gene expression similarity and high number of interactions between the molecules. A lot of methods have been presented for this task, as described in Sections 2.1.4 and 3.3.

In Publication I a probabilistic model for network data, called *interaction component model* (ICM), was applied to searching for tightly interconnected clusters from protein interaction networks and further extended to incorporate gene expression data for improved module discovery. It was compared to two alternative clustering methods that use both expression and protein interaction data, called Matisse [128] and Hidden modular random field (HMoF) [108].

Matisse combines gene expression and protein interaction data by first transforming in the expression profiles into similarity values and then seeking connected sub-networks based on both expression similarity and protein interaction data [128]. The Matisse method was shown to outperform both Co-clustering [51] and CLICK [105]. The HMoF method for-

mulates a joint probabilistic model based on a Markov random field and network modularity [108].

Matisse is a non-probabilistic method and does not involve any kind of a noise model for the interaction data, which makes it sensitive to noise in the input data. However, Matisse does not need to include all genes into the clusters if they do not seem to fit any of them, which can be beneficial as not all genes participate in any functional modules. HMoF in turn does contain a probabilistic noise model for both interactions and expression data, providing robustness against noise. Both Matisse and HMoF assign each gene to exactly one cluster. This is a serious limitation, as many genes and proteins are known to participate in multiple functions. The probabilistic ICM, presented in Publication I, can cope with noisy input data and additionally provides overlapping gene modules, better matching the underlying biological phenomena.

4.2 Interaction component models for protein interaction and gene expression data

The Interaction component model (ICM) was designed to capture tightly interconnected clusters of nodes, and it has earlier been applied to finding communities from large social networks [112, 113]. In Publication I, ICM was used to search for interconnected clusters from protein interaction networks, the clusters being then interpretable as functional gene modules. Additionally, two extensions to the model were introduced for integrating gene expression data into the analysis. A one-to-one matching between genes and proteins was assumed, and the terms are thus treated interchangeably.

ICM defines a generative probabilistic model for network data. It assumes that each interaction in the network, or edge in a graph, is generated from a latent component. A component is defined as a probability distribution over the nodes in the graph. A plate diagram of the model is shown in Figure 4.1A. In detail, the links are generated by first drawing the component z from the multinomial distribution θ, and then drawing the end nodes i and j of the link from the multinomial distribution ϕ_z of the component z. In the generative process each link belongs to exactly one component, while nodes may belong to several and thus the model allows for overlapping clusters.

In the first extension the gene expression data was transformed into

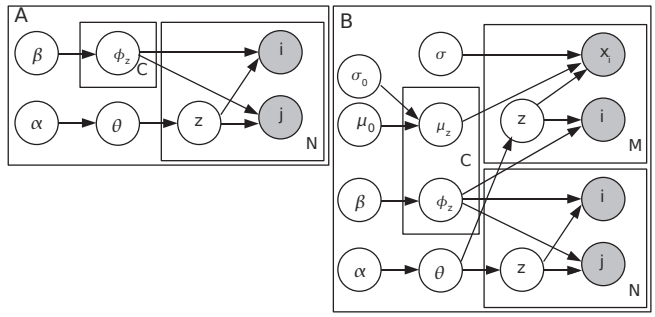


Figure 4.1. Plate diagrams for interaction component models. **(A)** Plain ICM. Each interaction is generated from a component z by sampling the end points i and j from the multinomial distribution ϕ_z. **(B)** Extension (ICMg2) with gene expression data (the bottom part of the plate is the same as in A). Node data x_i are generated from the same components z as the interactions, by sampling each node i from ϕ_z and then its gene expression profile from a component-specific Gaussian distribution $\mathcal{N}(\mu_z, \sigma^2 I)$. The component-specific means μ_z have a Gaussian prior $\mathcal{N}(\mu_0, \sigma_0^2 I)$. There are in total C components, N links and M nodes. Figure adapted from Publication I.

additional similarity-based links and pooled together with the protein interactions. The idea is that highly correlated genes are likely to be co-regulated and thus involved in the same functional processes. In practice, the Pearson correlation was computed between each pair of genes, and a link was added between those pairs with correlation exceeding 0.85 in the protein interaction network. This extension was denoted as ICMg1, g referring to genes.

The second model variant extends the generative process of the ICM to incorporate gene expression data for the nodes. The assumption is that the components represent modules that are both strongly interconnected and similar in terms of gene expression. The plate diagram for this extended model is shown in Figure 4.1B. In practice, for each node the component z is first generated from the same distribution θ_z as the link components, and finally the gene expression profile x_k is sampled from the component-specific Gaussian distribution. For computational simplicity the fact that each node has exactly one gene expression profile was not included as a constraint. This extension was denoted as ICMg2.

Model inference for ICM is carried out with collapsed Gibbs sampling, where some of the variables are integrated out. In short, the assignments of the data points to the latent components z are sampled one data point at a time, holding all other assignments fixed, providing a reasonably simple and fast sampling scheme. Details are given in Publication I.

4.3 Results

The three ICM variants were applied to a collection of protein interaction data and two gene expression data sets, resulting from osmotic shock response and DNA damage experiments from yeast *Saccharomyces cerevisiae* [128]. Two comparison methods for combining the data types were included: HMoF [108] and Matisse [128]. All methods provided as output a set of gene clusters, which were then compared to external ground truth data to evaluate their biological relevance. The number of clusters for ICM variants and HMoF was set to the median of 20 Matisse runs. The relative weight for HMoF between the expression and network data was fixed to $\omega = 0.2$ as suggested in the original paper [108]. Matisse was run with default parameters.

The ground truths used in the evaluation were Gene Ontology (GO) [6], a hierarchical classification of known gene functions, and known protein complexes from the Comprehensive Yeast Genome Database [46]. First, perplexities were computed for predicting the standard gene classes derived from the GO. Second, standard hypergeometric test-based GO enrichment analysis was performed. Finally, the degree of coverage of the identified clusters to known protein complexes was computed.

The results showed that the three interaction component models outperformed the alternatives. Matisse performed clearly worse than the other methods, while HMoF was in some cases equally good as the ICM methods. From the ICM variants, the ICMg1 was in general slightly better than the two alternatives, but the plain ICM without the gene expression data performed surprisingly well. Additionally, the ability of the ICM methods to detect overlapping modules based on network data was demonstrated on an artificial data set with a known ground truth.

4.4 Discussion

Publication I introduced generative probabilistic models for functional module discovery from combinations of gene expression and protein interaction data. The interaction component model and its extensions outperformed a representative set of earlier methods for the task. The difference was clear to the non-probabilistic method Matisse that lacks a noise model for the interaction data, but smaller to the other probabilistic method, HMoF, indicating that the probabilistic formulation is beneficial in coping

with the highly noisy input data. Matisse has later been updated with an improved probabilistic model for the subnetwork connectivity [129], outperforming the earlier Matisse version.

Another key advantage of the ICM formulation is that the genes or proteins can belong to multiple, overlapping modules, reflecting their biological nature to participate in several functions. In Publication I this was demonstrated with artificial data, and future work could address this in real data, validated with a suitable ground truth. Since Publication I, other methods capable of detecting of overlapping modules have also been introduced, and a comparison would be interesting [83].

In its current form ICM requires the number of components to be specified beforehand, whereas some methods, such as Matisse, can estimate the number of clusters automatically. A natural extension for the ICM would be to use the Dirichlet Process [121], a common non-parametric prior for multinomial distributions for estimating the number of components from the data. Another possible improvement would be to add an explicit model for inter-cluster links, as ICM currently only models intra-cluster links.

In the current experiments, protein interactions seemed to be more informative than gene expression data, with the addition of gene expression providing only minor improvement over the performance of the ICM applied to the network data alone. This is consistent with earlier studies [87, 108], suggesting that functionally related proteins are more likely to interact together than to show highly similar gene expression profiles. Another possible reason is that the ground truth data, such as Gene Ontology or protein complex information, are biased towards protein interactions. In the current ICMg2 formulation, the interaction data gets more weight as each link provides one count for the node, whereas its gene expression profile only acts as one count. This weighting could probably be improved, increasing the relative importance of the gene expression data.

It is also possible that the gene expression data are so noisy and complex that the rather simple modelling assumptions used so far are not able to capture the relevant signals from the data. Transforming the gene expression data into links with a strict cutoff reduces the effect of the noise, but on the other hand discards also a lot of useful information. The assumption that the gene expression profiles should be similar across the whole genome, underlying all methods discussed here, is quite strict as gene expression can vary over conditions. A more structured probabilistic model

could be used to identify this condition-dependency, and would probably improve the functional module discovery performance also. Such a model was introduced by Lahti *et al.* [76], successfully identifying subnetworks with context-specific transcriptional responses. Such models could also be used to infer condition-dependency for the protein interactions, as discussed by Segal *et al.* [103].

5. Probabilistic components of drug responses and toxicity

In this chapter, contributions are presented on applying probabilistic component models to molecular drug response and toxicity analysis. Two application problems, toxicogenomics and drug connectivity mapping, are addressed with introductions and contributions on each problem. In Publication II, a component model is used for associating drug treatment gene expression measurements with corresponding toxicological outcomes, providing novel understanding of the genome-wide toxicogenomic responses in human cancer cell lines. In Publications III and IV, a probabilistic model-based data retrieval method is developed and applied to drug connectivity mapping, providing improved search method for relevant drug treatment experiments.

5.1 Probabilistic toxicogenomics

Toxicogenomics requires integration of multiple heterogeneous data sources and thus provides interesting challenges for methods development. As discussed in Section 2.2.2, the main obstacles for the development of toxicogenomics are lack of suitable data sets and sophisticated models that can distinguish the toxicology-associated signals from the noisy data. In Publication II, these problems were addressed by integrating the Connectivity Map drug treatment gene expression data [78] with toxicological profiles from the NCI60 Cancer cell line screen database [111], providing the so far largest toxicogenomics data set. Probabilistic modelling was then applied to identify toxicity-associated gene expression response patterns. These patterns were characterised in terms of underlying biological and toxicological mechanisms, and their predictive power was validated with independent high-throughput cell line screening data and with external human primary hepatocyte data from the TG-GATEs database [127].

5.1.1 Constructing the predictive toxicogenomics space

A novel toxicogenomic resource for studying the associations between gene expression and toxic responses was constructed by combining the drug treatment transcriptional response data from the Connectivity Map (CMap; [78]) with the NCI60 Cancer cell line screen data [111]. From the CMap, 3062 measurement *instances*, that is, chemical and cell line pairs, were included from the microarray platform HT HG-U133A for 1217 unique chemicals on three cell lines: MCF7 (breast), PC3 (prostate) and HL60 (leukaemia). Toxicological profiles from the NCI60 database were then obtained for a subset of 492 instances for 222 unique chemicals on the same cell lines, containing GI50 (50% growth inhibition), TGI (total growth inhibition), and LC50 (50% lethal concentration) values describing the toxic outcomes from the drug treatments. The matching between the two data sets was done based on chemical names.

Probabilistic component modelling was first applied to detect a set of robust drug treatment gene expression response components, following the method described by Caldas *et al.* [16]. Briefly, Gene set enrichment analysis (GSEA; [118]) was first applied to reduce the high dimensionality of the data and to incorporate prior knowledge, using 1321 curated gene sets from the Molecular Signature Database [118]. The GSEA output values for each instance and gene set-pair were then quantised, and the probabilistic decomposition was performed with the Latent Dirichlet allocation component model [12], earlier applied to topic modelling as described in Section 3.2.

The component model assumes that each instance i has a probabilistic assignment vector $p(z|i)$ to the components z, and each component has a probabilistic assignment vector $p(gs|z)$ to the gene sets gs. Thus, each instance and gene set can be associated with multiple components, following the polypharmacology assumption [56, 67]. Each component then represents a specific chemical-induced transcriptional response pattern, active for a subset of chemicals and cell lines. The components can be interpreted as biological processes through the gene set activities, and the assumption is that some of these capture the toxicity-associated responses. The number of components was set to 100 based on external functional similarity data of the chemicals. The inference on parameters was carried out with collapsed Gibbs sampling.

After identifying the response components from the full CMap data, the

components were associated to the toxicological profiles available for the subset of 492 instances. As the toxicological outcomes from the CMap drug treatments are a result of both the chemicals intrinsic toxic potential (GI50, etc.) and the dose with which the treatment was carried out, a concentration-dependent toxicity T_i for instance i was defined as the difference between the CMap concentration D and the $GI50$ values: $T_i = \log_{10} D - \log_{10} GI50$. For each component z, the mean concentration-dependent toxicity value T_z was computed as $T_z = \sum_i [p(i|z) \cdot T_i]$, where the probability of the instances to belong to the component z is computed as $p(i|z) = p(z|i) / \sum_{i'} p(z|i')$.

A subset of the 100 components was then selected as the Predictive toxicogenomic space (PTGS). The predictive performance was validated based on the 492 instances with GI50 values available using a cross-validation procedure to avoid overfitting. The 14 components, labelled A-N, with the highest performance in concentration-dependent toxicity prediction, were selected as the PTGS components as a tradeoff between interpretability and predictive power. The probability of an instance to belong to the 14 PTGS components was thereafter used as a toxicity-predictive score.

To characterise the PTGS, the top instances and genes were identified for each of the 14 PTGS components. The top genes were identified by evaluating the differential expression of each of the genes in the top gene sets, as given by $p(gs|z)$. Based on the top genes, biological and toxicological characterisation was performed using Gene Ontology [6] enrichment analysis (R package topGo; [4]) and Ingenuity Pathway Analysis.

5.1.2 Results

The constructed novel toxicogenomics dataset revealed a strong correlation between concentration-dependent chemical toxicity and transcriptional variation. The identified PTGS components covered primarily the transcriptional responses resulting from measurements at toxic doses. The PTGS components reflected responses induced by a broad set of chemical classes and mostly shared across the cell lines. The genes associated with the PTGS components showed an expected enrichment of various biological and metabolic processes and transcription factors associated with growth inhibition, and cellular toxicity and stress pathways. Ingenuity Pathway Analysis revealed toxicity effects in major internal organs affected by adverse drug reactions. The PTGS-associated genes included also many novel genes and transcription factors for toxicity prediction

The ability of the PTGS score to predict concentration-dependent toxicity was evaluated for a set of independent toxicological profile measurements for the CMap instances. The predictive performance of the PTGS scores was compared to partial least squares quantitative structure-activity relationships (QSAR) -based approach. The PTGS resulted in significantly higher predictive performance than QSAR, confirming the predictive power of the PTGS. Finally, the PTGS was evaluated outside the CMap data set, using the TG-GATEs database [127] of drug treatment measurements hepatocyte cells on human *in vitro* and rat *in vitro* and *in vivo*. In the human data set, the PTGS scores showed clear concentration-dependent behaviour, suggesting that the PTGS score is applicable also outside the three CMap cancer cell lines. However, for the rat data the concentration-dependent behaviour was not visible, suggesting that the PTGS is human-specific.

5.1.3 Discussion

In Publication II, a novel toxicogenomics dataset was constructed, revealing a strong correlation between the chemical toxicity and the chemical-induced transcriptional variation that has not been shown in this scale earlier. A probabilistic component model was used to identify a set of robust toxicogenomic response components that were then further characterised and shown to exhibit high concentration-dependent toxicity-predictive power. This study is the first to provide a broad view to the dose-dependent toxicogenomic responses. Remarkably, the effects are mostly shared across cell lines and further extend to hepatocyte cells as well, though being apparently human-specific.

Given the overlapping chemicals between CMap and NCI60, an alternative approach would have been to use some statistical method to identify genes that are predictive of the toxicological profiles. However, such an approach would likely provide a lot of false positives. Instead of analysing individual genes, it was possible to analyse sets of coherently behaving genes with the help of GSEA. The probabilistic model made then possible to associate both drugs and gene sets with multiple response components, reflecting the promiscuous nature of drug responses. Moreover, the probabilistic model allowed the use of the full CMap data to construct more robust gene expression response components.

The results showed clearly that the gene expression data is more informative of the toxicological outcomes than the chemical descriptors. This

is in line with earlier results with the TG-GATEs data [82]. It would be interesting to try combining these data sources in the future and see if better methods could bring any additional benefits from the chemical descriptors. Notably, another study using *in vitro* assays from the ToxCast project did not find a significantly different predictive power as compared to QSAR [122], suggesting that the specific assays did not cover enough toxicity-predictive genes. The top genes associated with the PTGS provide an effective starting point for developing more accurate toxicity-predictive assays.

5.2 Probabilistic drug connectivity mapping

In this Section, a novel method is introduced for retrieving relevant experiments from large collections of gene expression data for multiple cell types. The first contribution is to define the relevance for retrieval based on a probabilistic model. The second contribution is to use a multi-view model for proper data integration in retrieval applications where data from multiple sources are available, carrying partially relevant information for the retrieval task.

In Publications III and IV the novel probabilistic modelling-based data retrieval method was introduced, extending the retrieval of relevant experiments principle introduced by Caldas *et al.* [16] and described in Section 3.3.3. The idea is that a well-chosen probabilistic model can capture relevant response components from the noisy input data, and retrieval is then more accurate based on the model than the original input data.

The method was applied to the drug connectivity mapping task, introduced in Section 2.2.3. Compared to earlier alternatives, the method is better able to capture relevant activity components from the noisy input data, and further explicitly infer which components generalise across the cell types. In Publication III the task is to infer overall similarity for a query with data from all three cell types. In contrast, in Publication IV the query data is given for only one cell type, and the task is to use the information from the other cell types to improve the retrieval performance on the query cell type.

5.2.1 Data retrieval with group factor analysis

A general limitation in current connectivity mapping methods is that they simply aggregate data over all experimental factors, as discussed in Section 2.2.3. They assume that only the responses that are general across experimental factors are relevant, while any specific responses are discarded as noise. However, different cell types often respond differently to the same drugs, at least when considering a broader variety of cell types [7, 40]. In personalised cancer medicine the aim is to find a cure for a specific disease, in which case the focus should be on the responses characteristic to the corresponding cell type. On the other hand, toxicity analysis aims to detect effects that generalise across model organisms and ultimately to humans, and thus the analysis should focus on effects shared across all cell types.

To solve the data integration problem, a probabilistic multi-view model was used. Group factor analysis (GFA) [133], described in Section 3.2.3, provides a suitable decomposition of multiple cell type response data to shared and specific components, representing gene expression response patterns for subsets of the drugs and genes. Combining information from multiple cell types provides more robust components, as noise in the data is by definition specific to individual cell types. Additionally, the latent variable-based probabilistic model is able to deal with the high dimensionality and low sample size, typical in gene expression data analysis.

Details of the GFA model were described in Section 3.2.3. Applied to drug treatment gene expression measurements from multiple cell types, GFA provides latent component activities $\mathbf{z}_i$ for each treatment i and loadings $\mathbf{w}_{m,k}$ for each cell type m and component k. Due to the group-wise ARD prior, the loadings are made sparse on the cell type level, resulting in decomposition of the data where components are active in all or only a subset of the cell types. This is a key novelty of the GFA-based retrieval drug connectivity mapping method, compared to earlier alternatives which consider only the general responses to be relevant. Currently GFA uses simple Gaussian distributions to model the gene expression data, which seems to work well based on the good retrieval performance, but also other distributions could be used, possibly improving the performance further.

Given the GFA model, retrieval can be performed based on the shared or specific components, or both, depending on the application and user

interests. A suitable relevance measure between drugs i and j is the Pearson correlation between their latent variables $\mathbf{z}_i$ and $\mathbf{z}_j$ identified by GFA. This measure focuses on the non-zero factors, representing relevant activity for the query. For a new chemical-treatment sample outside the existing database, the latent variables can be estimated easily. The model-based retrieval method was coined in the work as *probabilistic connectivity mapping*.

5.2.2 Probabilistic drug connectivity mapping results

In Publication III the probabilistic connectivity mapping method was applied to the Connectivity Map (CMAP) [78] gene expression data with three cell lines. The performance of the method to retrieve functionally and chemically similar drugs was compared to alternative, rank- and correlation-based connectivity mapping methods [62, 78]. Additionally, alternative probabilistic model formulations were evaluated to study the benefits from the multi-view data integration capability of GFA, including sparse factor analysis with feature-wise sparsity and Bayesian PCA with a shared sparsity prior across all dimensions.

From the CMap chemicals, a subset of 718 chemicals measured with the HT-HG_U133A platform on all three cell lines were used. Instead of the full genome, a subset of 930 Landmark genes identified in the LINCS project were used. Retrieval performance was measured against two ground truth sets of functional and chemical drug similarity: shared Anatomic Therapeutic Chemical (ATC) classification codes and Tanimoto similarity. Two retrieval performance measures were used: partial area under the ROC curve and top-10 mean average precision.

The results showed clearly how the probabilistic connectivity mapping with GFA and sparse FA outperformed the other methods on both ground truths and goodness measures. Bayesian PCA performed clearly worse than the sparse models, confirming the benefits from the sparsity assumptions for capturing relevant activity from noisy data. Additionally, the shared GFA components were shown to be relatively more important than the cell line-specific components from either GFA or sparse FA. This indicates that the data integration approach that separates the shared and specific effects is beneficial for connectivity mapping.

5.2.3 Cross-organism toxicity analysis results

A central limitation in toxicogenomics is the lack of direct chemical treatment measurements from humans, as discussed in Section 2.2.2. One solution to this problem has been to use connectivity mapping -type retrieval methods on gene expression data measured on various model organisms, and interpreting the results based on known drug toxicity effects on humans [15, 144].

So far the connectivity mapping methods for toxicity analysis have allowed the use of only one cell type at a time, whereas databases such as TG-GATEs offer data on cells from multiple organisms (human *in vitro*, rat *in vitro* and *in vivo*). As the goal is to predict chemical toxicity in humans, it would make sense to integrate the responses across the cell types and especially identify those responses that are conserved across the organisms, as those are most likely to generalise to humans as well.

Toxicity analysis thus offered another interesting application for evaluating the benefits of probabilistic connectivity mapping with group factor analysis. This was studied in Publication IV with the TG-GATEs data. As ground truth for evaluating the retrieval performance, the drug-induced liver injury (DILI) label and concern classes [22] were used, representing liver-related toxicity risk information for drugs already in clinical use. Additionally, the ATC codes were used to give more detailed information about the drug mechanisms of action.

GFA was applied to identify shared gene expression responses across the three model organisms. The GFA model was extended with additional element-wise sparsity on both the factors $\mathbf{Z}$ and factor loadings $\mathbf{W}$, assuming that the identified factors represent drug response patterns specific for a small number of drugs and genes. Details of the GFA variant are given in Publication IV. The measurements from the human *in vitro* hepatocyte cells were used for retrieval, reflecting a practical connectivity mapping use case where the treatment effects for a novel chemical are measured in only a single cell type. The performance of the probabilistic connectivity mapping with shared components to retrieve relevant chemicals in terms of the ground truth data was evaluated against the standard connectivity mapping methods. Retrieval performance with the shared components was indeed better than with earlier methods using only a single data source, suggesting that learning and using the cross-organism responses is suitable for connectivity mapping-based toxicity analysis.

5.2.4 Discussion

Publications III and IV introduced group factor analysis-based data retrieval method, called *probabilistic connectivity mapping*, with two main contributions: First, using a suitable probabilistic model it is possible to capture the relevant drug response components from the noisy gene expression data and use these components to define the relevance in the connectivity mapping task. Second, using a model capable of integrating the data from multiple available cell types, and especially learning which responses are specific to cell types and which are shared by two or more of them, improves the retrieval performance.

In Publication III probabilistic connectivity mapping was shown to outperform alternatives in finding functionally and chemically similar drugs from measurements over multiple cell lines. In Publication IV the method was shown to improve toxicity-related retrieval results for a single-cell query by using shared responses identified from a database with multiple cell types. The multi-view model-based retrieval method thus provides a promising direction for methods development for drug repositioning and toxicity analysis applications.

Probabilistic connectivity mapping gives the user the explicit choice of focusing on the shared or specific responses, making it applicable in both personalised medicine and general toxicity analysis. In both Publications III and IV, the shared components were shown to be more informative, but also some of the cell line-specific components contained some relevant information. With the number of available measurements growing in the future data collections, such as LINCS, the benefits from the multi-view decomposition and especially the shared components are expected to become even more apparent.

In the current works the model-based relevance was defined in a very simple manner as the Pearson correlation between point estimates of the latent variables. As discussed in Section 3.3.3, the relevance could be formulated in a more Bayesian manner in terms of probabilities with either point estimates as $p(x_q|z_i)$, or over the full posterior $p(x_q|x_i) = \int p(x_q|z_i)p(z_i|x_i)dz$. Both alternatives were studied also for probabilistic connectivity mapping, but the retrieval performance turned out to be inferior to the simple correlation. One possible reason for this is that the distributional assumptions made in GFA do not fully match to the data, and thus the probability-based relevance measures that take the partly

misspecified uncertainty into account make things worse, which is something to consider in future extensions.

6. Model-based graph visualisation

In this chapter, contributions are presented on applying probabilistic models for information visualisation. In particular, a novel model-based method is introduced for graph visualisation, which is an important exploratory tool in many biomedical and other applications. The method enables the visualisation to focus on relevant global structures in the data, in contrast to earlier graph visualisation methods that focus on local properties.

In Publication V a probabilistic latent variable model for graph data was first used to capture important link distributions in the data, and then combined with a recent information retrieval-based visualisation method to provide a two-dimensional graph layout. The result is a novel graph visualisation method that can reveal complex structures in large graphs, as demonstrated with word-adjacency graphs.

6.1 Graph visualisation

In many problem domains, data can naturally be represented as a network, where nodes represent objects and edges represent relationships between the objects [44]. For example, proteins and other molecules interact to form complex networks that drive cellular actions, as discussed in Chapter 2. Also interactions between humans can be represented and analysed as social networks.

A popular way to explore and analyse network data is visualisation, which can often convey more information than purely numerical data. A network or graph visualisation is defined by the layout of the nodes and edges in two dimensions. Different visualisation methods focus on different aspects of the layout. Traditionally, graph drawing has been formulated as producing the layout "according to some generally accepted aesthetic criteria" [38]. Algorithms that focus on local aesthetic criteria

may produce visually nice graphs, while completely ignoring any global structure underlying the graph data.

A common aesthetic criterion for a graph drawing has been to position nodes connected by an edge close to each other, but not so close that they would overlap. This principle is followed by the force-based graph drawing algorithms [27, 38, 55]. Traditional force-based methods are very slow for large graphs, and thus faster variants have been proposed [48, 135]. Another family of graph drawing algorithms with essentially the same goal are spectral methods [50, 73]. They operate based on the spectral decomposition of some matrix derived from the graph, and are much faster than force-based methods, but tend to produce layouts where many nodes overlap [47].

An alternative to aesthetic goal-based graph drawing is to focus on interesting global structures, such as clusters of tightly interconnected nodes. An example of a graph drawing method aiming to show graph clusters is LinLog introduced by Noack [88]. The author shows that this task is in conflict with the traditional aesthetic goals, and produces highly different kinds of layouts. As discussed in Chapter 2.1.3, clusters of similar genes are also studied a lot in molecular biology, and thus a lot of graph clustering and visualisation methods have been applied to analysis of biological networks [37, 41].

6.2 Probabilistic model-based graph visualisation

The novel probabilistic model-based graph visualisation method introduced in Publication V consists of three key steps: 1) using a suitable latent variable model for capturing essential structure from graph data, 2) computing distances between the nodes in the graph based on the model, and 3) using a non-linear dimensionality reduction method to visualise the data in two dimensions.

As a probabilistic generative model for graphs, the Simple social network LDA (SSN-LDA) was chosen. It is a modification of the topic model described in Section 3.2 for graph data. In SSN-LDA, the nodes are associated with a membership distribution over latent components, and each component is in turn associated with a distribution over the nodes in the graph. Edges are generated by first drawing a component for the starting node and then drawing the receiving node from the component-specific distribution. The components thus capture specific link distribution pat-

terns in the graph, representing nodes that link to the same set of other nodes.

Given the learnt SSN-LDA model for the graph, a similarity matrix is computed between the nodes based on their probability distributions over the components. As a similarity measure between the probability distributions p_i and q_i over the latent variables i, Hellinger distance was used:

$$d_H(p,q) = \sqrt{\sum_{i=1}^{n} (\sqrt{p_i} - \sqrt{q_i})^2} \,. \tag{6.1}$$

It has been shown to be useful for topic models [11]. Finally, the similarity matrix is used for input for a non-linear dimensionality reduction method to produce a two-dimensional visualisation. Here a recently introduced method called *neighbour retrieval visualiser* (NeRV) [131] was used. NeRV offers a principled way to control the tradeoff between precision and recall in information visualisation and has been shown to outperform other non-linear dimensionality reduction methods.

Compared to the earlier graph visualisation methods, the proposed method allows an explicit choice of what aspects the visualisation should focus on, by choosing a suitable generative model. This allows the visualisation to focus on global structures such as clusters, and by changing the model the visualisation could be made to focus on other things. SSN-LDA was chosen here as it can capture both tightly interconnected clusters of nodes and also clusters with more complicated linking patterns.

6.3 Graph visualisation results

In Publication V the graph visualisation method was applied to a set of graphs with various underlying structures that match to known ground truth labels. The method was compared to Walshaw's force-based method [135], Kruskal's spectral method [73], and Noack's LinLog [88].

The methods were first applied to a graph where nodes represented football teams in different conferences, and edges represented games between the teams. The teams in the same conference typically played more often against each other than teams from other conferences. Both LinLog and the model-based method were able to capture a cluster structure matching very well to the conferences. Neither the force-based nor the spectral method showed any clear clusters, but teams in the same conferences were still mostly positioned close to each other.

Next, word-adjacency graphs with more complex structure were studied. In these graphs, nodes represent words and they are connected with an edge if they occur frequently next to each other in text. Word-adjacency graphs have been shown to exhibit *disassortative* structure [85]: Words from the same word classes, such as noun, typically appear more often next to words from different classes than words from the same class. A small graph consisting of only adjectives and nouns was first studied. A larger word-adjacency graph based on seven novels by Jane Austen was also constructed, using adjectives, nouns, and verbs.

In the word-adjacency graphs the ability of the model-based method to detect also complex linking structures, such as disassortative clusters, became evident. While the three comparison methods failed to show any meaningful structure in the graphs, the model-based method was able to show groups of words that match well to the known classes. This is also confirmed with quantitative validation: The model-based visualisations showed clearly superior k-nearest neighbour classification accuracy for the word classes. In addition to grouping similar words together, the model-based visualisations showed clear edge bundles between the word groups, revealing the underlying linking patterns.

6.4 Discussion

In Publication V a novel graph visualisation method was introduced that is able to capture and show relevant structure in graph data by using a suitable probabilistic model. It was shown to find both tightly interconnected clusters and disassortative structure from the studied example graphs, where earlier graph drawing methods failed. By changing the model, the visualisation could be focused on other things. For example, using the interaction component model used and extended in Chapter 4 instead of SSN-LDA, the visualisation would focus only on interconnected clusters, perhaps performing better than SSN-LDA in that particular task, but would then miss the disassortative structure in the word-adjacency graphs.

The proposed method is analogous to the model-based retrieval method described in Sections 3.3.3 and 5.2 in that a probabilistic model is used to define what is important in the data and then the model output is used to infer similarity or relevance between the data points. These are in fact different views to the similarity structure captured by the model: Re-

trieval focuses on the most relevant results for a given query data point, whereas a visualisation shows the whole data set, trying to keep similar points close to each other. The model-based visualisation method could thus be used for example to show the overall similarity structure of the drugs in the CMap data, and in general the introduced principles should be widely applicable in visualising molecular interactions and other types of biomedical network data.

7. Conclusions

In this thesis contributions were summarised on applying probabilistic component models for analysing molecular interactions and drug responses. The studied applications represent timely problems in molecular biology and medicine. Solving such problems requires advanced computational analysis methods that can cope with multiple noisy data sources with small sample sizes, and here probabilistic component models have been proven highly useful. The publications in this thesis show how the component models can be used in many practical data analysis tasks, such as interpretation, prediction, retrieval and visualisation.

Identifying functionally coherent gene modules from combinations of gene expression and protein interaction data is an important task for predicting functions for unknown genes and proteins and identifying molecular response mechanisms activated by diseases and drug treatments. In Publication I, a probabilistic interaction component model was applied to this task, outperforming earlier alternatives. The joint probabilistic model was able to effectively integrate the two noisy data sources, and furthermore provide overlapping gene modules.

With probabilistic component models, it is also possible to model high-throughput molecular profiling measurements from drug treatment samples and provide novel understanding and predictive models. In Publication II, a component model was used to identify and associate drug response patterns from gene expression data with toxic responses. This so far largest toxicogenomic study provided both detailed hypotheses for molecular response mechanisms leading to chemical toxicity and also high performance in toxicity prediction.

Probabilistic components identified from drug response data can also be used for improved retrieval of functionally and chemically similar drugs in drug connectivity mapping tasks, as was shown in Publications III and

IV. Moreover, using a suitable probabilistic multi-view model it is possible to separate drug responses shared across cell types and even organisms from specific responses, which is crucial for developing connectivity mapping methods for personalised medicine applications. The model-based similarities can also be used for visualisation of interesting latent structures in the data, as demonstrated in Publication V.

The research summarised in this thesis was motivated mainly by applications in molecular biology and medicine, where the development of measurement technologies provides ever growing data sources with interesting analysis challenges. Similar development can however be seen in many other fields, and thus also the models applied and developed in the thesis would be broadly applicable outside the biomedical domain. For example in computational social sciences, the probabilistic component models could be used to understand underlying patterns in social systems and predict future events.

The work presented in the thesis also opens interesting future research directions. Several of the publications involved integration of multiple data sources, with promising results. This could be taken a step further by for example jointly modelling the drug treatment gene expression and toxicological profile data from multiple cell lines. This could also include measurements from multiple organisms and different toxicological outcome observations, as provided in the TG-GATEs database. Here the probabilistic multi-view methods, such as group factor analysis (GFA), would be readily applicable. The resulting components could capture both specific and shared toxicity-associated drug response patterns across multiple cell types. Such approach has been already demonstrated with the Connectivity Map data [65], and a large-scale analysis for example on the LINCS data could potentially provide a wealth of novel understanding on specific and general pharmacological and toxicological drug response mechanisms.

A lot of research has been conducted in predicting tissue-specific drug sensitivity combining multiple genomic data sources over hundreds of cell lines [7, 40]. Integrating the available drug treatment gene expression data to such analyses could provide improved predictions. Probabilistic multi-source models should be useful in integrating such disparate data sources, especially as even the state-of-the-art measurement technologies provide only partially comparable data [139]. Further incorporating chemical structure data into the analysis could provide direct links with

the particular substructures responsible for the molecular responses.

Identifying the drug response components could also benefit from combining gene expression with other high-throughput data such as protein interactions. The multi-view models such as GFA could be extended to use molecular network data for inferring structural relationships between the genes, and here the interaction component models used in Publication I would be a good starting point. Such a model could be used for interpreting the toxicogenomic molecular mechanisms as in Publication II, and also for improved connectivity mapping, as in Publications III and IV. Moreover, the model-based information visualisation method introduced in Publication V could be used to provide a comprehensive overview of the structure the drug response space.

These studies and ideas also give motivation further model development. As the high-throughput databases contain a lot of missing observations over the measured data and cell types and organisms, it would be important for the probabilistic models to handle the missing data properly. Additionally, the data often has some known underlying structure that should be optimally included in the model. For example, the GFA model assumes that all dimensions in the data views are independent, whereas the genes across cell types are known to mostly behave in a similar manner. Thus a more restricted model could better match the biological phenomenon while also helping in model identifiability. Also the sparsity assumptions and technical implementation could be improved further based on prior knowledge of the biological systems.

This thesis provides scientific advances in both molecular biology and medicine applications and probabilistic modelling, and highlights the value of interdisciplinary research where applied research and methods development feed each other.

Bibliography

[1] E. M. Airoldi. Getting Started in Probabilistic Graphical Models. *PLoS Compututational Biology*, 3(12), 2007.

[2] E. M. Airoldi, S. E. Fienberg, and E. P. Xing. Mixed membership analysis of genome-wide expression data. *ArXiv e-print arXiv:0711.2520*, 2007.

[3] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, New York, fourth edition, 2002.

[4] A. Alexa, J. Rahnenführer, and T. Lengauer. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22(13):1600–1607, 2006.

[5] C. Archambeau and F. Bach. Sparse probabilistic projections. In D. Koller, D. Schuurmans, Y. Bengio, L. Bottou, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 73–80. MIT Press, Cambridge, MA, 2009.

[6] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–29, 2000.

[7] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehar, G. V. Kryukov, D. Sonkin, A. Reddy, M. Liu, L. Murray, M. F. Berger, J. E. Monahan, P. Morais, J. Meltzer, A. Korejwa, J. Jane-Valbuena, F. A. Mapa, J. Thibault, E. Bric-Furlong, P. Raman, A. Shipway, I. H. Engels, J. Cheng, G. K. Yu, J. Yu, P. Aspesi, M. de Silva, K. Jagtap, M. D. Jones, L. Wang, C. Hatton, E. Palescandolo, S. Gupta, S. Mahan, C. Sougnez, R. C. Onofrio, T. Liefeld, L. MacConaill, W. Winckler, M. Reich, N. Li, J. P. Mesirov, S. B. Gabriel, G. Getz, K. Ardlie, V. Chan, V. E. Myer, B. L. Weber, J. Porter, M. Warmuth, P. Finan, J. L. Harris, M. Meyerson, T. R. Golub, M. P. Morrissey, W. R. Sellers, R. Schlegel, and L. A. Garraway. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–307, 2012.

[8] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muertter, M. Holko, O. Ayanbule, A. Yefanov, and A. Soboleva. NCBI GEO:

archive for functional genomics data sets—10 years on. *Nucleic Acids Research*, 39(suppl 1):D1005–D1010, 2011.

[9] C. M. Bishop. Variational principal components. In *Proceedings Ninth International Conference on Artificial Neural Networks*, volume 1, pages 509–514. IEE, 1999.

[10] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer Science+Bussiness Media LLC, New York, first edition, 2006.

[11] D. M. Blei and J. D. Lafferty. A correlated topic model of Science. *The Annals of Applied Statistics*, 1(1), 2007.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(4-5):993–1022, 2003.

[13] W. Buntine and A. Jakulin. Applying discrete PCA in data analysis. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 59–66. AUAI Press, Arlington, VA, 2004.

[14] A. J. Butte and I. S. Kohane. Creation and implications of a phenome-genome network. *Nature Biotechnology*, 24(1):55–62, 2006.

[15] F. Caiment, M. Tsamou, D. Jennen, and J. Kleinjans. Assessing compound carcinogenicity in vitro using connectivity mapping. *Carcinogenesis*, 35(1):201–207, 2014.

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

[17] J. Caldas, N. Gehlenborg, E. Kettunen, A. Faisal, M. Rönty, A. G. Nicholson, S. Knuutila, A. Brazma, and S. Kaski. Data-driven information retrieval in heterogeneous collections of transcriptomics data links SIM2s to malignant pleural mesothelioma. *Bioinformatics*, 28(2):246–253, 2012.

[18] J. Caldas and S. Kaski. Bayesian biclustering with the plaid model. In *IEEE Workshop on Machine Learning for Signal Processing*, pages 291–296. IEEE, 2008.

[19] J. Caldas and S. Kaski. Hierarchical generative biclustering for microRNA expression analysis. *Journal of Computational Biology*, 18(3):251–261, 2011.

[20] C. M. Carvalho, J. Chang, J. E. Lucas, J. R. Nevins, Q. Wang, and M. West. High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics. *Journal of the American Statistical Association*, 103(484):1438–1456, 2008.

[21] B.-J. J. Chen, H. C. Causton, D. Mancenido, N. L. Goddard, E. O. Perlstein, and D. Pe'er. Harnessing gene expression to identify the genetic basis of drug resistance. *Molecular Systems Biology*, 5, 2009.

[22] M. Chen, V. Vijay, Q. Shi, Z. Liu, H. Fang, and W. Tong. FDA-approved drug labeling for the study of drug-induced liver injury. *Drug Discovery Today*, 16(15-16):697–703, 2011.

[23] M. Chen, M. Zhang, J. Borlak, and W. Tong. A decade of toxicogenomic research and its contribution to toxicological science. *Toxicological Sciences*, 130(2):217–228, 2012.

[24] F. S. Collins, G. M. Gray, and J. R. Bucher. Toxicology. Transforming environmental health protection. *Science*, 319(5865):906–907, 2008.

[25] J. Corander, T. Aittokallio, S. Ripatti, and S. Kaski. The rocky road to personalized medicine: computational and statistical challenges. *Personalized Medicine*, 9(2):109–114, 2012.

[26] F. Crick. Central Dogma of Molecular Biology. *Nature*, 227(5258):561–563, 1970.

[27] W. Cui, H. Zhou, H. Qu, P. C. Wong, and X. Li. Geometry-Based Edge Clustering for Graph Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1277–1284, 2008.

[28] X. Cui and G. Churchill. Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology*, 4(4):210+, 2003.

[29] R. De Bin and D. Risso. A novel approach to the clustering of microarray data via nonparametric density estimation. *BMC Bioinformatics*, 12(1):49+, 2011.

[30] P. D'haeseleer. How does gene expression clustering work? *Nature Biotechnology*, 23(12):1499–1501, 2005.

[31] X. Ding, L. He, and L. Carin. Bayesian Robust Principal Component Analysis. *IEEE Transactions on Image Processing*, 20(12):3419–3430, 2011.

[32] J. T. Dudley, R. Tibshirani, T. Deshpande, and A. J. Butte. Disease signatures are robust across tissues and experiments. *Molecular systems biology*, 5(1), 2009.

[33] S. Dudoit, J. Schaffer, and J. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103, 2003.

[34] R. Edgar, M. Domrachev, and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002.

[35] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 1998.

[36] F. Farooq, B. Krishnapuram, R. Rosales, S. Yu, J. Shavlik, and R. Kucherlapati. Predictive Models in Personalized Medicine: Neural Information Processing Systems, 2010 Workshop Report. *SIGHIT Record*, 1(1):23–25, 2011.

[37] T. C. Freeman, L. Goldovsky, M. Brosch, S. van Dongen, P. Mazière, R. J. Grocock, S. Freilich, J. Thornton, and A. J. Enright. Construction, Visualisation, and Clustering of Transcription Networks from Microarray Expression Data. *PLoS Computational Biology*, 3(10):e206–2042, 2007.

[38] T. M. J. Fruchterman and E. M. Reingold. Graph drawing by force-directed placement. *Software - Practice & Experience*, 21(11):1129–1164, 1991.

[39] B. Ganter, S. Tugendreich, C. I. Pearson, E. Ayanoglu, S. Baumhueter, K. A. Bostian, L. Brady, L. J. Browne, J. T. Calvin, G.-J. J. Day, N. Breckenridge, S. Dunlea, B. P. Eynon, M. M. Furness, J. Ferng, M. R. Fielden, S. Y. Fujimoto, L. Gong, C. Hu, R. Idury, M. S. Judo, K. L. Kolaja, M. D. Lee, C. McSorley, J. M. Minor, R. V. Nair, G. Natsoulis, P. Nguyen, S. M. Nicholson, H. Pham, A. H. Roter, D. Sun, S. Tan, S. Thode, A. M. Tolley, A. Vladimirova, J. Yang, Z. Zhou, and K. Jarnagin. Development of a large-scale chemogenomics database to improve drug candidate selection and to understand mechanisms of chemical toxicity and action. *Journal of Biotechnology*, 119(3):219–244, 2005.

[40] M. J. Garnett, E. J. Edelman, S. J. Heidorn, C. D. Greenman, A. Dastur, K. W. Lau, P. Greninger, I. R. Thompson, X. Luo, J. Soares, Q. Liu, F. Iorio, D. Surdez, L. Chen, R. J. Milano, G. R. Bignell, A. T. Tam, H. Davies, J. A. Stevenson, S. Barthorpe, S. R. Lutz, F. Kogera, K. Lawrence, A. McLaren-Douglas, X. Mitropoulos, T. Mironenko, H. Thi, L. Richardson, W. Zhou, F. Jewitt, T. Zhang, P. O/'Brien, J. L. Boisvert, S. Price, W. Hur, W. Yang, X. Deng, A. Butler, H. G. Choi, J. W. Chang, J. Baselga, I. Stamenkovic, J. A. Engelman, S. V. Sharma, O. Delattre, J. Saez-Rodriguez, N. S. Gray, J. Settleman, P. A. Futreal, D. A. Haber, M. R. Stratton, S. Ramaswamy, U. McDermott, and C. H. Benes. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570–575, 2012.

[41] N. Gehlenborg, S. I. O'Donoghue, N. S. Baliga, A. Goesmann, M. A. Hibbs, H. Kitano, O. Kohlbacher, H. Neuweger, R. Schneider, D. Tenenbaum, and A.-C. C. Gavin. Visualization of omics data for systems biology. *Nature Methods*, 7(3 Suppl):S56–S68, 2010.

[42] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, third edition, 2013.

[43] G. K. Gerber, R. D. Dowell, T. S. Jaakkola, and D. K. Gifford. Automated Discovery of Functional Generality of Human Gene Expression Programs. *PLoS Computational Biology*, 3(8):e148+, 2007.

[44] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.

[45] H. Goodarzi, O. Elemento, and S. Tavazoie. Revealing Global Regulatory Perturbations across Human Cancers. *Molecular Cell*, 36(5):900–911, 2009.

[46] U. Güldener, M. Münsterkötter, G. Kastenmüller, N. Strack, J. van Helden, C. Lemer, J. Richelles, S. J. Wodak, J. García-Martínez, J. E. Pérez-Ortín, H. Michael, A. Kaps, E. Talla, B. Dujon, B. André, J. L. Souciet, J. De Montigny, E. Bon, C. Gaillardin, and H. W. Mewes. CYGD: the Comprehensive Yeast Genome Database. *Nucleic Acids Research*, 33(Database issue):D364+, 2005.

[47] S. Hachul and M. Jünger. Large-Graph Layout Algorithms at Work: An Experimental Study. *Journal of Graph Algorithms and Applications*, 11(2), 2007.

[48] R. Hadany and D. Harel. A Multi-Scale Algorithm for Drawing Graphs Nicely. In P. Widmayer, G. Neyer, and S. Eidenbenz, editors, *Graph-Theoretic Concepts in Computer Science*, volume 1665 of *Lecture Notes in Computer Science*, pages 262–277. Springer Berlin Heidelberg, 1999.

[49] A. Hahn, J. Rahnenfuhrer, P. Talwar, and T. Lengauer. Confirmation of human protein interaction data by human expression data. *BMC Bioinformatics*, 6(1):112+, 2005.

[50] K. M. Hall. An r-Dimensional Quadratic Placement Algorithm. *Management Science*, 17(3), 1970.

[51] D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer. Co-clustering of biological networks and gene expression data. *Bioinformatics*, 18(suppl 1):S145–S154, 2002.

[52] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical Correlation Analysis: An Overview with Application to Learning Methods. *Neural Computation*, 16(12):2639–2664, 2004.

[53] B. Hardy, G. Apic, P. Carthew, D. Clark, D. Cook, I. Dix, S. Escher, J. Hastings, D. J. Heard, N. Jeliazkova, P. Judson, S. Matis-Mitchell, D. Mitic, G. Myatt, I. Shah, O. Spjuth, O. Tcheremenskaia, L. Toldo, D. Watson, A. White, and C. Yang. Toxicology ontology perspectives. *ALTEX*, 29(2):139–156, 2012.

[54] T. Hartung, E. van Vliet, J. Jaworska, L. Bonilla, N. Skinner, and R. Thomas. Food for thought ... systems toxicology. *ALTEX*, 29(2):119–128, 2012.

[55] I. Herman, G. Melancon, and M. S. Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):24–43, 2000.

[56] A. Hopkins, J. Mason, and J. Overington. Can we rationally design promiscuous drugs? *Current Opinion in Structural Biology*, 16(1):127–136, 2006.

[57] H. Hotelling. Relations Between Two Sets of Variates. *Biometrika*, 28(3/4):321–377, 1936.

[58] G. Hu and P. Agarwal. Human Disease-Drug Network Based on Genomic Expression Profiles. *PLoS ONE*, 4(8):e6536+, 2009.

[59] D. Huang, B. Sherman, Q. Tan, J. Collins, W. G. Alvord, J. Roayaei, R. Stephens, M. Baseler, H. C. Lane, and R. Lempicki. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biology*, 8(9):R183+, 2007.

[60] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102(1):109–126, 2000.

[61] I. Huopaniemi, T. Suvitaival, J. Nikkilä, M. Orešič, and S. Kaski. Multivariate multi-way analysis of multi-source data. *Bioinformatics*, 26(12):i391–i398, 2010.

[62] F. Iorio, R. Bosotti, E. Scacheri, V. Belcastro, P. Mithbaokar, R. Ferriero, L. Murino, R. Tagliaferri, N. Brunetti-Pierri, A. Isacchi, and D. 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.

[63] F. Iorio, T. Rittman, H. Ge, M. Menden, and J. Saez-Rodriguez. Transcriptional data: a new gateway to drug repositioning? *Drug Discovery Today*, 18(7-8):350–357, 2013.

[64] M. Iskar, M. Campillos, M. Kuhn, L. J. Jensen, V. van Noort, and P. Bork. Drug-Induced Regulation of Target Expression. *PLoS Computational Biology*, 6(9):e1000925+, 2010.

[65] M. Iskar, G. Zeller, P. Blattmann, M. Campillos, M. Kuhn, K. H. Kaminska, H. Runz, A.-C. C. Gavin, R. Pepperkok, V. van Noort, and P. Bork. Characterization of drug-induced transcriptional modules: towards drug repositioning and functional understanding. *Molecular Systems Biology*, 9(1), 2013.

[66] D. Jiang, C. Tang, and A. Zhang. Cluster analysis for gene expression data: a survey. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1370–1386, 2004.

[67] M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran, R. Whaley, R. A. Glennon, J. Hert, K. L. Thomas, D. D. Edwards, B. K. Shoichet, and B. L. Roth. Predicting new molecular targets for known drugs. *Nature*, 462(7270):175–181, 2009.

[68] S. Khan, A. Faisal, J. Mpindi, J. Parkkinen, T. Kalliokoski, A. Poso, O. Kallioniemi, K. Wennerberg, and S. Kaski. Comprehensive data-driven analysis of the impact of chemoinformatic structure on the genome-wide biological response profiles of cancer cells to 1159 drugs. *BMC Bioinformatics*, 13(1):112+, 2012.

[69] A. Klami and S. Kaski. Local Dependent Components. In Z. Ghahramani, editor, *Proceedings of the 24th International Conference on Machine Learning*, pages 425–432. ACM, New York, NY, 2007.

[70] A. Klami, S. Virtanen, and S. Kaski. Bayesian Canonical Correlation Analysis. *Journal of Machine Learning Research*, 14:965–1003, 2013.

[71] D. Knowles and Z. Ghahramani. Nonparametric Bayesian sparse factor models with application to gene expression modeling. *The Annals of Applied Statistics*, 5(2B):1534–1552, 2011.

[72] P. Kohonen, E. Benfenati, D. Bower, R. Ceder, M. Crump, K. Cross, R. C. Grafström, L. Healy, C. Helma, N. Jeliazkova, V. Jeliazkov, S. Maggioni, S. Miller, G. Myatt, M. Rautenberg, G. Stacey, E. Willighagen, J. Wiseman, and B. Hardy. The ToxBank Data Warehouse: Supporting the Replacement of In Vivo Repeated Dose Systemic Toxicity Testing. *Molecular Informatics*, 32(1):47–63, 2013.

[73] J. B. Kruskal and J. B. Seery. Designing network diagrams. In *Proceedings of the First General Conference on Social Graphics*, pages 22–50. U.S. Department of Commerce, Bureau of the Census, 1980.

[74] Z. Kutalik, J. S. Beckmann, and S. Bergmann. A modular approach for integrative analysis of large-scale gene-expression and drug-response data. *Nature Biotechnology*, 26(5):531–539, 2008.

[75] K. Lage, N. T. Hansen, E. O. Karlberg, A. C. Eklund, F. S. Roque, P. K. Donahoe, Z. Szallasi, T. S. Jensen, and S. Brunak. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proceedings of the National Academy of Sciences*, 105(52):20870–20875, 2008.

[76] L. Lahti, J. E. A. Knuuttila, and S. Kaski. Global modeling of transcriptional responses in interaction networks. *Bioinformatics*, 26(21):2713–2720, 2010.

[77] L. Lahti, S. Myllykangas, S. Knuutila, and S. Kaski. Dependency detection with similarity constraints. In *IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE, 2009.

[78] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong, S. J. Haggarty, P. A. Clemons, R. Wei, S. A. Carr, E. S. Lander, and T. R. Golub. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science*, 313(5795):1929–1935, 2006.

[79] A. C. Lee, K. Shedden, G. R. Rosania, and G. M. Crippen. Data Mining the NCI60 to Predict Generalized Cytotoxicity. *Journal of Chemical Information and Modeling*, 48(7):1379–1388, 2008.

[80] E. Lee, H.-Y. Chuang, J.-W. Kim, T. Ideker, and D. Lee. Inferring Pathway Activity toward Precise Disease Classification. *PLoS Computational Biology*, 4(11):e1000217+, 2008.

[81] D. Lin, J. Zhang, J. Li, V. Calhoun, H. W. Deng, and Y. P. Wang. Group sparse canonical correlation analysis for genomic data integration. *BMC Bioinformatics*, 14(1):245+, 2013.

[82] Y. Low, T. Uehara, Y. Minowa, H. Yamada, Y. Ohno, T. Urushidani, A. Sedykh, E. Muratov, V. Kuz'min, D. Fourches, H. Zhu, I. Rusyn, and A. Tropsha. Predicting Drug-Induced Hepatotoxicity Using QSAR and Toxicogenomics Approaches. *Chemical Research in Toxicology*, 24(8):1251–1262, 2011.

[83] I. Maraziotis, K. Dimitrakopoulou, and A. Bezerianos. An in silico method for detecting overlapping functional modules from composite biological networks. *BMC Systems Biology*, 2(1):93+, 2008.

[84] E. R. Mardis. Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, 9(1):387–402, 2008.

[85] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, and U. Alon. Superfamilies of Evolved and Designed Networks. *Science*, 303(5663):1538–1542, 2004.

[86] C. L. Myers, M. J. Dunham, S. Y. Kung, and O. G. Troyanskaya. Accurate detection of aneuploidies in array CGH and gene expression microarray data. *Bioinformatics*, 20(18):3533–3543, 2004.

[87] N. Nariai, E. D. Kolaczyk, and S. Kasif. Probabilistic protein function prediction from heterogeneous genome-wide data. *PLoS ONE*, 2(3):e337+, 2007.

[88] A. Noack. Energy Models for Graph Clustering. *Journal of Graph Algorithms and Applications*, 11(2):453–480, 2007.

[89] P. Nurse and J. Hayles. The cell in an era of systems biology. *Cell*, 144(6):850–854, 2011.

[90] H. Parkinson, M. Kapushesky, N. Kolesnikov, G. Rustici, M. Shojatalab, N. Abeygunawardena, H. Berube, M. Dylag, I. Emam, A. Farne, E. Holloway, M. Lukk, J. Malone, R. Mani, E. Pilicheva, T. F. Rayner, F. Rezwan, A. Sharma, E. Williams, X. Z. Bradley, T. Adamusiak, M. Brandizi, T. Burdett, R. Coulson, M. Krestyaninova, P. Kurnosov, E. Maguire, S. G. Neogi, P. Rocca-Serra, S.-A. Sansone, N. Sklyar, M. Zhao, U. Sarkans, and A. Brazma. ArrayExpress update—from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Research*, 37(suppl 1):D868–D872, 2009.

[91] X. A. Qu and D. K. Rajpal. Applications of Connectivity Map in drug discovery and development. *Drug Discovery Today*, 17(23-24):1289–1298, 2012.

[92] J. Quackenbush. Microarray data normalization and transformation. *Nature Genetics*, 32 Suppl:496–501, 2002.

[93] P. Ranganathan. From Microprocessors to Nanostores: Rethinking Data-Centric Systems. *Computer*, 44(1):39–48, 2011.

[94] J. M. Raser and E. K. O'Shea. Noise in Gene Expression: Origins, Consequences, and Control. *Science*, 309(5743):2010–2013, 2005.

[95] A. W. Rives and T. Galitski. Modular organization of cellular networks. *Proceedings of the National Academy of Sciences*, 100(3):1128–1133, 2003.

[96] J. Rung and A. Brazma. Reuse of public genome-wide gene expression data. *Nature Reviews Genetics*, 14(2):89–99, 2012.

[97] C. Sander. Genomic Medicine and the Future of Health Care. *Science*, 287(5460):1977–1978, 2000.

[98] R. S. Savage, Z. Ghahramani, J. E. Griffin, B. J. de la Cruz, and D. L. Wild. Discovering transcriptional modules by Bayesian data integration. *Bioinformatics*, 26(12):i158–i167, 2010.

[99] A. Schulze and J. Downward. Navigating gene expression using microarrays [mdash] a technology review. *Nature Cell Biology*, 3(8):E190–E195, 2001.

[100] E. Segal, N. Friedman, D. Koller, and A. Regev. A module map showing conditional activity of expression modules in cancer. *Nature Genetics*, 36(10):1090–1098, 2004.

[101] E. Segal, T. Raveh-Sadka, M. Schroeder, U. Unnerstall, and U. Gaul. Predicting expression patterns from regulatory sequence in Drosophila segmentation. *Nature*, 451(7178):535–540, 2008.

[102] E. Segal, B. Taskar, A. Gasch, N. Friedman, and D. Koller. Rich probabilistic models for gene expression. *Bioinformatics*, 17 Suppl 1(suppl 1):S243–S252, 2001.

[103] E. Segal, H. Wang, and D. Koller. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, 19(suppl 1):i264–i272, 2003.

[104] E. Segal, R. Yelensky, and D. Koller. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, 19(suppl 1):i273–i282, 2003.

[105] R. Sharan and R. Shamir. CLICK: a clustering algorithm with applications to gene expression analysis. In *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 307–316. AAAI Press, 2000.

[106] R. Sharan, I. Ulitsky, and R. Shamir. Network-based prediction of protein function. *Molecular Systems Biology*, 3(1), 2007.

[107] R. Shen, A. B. Olshen, and M. Ladanyi. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906–2912, 2009.

[108] M. Shiga, I. Takigawa, and H. Mamitsuka. Annotating gene function by combining expression data with a modular gene network. *Bioinformatics*, 23(13):i468–478, 2007.

[109] D. Shigemizu, Z. Hu, J.-H. Hung, C.-L. Huang, Y. Wang, and C. DeLisi. Using Functional Signatures to Identify Repositioned Drugs for Breast, Myelogenous Leukemia and Prostate Cancer. *PLoS Computational Biology*, 8(2):e1002347+, 2012.

[110] Y. Shiraishi, S. Kimura, and M. Okada. Inferring cluster-based networks from differently stimulated multiple time-course gene expression data. *Bioinformatics*, 26(8):1073–1081, 2010.

[111] R. H. Shoemaker. The NCI60 human tumour cell line anticancer drug screen. *Nature Reviews Cancer*, 6(10):813–823, 2006.

[112] J. Sinkkonen, J. Aukia, and S. Kaski. Inferring vertex properties from topology in large networks. In *Working Notes of the 5th International Workshop on Mining and Learning with Graphs*. Universita degli Studi di Firenze, 2007.

[113] J. Sinkkonen, J. Aukia, and S. Kaski. Component models for large networks. *ArXiv eprint arXiv:0803.1628*, 2008.

[114] M. Sirota, J. T. Dudley, J. Kim, A. P. Chiang, A. A. Morgan, A. Sweet-Cordero, J. Sage, and A. J. Butte. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Science Translational Medicine*, 3(96):96ra77, 2011.

[115] J. L. Smalley, T. W. Gant, and S.-D. Zhang. Application of connectivity mapping in predictive toxicology based on gene-expression similarity. *Toxicology*, 268(3):143–146, 2010.

[116] S. C. Smith, A. S. Baras, J. K. Lee, and D. Theodorescu. The COXEN principle: translating signatures of in vitro chemosensitivity into tools for clinical outcome prediction and drug discovery in cancer. *Cancer Research*, 70(5):1753–1758, 2010.

[117] B. P. Sorrentino. Gene therapy to protect haematopoietic cells from cytotoxic cancer drugs. *Nature Reviews Cancer*, 2(6):431–441, 2002.

[118] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. 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.

[119] S. Suthram, J. T. Dudley, A. P. Chiang, R. Chen, T. J. Hastie, and A. J. Butte. Network-Based Elucidation of Human Disease Similarities Reveals Common Functional Modules Enriched for Pluripotent Drug Targets. *PLoS Computational Biology*, 6(2):e1000662+, 2010.

[120] A. Tanay, R. Sharan, and R. Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18(suppl 1):S136–S144, 2002.

[121] Y. W. Teh. Dirichlet processes. In C. Sammut and G. I. Webb, editors, *Encyclopedia of Machine Learning*. Springer, New York, 2011.

[122] R. S. Thomas, M. B. Black, L. Li, E. Healy, T.-M. M. Chu, W. Bao, M. E. Andersen, and R. D. Wolfinger. A comprehensive statistical analysis of predicting in vivo hazard using high-throughput in vitro screening. *Toxicological Sciences*, 128(2):398–417, 2012.

[123] L. Thurstone. Multiple factor analysis. *Psychological Review*, 38(5):406–427, 1931.

[124] T. T. Torres, M. Metta, B. Ottenwälder, and C. Schlötterer. Gene expression profiling by massively parallel sequencing. *Genome Research*, 18(1):172–177, 2008.

[125] R. J. Trent. *Molecular Medicine: Genomics to Personalized Healthcare*. Elsevier, fourth edition, 2012.

[126] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, Reading, 1977.

[127] T. Uehara, A. Ono, T. Maruyama, I. Kato, H. Yamada, Y. Ohno, and T. Urushidani. The Japanese toxicogenomics project: Application of toxicogenomics. *Molecular Nutrition & Food Research*, 54(2):218–227, 2010.

[128] I. Ulitsky and R. Shamir. Identification of functional modules using network topology and high-throughput data. *BMC Systems Biology*, 1(1):8+, 2007.

[129] I. Ulitsky and R. Shamir. Identifying functional modules using expression profiles and confidence-scored protein interactions. *Bioinformatics*, 25(9):1158–1164, 2009.

[130] R. J. P. van Berlo, L. F. A. Wessels, S. D. C. Martes, and M. J. T. Reinders. Predicting gene function by combining expression and interaction data. In *Computational Systems Bioinformatics Conference, Workshops and Poster Abstracts*, pages 166–167. IEEE, 2005.

[131] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski. Information Retrieval Perspective to Nonlinear Dimensionality Reduction for Data Visualization. *Journal of Machine Learning Research*, 11:451–490, 2010.

[132] S. Virtanen, A. Klami, and S. Kaski. Bayesian CCA via group sparsity. In L. Getoor and T. Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning*, pages 457–464. ACM, New York, NY, 2011.

[133] S. Virtanen, A. Klami, S. A. Khan, and S. Kaski. Bayesian Group Factor Analysis. In N. D. Lawrence and M. A. Girolami, editors, *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, volume 22 of *JMLR W&CP*, pages 1269–1277, 2012.

[134] C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields, and P. Bork. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887):399–403, 2002.

[135] C. Walshaw. A Multilevel Algorithm for Force-Directed Graph Drawing. In J. Marks, editor, *Graph Drawing*, volume 1984 of *Lecture Notes in Computer Science*, pages 171–182. Springer Berlin Heidelberg, 2001.

[136] C. Wang. Variational Bayesian Approach to Canonical Correlation Analysis. *IEEE Transactions on Neural Networks*, 18(3):905–910, 2007.

[137] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[138] J. N. Weinstein. Spotlight on molecular profiling: "Integromic" analysis of the NCI-60 cancer cell lines. *Molecular Cancer Therapeutics*, 5(11):2601–2605, 2006.

[139] J. N. Weinstein and P. L. Lorenzi. Cancer: Discrepancies in drug sensitivity. *Nature*, 504(7480):381–383, 2013.

[140] M. West. Bayesian Factor Regression Models in the "Large p, Small n" Paradigm. In *Bayesian Statistics*, pages 723–732. Oxford University Press, 2003.

[141] P. M. Williams. Bayesian Regularization and Pruning Using a Laplace Prior. *Neural Computation*, 7(1):117–143, 1995.

[142] E. L. Willighagen, R. Wehrens, and L. M. C. Buydens. Molecular Chemometrics. *Critical Reviews in Analytical Chemistry*, 36(3-4):189–198, 2006.

[143] D. M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

[144] L. Xing, L. Wu, Y. Liu, N. Ai, X. Lu, and X. Fan. LTMap: a web server for assessing the potential liver toxicity by genome-wide transcriptional expression data. *Journal of Applied Toxicology*, 2013.

[145] Y. Yamanishi, J. P. Vert, A. Nakaya, and M. Kanehisa. Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. *Bioinformatics*, 19(suppl 1):i323–i330, 2003.

DISSERTATIONS IN INFORMATION AND COMPUTER SCIENCE

Aalto-DD44/2013 Pylkkönen, Janne
Towards Efficient and Robust Automatic Speech Recognition: Decoding Techniques and Discriminative Training. 2013.

Aalto-DD47/2013 Reyhani, Nima
Studies on Kernel Learning and Independent Component Analysis. 2013.

Aalto-DD70/2013 Ylipaavalniemi, Jarkko
Data-driven Analysis for Natural Studies in Functional Brain Imaging. 2013.

Aalto-DD61/2013 Kandemir, Melih
Learning Mental States from Biosignals. 2013.

Aalto-DD90/2013 Yu, Qi
Machine Learning for Corporate Bankruptcy Prediction. 2013.

Aalto-DD128/2013 Ajanki, Antti
Inference of relevance for proactive information retrieval. 2013.

Aalto-DD205/2013 Lijffijt, Jefrey
Computational methods for comparison and exploration of event sequences. 2013.

Aalto-DD21/2014 Cho, Kyunghyun
Foundations and Advances in Deep Learning. 2014.

Aalto-DD49/2014 Lindh-Knuutila, Tiina
Computational Modeling and Simulation of Language and Meaning: Similarity-Based Approaches. 2014.

Aalto-DD80/2014 Toivola, Janne
Advances in Wireless Damage Detection for Structural Health Monitoring. 2014.

Molecular medicine studies how cancer and other complex diseases operate on the molecular level. Identifying the detailed mechanisms and interactions of how diseases progress and respond to drug treatments is essential for developing effective therapies. High-throughput molecular profiling technologies have provided vast amounts of measurement data of these phenomena. However, making sense of these masses of data is far from straightforward and requires advanced computational analysis methods. Probabilistic component models have been proven an effective tool in analysing and integrating high-dimensional and noisy molecular profiling data sources, such as gene expression. Such models can identify coherent components from the data, and interpreting these components provides insights about the underlying biological processes, such as disease progression and drug responses. In this thesis, probabilistic component models are applied and extended to identify and analyse molecular interaction and drug response patterns.

Aalto-DD 105/2014

ISBN 978-952-60-5773-6
ISBN 978-952-60-5774-3 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934
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