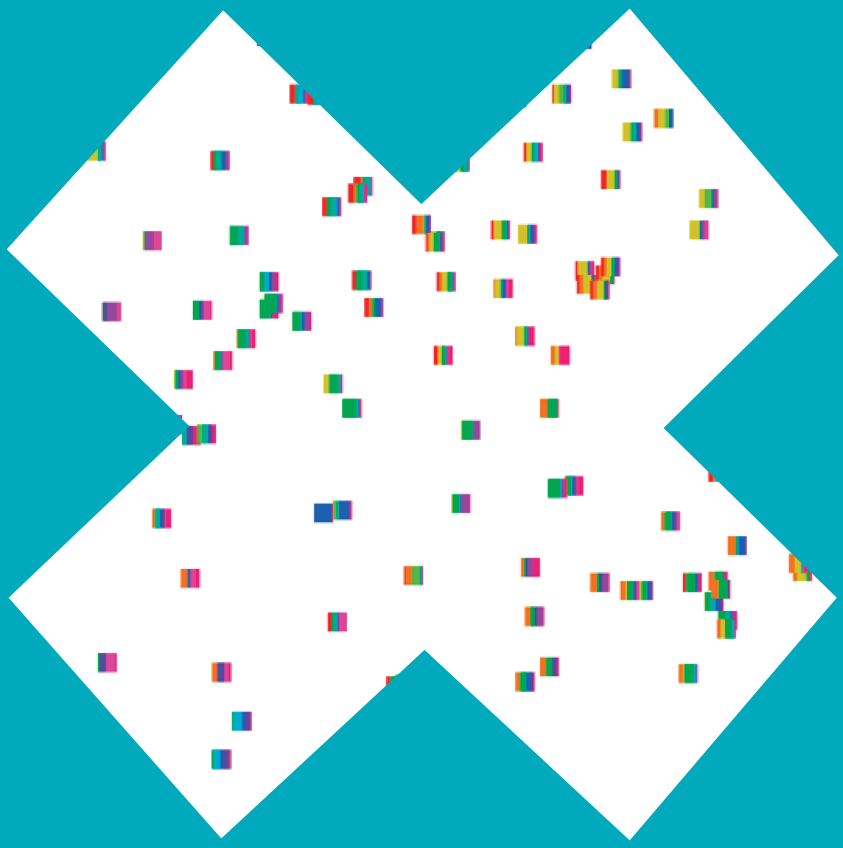
Retrieval of Gene Expression Measurements with Probabilistic Models

Ali Faisal



Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 108/2014

Retrieval of Gene Expression Measurements with Probabilistic Models

Ali Faisal

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 Auditorium AS1 of the school on 15th of August 2014 at 12.

Aalto University
School of Science
Department of Information and Computer Science

Supervising professor
Prof. Samuel Kaski

Thesis advisor
Dr. Jaakko Peltonen

Preliminary examiners
Dr. Reija Autio, Tampere University of Technology, Finland
Dr. Julio Saez-Rodriguez, European Bioinformatics Institute, Cambridge, United Kingdom

Opponent
Prof. Hiroshi Mamitsuka, Kyoto University, Gokasho, Japan

Aalto University publication series
DOCTORAL DISSERTATIONS 108/2014

ISBN 978-952-60-5780-4
ISBN 978-952-60-5781-1 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-5781-1

Unigrafia Oy
Helsinki 2014

Finland

441 697
Printed matter

Aalto University

Abstract

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

Author
Ali Faisal

Name of the doctoral dissertation
Retrieval of Gene Expression Measurements with Probabilistic Models

Publisher School of Science

Unit Department of Information and Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 108/2014

Field of research Information and Computer Science

Manuscript submitted 12 May 2014 **Date of the defence** 15 August 2014

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

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

Abstract

A crucial problem in current biological and medical research is how to utilize the diverse set of existing biological knowledge and heterogeneous measurement data in order to gain insights on new data. As datasets continue to be deposited in public repositories it is becoming important to develop search engines that can efficiently integrate existing data and search for relevant earlier studies given a new study. The search task is encountered in several biological applications including cancer genomics, pharmacokinetics, personalized medicine and meta-analysis of functional genomics.

Most existing search engines rely on classical keyword or annotation based retrieval which is limited to discovering known information and requires careful downstream annotation of the data. Data-driven model-based methods, that retrieve studies based on similarities in the actual measurement data, have a greater potential for uncovering novel biological insights. In particular, probabilistic modeling provides promising model-based tools due to its ability to encode prior knowledge, represent uncertainty in model parameters and handle noise associated to the data. By introducing latent variables it is further possible to capture relationships in data features in the form of meaningful biological components underlying the data.

This thesis adapts existing and develops new probabilistic models for retrieval of relevant measurement data in three different cases of background repositories. The first case is a background collection of data samples where each sample is represented by a single data type. The second case is a collection of multimodal data samples where each sample is represented by more than one data type. The third case is a background collection of datasets where each dataset, in turn, is a collection of multiple samples. In all three setups the proposed models are evaluated quantitatively and with case studies the models are demonstrated to facilitate interpretable retrieval of relevant data, rigorous integration of diverse information sources and learning of latent components from partly related dataset collections.

Keywords Machine learning, Bioinformatics, Probabilistic modeling, Information retrieval, Bayesian generative models

ISBN (printed) 978-952-60-5780-4 **ISBN (pdf)** 978-952-60-5781-1

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

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

Pages 255 **urn** http://urn.fi/URN:ISBN:978-952-60-5781-1

Preface

This thesis work has been carried out in the Statistical Machine Learning and Bioinformatics (MI) group in the Department of Information and Computer Science, Aalto University School of Science (formerly known as Helsinki University of Technology). I also have had the pleasure of being a part of Helsinki Institute of Information Technology HIIT and Finnish Center of Excellence in Computational Inference (COIN). My research work has been funded by Finnish Doctoral Programme in Computational Sciences (FICS), the Finnish Funding Agency for Technology and Innovation (TEKES, grant no. 40101/07) and the Pattern Analysis, Statistical Modeling and Computational Learning Network of Excellence (PASCAL2 EU Network of Excellence).

I would like to thank my supervisor Prof. Samuel Kaski for his invaluable guidance and encouragement throughout my research and for providing several excellent collaborative opportunities with field experts in systems biology. I am particularly grateful to Dr. Krister Wennerberg, Dr. Johan Rung, Prof. Olli Kallioniemi, Prof. Sampsa Hautaniemi and Prof. Sakari Knuutila for their guidance and enthusiasm in the collaborative research work.

I would especially like to thank my instructor Dr. Jaakko Peltonen for the many worthwhile mathematical discussions and his patience in helping me in very practical details. The resulting research environment has enabled me to acquire skills in both statistical machine learning and computational biology. I wish to express my gratitude to Prof. Dirk Husmeier who has consistently encouraged me and introduced me to the new interdisciplinary topics in computational biology and statistical machine learning.

I would like to thank the reviewers of this thesis, Doctor Reija Autio and Doctor Julio Saez-Rodriguez for their valuable feedback.

I am very grateful to all co-authors and all current and former members of the MI research group (especially Suleiman Khan, Leo Lahti, Melih Kandemir, Elisabeth Georgii, Sohan Seth, Jussi Gillberg, Tommi Suvitaival, Juuso Parkkinen, Jose Caldas and Gayle Leen) for valuable discussions on both science and life in general.

Finally, thanks to all my friends for their support and interest in the state of my research work. I also wish to thank all members of the Imaging Language group of Aalto University's Brain Research Unit, headed by Prof. Riitta Salmelin. Getting to know each of you while learning a new research area has been an exciting experience.

I am extremely grateful to my parents for their consistent support and encouragement throughout my doctoral studies. Most of all I wish to thank my wife Sobia and our dear little happy bird Abdul Moiz, for absolutely everything.

Espoo, June 20, 2014,

Ali Faisal

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 José Caldas, Nils Gehlenborg, Ali Faisal, Alvis Brazma and Samuel Kaski. Probabilistic retrieval and visualization of biologically relevant microarray experiments. *Bioinformatics*, 25(12):i145–i153, 2009.

II Ali Faisal, Frank Dondelinger, Dirk Husmeier, Colin M. Beale. Inferring species interaction networks from species abundance data: A comparative evaluation of various statistical and machine learning methods. *Ecological Informatics*, 5(6):451–464, 2010.

III José Caldas, Nils Gehlenborg, Eeva Kettunen, Ali Faisal, Mikko Rönty, Andrew G. Nicholson, Sakari Knuutila, Alvis Brazma and Samuel Kaski. Data-driven information retrieval in heterogeneous collections of transcriptomics data links *SIM2s* to malignant pleural mesothelioma. *Bioinformatics*, 28(2):246–253, 2012.

IV Suleiman A Khan, Ali Faisal, John P. Mpindi, Juuso A. Parkkinen, Tuomo Kalliokoski, Antti Poso, Olli P. Kallioniemi, Krister Wennerberg and Samuel 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*, 13:112, 2012.

V Riku Louhimo, Viljami Aittomaki*, Ali Faisal*, Marko Laakso*, Ping Chen, Kristian Ovaska, Erkka Valo, Leo Lahti, Vladimir Rogojin, Samuel

Kaski and Sampsa Hautaniemi. Systematic use of computational methods allows stratification of treatment responders in glioblastoma multiforme. *Systems Biomedicine*, 1(2):130–136, 2013.

VI Ali Faisal, Jussi Gillberg, Gayle Leen and Jaakko Peltonen. Transfer Learning using a Nonparametric Sparse Topic Model. *Neurocomputing*, 112:124–137, 2013.

VII Ali Faisal, Jaakko Peltonen, Elisabeth Georgii, Johan Rung and Samuel Kaski. Toward computational cumulative biology by combining models of biological datasets. *Submitted to a journal*, 6 pages, 2013.

Author's Contribution

Publication I: "Probabilistic retrieval and visualization of biologically relevant microarray experiments"

The author co-designed and implemented the underlying background model (topic model) and the retrieval subsystem, carried out the experiments on performance evaluation and co-wrote the manuscript.

Publication II: "Inferring species interaction networks from species abundance data: A comparative evaluation of various statistical and machine learning methods"

The author implemented the different algorithms for network reconstruction, made the data-analysis for the real dataset and co-wrote the manuscript.

Publication III: "Data-driven information retrieval in heterogeneous collections of transcriptomics data links *SIM2s* to malignant pleural mesothelioma"

The author implemented model selection and evaluation schemes, analyzed the results, facilitated the collaboration for the *SIM2s*-mesothelioma case study and co-wrote the manuscript.

Publication IV: "Comprehensive data-driven analysis of the impact of chemoinformatic structure on the genome-wide biological response profiles of cancer cells to 1159 drugs"

The author designed the retrieval study, collaborated with the first author in implementing the dependency analysis pipeline, generated different visualizations for the interpretation of the results, performed component level enrichment analysis and co-wrote the manuscript.

Publication V: "Systematic use of computational methods allows stratification of treatment responders in glioblastoma multiforme"

The author carried out the dependency and survival association analysis, co-designed the corresponding experiment and wrote related parts in the manuscript.

Publication VI: "Transfer Learning using a Nonparametric Sparse Topic Model"

The author carried out the experiments and had a key role in model derivation. The experimental design was a collaborative effort. The first and last author co-wrote the manuscript.

Publication VII: "Toward computational cumulative biology by combining models of biological datasets"

The model and experimental setup were designed together. The author implemented the model, carried out the experiments and performed data analysis. The writing of the article was a collaborative effort.

List of Abbreviations and Symbols

In this thesis boldface symbols are used to denote matrices and vectors. Uppercase symbols signify matrices and lowercase symbols column vectors. Normal lowercase symbols indicate scalar variables.

$\mathbb{R}$	Real domain
x, y	Scalars, data samples
$\mathbf{x}$, $\mathbf{y}$	Vectors, Multidimensional samples
X, Y	Scalar random variables
$\mathbf{X}$, $\mathbf{Y}$	Matrix, Multidimensional random variables
$\mathbf{z}$	Vector, Multidimensional latent variable
θ, ψ	Vector, parameters of a given model
$\mathcal{N}(\mu, \sigma^2)$	Gaussian distribution with mean μ and variance σ^2
$P(X)$	Probability mass function when X is discrete
$p(X)$	Probability density function when X is continuous
$p(X\|Y)$	Conditional probability mass function of X given Y
$\mathbf{I}$	Identity matrix
λ	Scalar, regularization parameter
$\log(.)$	Logarithmic function
$\mathbb{E}(X)$	Expected value of random variable X
$\mathbf{X}^T$	Transpose of matrix $\mathbf{X}$
$\mathbf{X}^{-1}$	Inverse of matrix $\mathbf{X}$
$\mathrm{Tr}(\mathbf{X})$	Trace of a matrix $\mathbf{X}$
$\mathcal{L}(\theta; \mathcal{D})$	Likelihood function; $P(\mathcal{D}\|\theta)$
$\arg\max_\theta f(\theta)$	Argument θ for which f has its maximum value
$\arg\min_\theta f(\theta)$	Argument θ for which f has its minimum value
aCGH	array-Comparative Genomic Hybridization
ARD	Automatic Relevance Determination
ATC	Anatomical Therapeutic Chemical codes
AUC	Area Under Curve

BN	Bayesian Network
CCA	Canonical Correlation Analysis
cDNA	Complementary DNA
CNV	Copy Number Variations
DNA	Deoxyribonucleic Acid
DP	Dirichlet Process
EFO	Experimental Factor Ontology
EM	Expectation Maximization
ES	Enrichment score
E-step	Expectation step
FP	False Positive
GBM	Glioblastoma Multiforme
GEO	Gene Expression Omnibus
GGM	Graphical Gaussian Model
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
HDP	Hierarchical Dirichlet Process
HM	Harmonic Mean
IBP	Indian Buffet Process
KL-divergence	Kullback-Leibler divergence
KM	Kaplan-Meier
LASSO	Least Absolute Shrinkage and Selection Operator
LDA	Latent Dirichlet Allocation
MAP	Maximum-a-Posteriori Estimation
MCMC	Markov Chain Monte Carlo
ML	Maximum Likelihood Estimation
MPM	Malignant Pleural Mesothelioma
M-step	Maximization step
MSigDB	Molecular Signature Database
MT-HDPLDA	Multi-Task HDP LDA topic model
NDCG	Normalized Discounted Cumulative Gain
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
RT-PCR	Real-Time Polymerase Chain Reaction
SBR	Sparse Bayesian Regression
TCGA	The Cancer Genome Atlas
TP	True Positive
TPFP5	TP rate at 5% FP rate

1. Introduction

With the rapid adaptation of high-throughput technology huge databases of biological measurements have become freely available. A crucial problem in current biological research is how to utilize the diverse sources of existing knowledge and heterogeneous measurement data to make accurate inferences about new data [1, 2]. Classical gene expression measurements hold great potential and continue to constitute a major part of the available databases. Figure 1.1 shows the number of microarray experiments in one of the biggest public databases (*ArrayExpress*; [3]) has doubled every two years since 2008.

As ArrayExpress and other repositories of genome-wide experiments are reaching a mature size, it is becoming more meaningful to utilize the diverse sources of existing knowledge and heterogeneous measurement data and search for related experiments given a new study. Existing solutions to retrieve relevant experiments either utilize the meta-data, such as annotations and descriptions of arrays and genes [4], or use the sample features directly without modeling shared biological patterns from the experiments [5]. The former considerably restricts the performance especially if the query sample has measurement data, while the latter fails to facilitate biological interpretation of the retrieval results.

Probabilistic modeling provides a flexible approach that uses the mathematics of probability theory to express noise in the data and uncertainty in the model parameters [6, 7]. The proposed models in this thesis complement task-dependent bioinformatics methods, which are naturally required in all biological and medical research problems as well, with methods that can efficiently integrate existing data and search for relevant studies given measurement data of user interest.

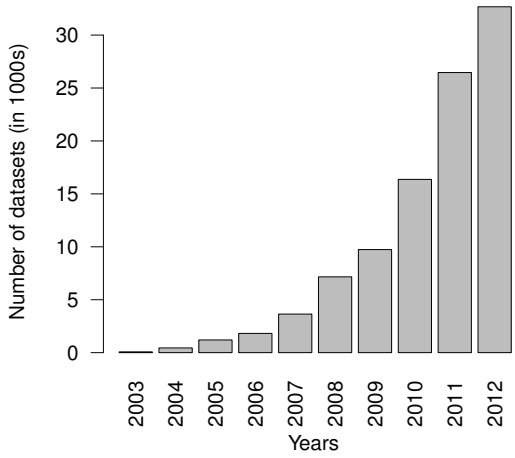


Figure 1.1. ArrayExpress database growth - number of microarray datasets deposited per year. *Statistics taken from ArrayExpress*

1.1 Contributions of the thesis

Given query data of user interest the thesis adapts existing and proposes new model-driven approaches to retrieve a ranked list of most relevant data from a background repository of earlier data measurements. The retrieval results are interpretable as they are based on the hidden high-level features (biological processes) modeled by the proposed methods. The contributions of the thesis can be grouped into three different scenarios for the background data repository: A) collection of single-view data samples, where each data sample is represented by a single data type, B) collection of multi-modal (or multi-view) samples, where each data sample is represented by multiple data types and C) collection of datasets, where each dataset is itself a collection of samples. The three scenarios and corresponding key contributions are summarized below:

The first set of contributions is a model-based measure of relevance between a query and a background sample (Publications I and III) and evaluation of different data-driven models (Publications II and III) for a background collection of single-view data samples. For the former, Publication I adapts a classical probabilistic latent variable model that yields the model-based relevance measure between gene expression profiles under a modeling assumption that the expression patterns in an individual sample are generated from multiple underlying processes. Publication III

compares the performance of the model-based retrieval on a larger data collection and also extends the underlying probabilistic model for a background data repository. While the models in Publications I and III are useful to quantify relevance at a global genomic scale, an alternative challenge is when there exist a few genes of interest and a user would like to 1. infer regulatory interactions among the genes (data features) and 2. focus the search on earlier samples that exhibit similar regulatory interactions for the genes of interest. Publication II addresses the former by comparing various models that can reconstruct the interaction network among a given set of data features. Specifically, the study uses both simulated and real world data to provide a rigorous evaluation of four commonly used reverse engineering methods. The second related task, that is, modeling relevance between samples given a set of potentially interesting features and a modeled regulatory network, has been investigated in an earlier study [8], briefly discussed in Section 4.3.3; the study utilizes one of the better performing reconstruction models, as evaluated in Publication II, and performs the search task using a predictive likelihood based relevance measure that is sensitive to the relationships (interaction strength and direction) among features.

The second contribution of the thesis is adaptation of a multi-view model to utilize a background collection of multi-modal samples to retrieve the most relevant earlier data samples (Publication IV) and search for chromosomal regions that are predictive of patient survival (Publication V). In this case a multi-modal (also referred to as multi-view) sample is represented by more than one data type. For example, a bi-modal representation of a tumor tissue can be its gene expression profile and copy number changes measured from the tumor tissue. Specifically Publication IV presents a data-driven approach to perform retrieval of relevant samples with the assumption that there exist multiple hidden linear combinations of features (represented as *latent space*) in the first data type that are maximally correlated with the hidden combinations of features in the second data type. Relevance between samples is computed using their projections to the latent spaces. The study presents different case studies on drug chemical and biological responses and finds the added benefit of another data type improving the retrieval performance in the multi-view setting as compared to a single-view setting. Another common case for multi-view repositories is when the multiple data types have feature spaces that can be mapped using a one-to-one correspondence of features

across data types. For example, a common case in cancer genomics is measurement of copy number amplifications and methylation changes; two chromosomally continuous data types where chromosomal regions can be mapped to individual genes in gene expression. Publication V adapts an existing multi-view model with the aim to extract relevant survival-associated multi-view features (potential biomarkers) that are predictive of patient survival. The problem is relevant for targeted repositories that contain multiple experiments, all measuring the same disease. Using a constrained version of the multi-view model used in Publication IV, Publication V first identifies chromosomal regions that are highly dependent across multiple data types and then performs a survival association analysis to further filter out those dependent regions that can effectively stratify patients into high and low survival groups.

Two novel and general-purpose approaches to model and relate a collection of datasets in Publication VI and Publication VII form the third set of contributions for the thesis. A typical dataset in transcriptomics corresponds to a microarray experiment that contains multiple microarray samples. The work in Publication VI considers the question under the assumption that the dataset-of-interest (query dataset, also called the *task of interest*) has a limited number of samples. Since each sample corresponds to a specific dataset, from a modeling point of view it builds a structured model for each dataset; in particular, a novel Bayesian generative transfer learning model is proposed that represents similarity across datasets by sparse sharing of latent components controlled by a non-parametric prior. The use of a non-parametric prior does not require one to pre-specify the number of latent components unlike in Publication I and Publication III. The method outperforms competing models on both simulated and real data with small numbers of samples. While traditional multi-task learning and the work in Publication VI take the approach of building a single unified model of all the data, as the number of datasets keeps increasing and the amount of quantitative biological knowledge keeps accumulating, the complexity of the task of building an accurate unified global model becomes increasingly prohibitive [9]. Publication VII introduces a novel general purpose and scalable method to relate a collection of datasets. Assuming that in the future researchers will increasingly develop their hypothesis in terms of (probabilistic) models of their own data which would allow them to take properly into account both the uncertainty in the data and the existing biological knowledge,

the study proposes a mixture model of existing models that decomposes a given query dataset into contributions from relevant background models. The parameters of the mixture model specify the amount of variation in the query dataset that is explained by a background model learned from a background dataset. These mixture weights are directly used as a proxy for relevance between two datasets. The data-driven decomposition identified a network of interrelated datasets from a large collection of human gene expression microarray experiments where tissue and disease were found to be the major factor in determining relevant datasets. Further, the findings from the case study were able to correctly identify inconsistencies in the public repositories.

From a machine learning perspective, the key contributions of the thesis are:

1. A retrieval model for transcriptomics where a topic model family is adapted to compute a probabilistic model-driven measure of relevance (Publication I).

2. A comparative evaluation of various network reconstruction approaches that can infer relationships among different features or genes (Publication II).

3. Adaptations of an existing canonical correlation model family where the latent components are used to a) perform multi-modal retrieval (Publication IV)and b) used to validate identified chromosomal regions based on their power to predict patient survival (Publication V).

4. A novel multi-task Bayesian topic model that is able to relate datasets and performs better than state-of-the-art non-parametric multi-task topic models. (Publication VI)

5. A novel mixture model of data models that is able to decompose a given query dataset into earlier datasets and is both scalable and rapidly computable. (Publication VII)

Table 1.1 summarizes the relationships between the publications and contribution areas.

Table 1.1. Publications and the main contribution areas.

	Publications						
	I	II	III	IV	V	VI	VII
Model-based retrieval for data samples	X	X	X				
Multi-view retrieval with biological samples as queries				X	X		
Multi-task learning and retrieval of datasets						X	X

1.2 Organization of the thesis

In the thesis an overview of the computational methods is provided and the main contributions are highlighted. Chapter 2 presents an overview of basic molecular biology, functional genomics and genomic data resources that are used in the subsequent studies. Chapter 3 builds a methodological background by starting from basic probability theory, followed by density estimation and finally motivates probabilistic latent variable models and Bayesian inference techniques. Chapters 4-6 describe the key contributions of the thesis on information retrieval, in three different scenarios for the background data collection, that is, single-view collection of multiple samples, a multi-modal or multi-view collection of multiple samples, and a collection of multiple datasets, respectively. Lastly, Chapter 7 presents conclusions and discusses potential future directions.

2. Molecular Biology

This chapter provides an introduction to the basic concepts of molecular biology primarily used in the thesis - cells, molecules, genes and functional genomics. The brief introduction is written with emphasis on genomics. It is intended for those who do not have a strong biological background. For futher background in molecular biology see [10–12].

2.1 Organization of genetic information

Cells are the basic building blocks of living organisms. All organisms consist of small cells that are typically too small to be seen by a naked eye. There are various different cell types in the human body including skin cells, muscle cells, red blood cells and brain cells (neurons) etc. Each cell carries a copy of the heritable genetic code, *the genome*. The human genome is the complete set of the 23 pairs of chromosomes where every chromosome contains coiled-up deoxyribonucleic acid (*DNA*) molecules [10]. DNA is the main information carrier molecule in a cell and in eukaryotic organisms (e.g., humans and plants) most DNA is stored in an enclosed cellular compartment known as the nucleus.

The genetic information in the DNA is encoded as a sequence containing four organic bases: adenine (A), thymine (T), guanine (G) and cytosine (C). These nucleotide bases are attached to a sugar-phosphate backbone containing two strands of complementary nucleotide sequences where in the opposing strands only A-T and G-C pairs can hybridize with each other. This leads to the well-known double-helix structure of the DNA and forms the basis for transmission of genetic information. The *central dogma of molecular biology* [13] explains the flow of information from DNA as an irreversible process of *protein synthesis* where the DNA acts as the template for its own replication and encodes the information

for the construction of a protein. This is a simplified framework for understanding the transfer of information at the cellular level in living organisms.

2.1.1 Protein synthesis

Scattered along the DNA molecule are particularly important sequences of bases known as *genes*. Genes are the basic functional units of genetic information that contain a particular set of instructions, usually coding for a particular *protein* or a particular function [10]. Variation and regulation in the gene activity therefore have major phenotypic consequences. Proteins are the fundamental entities in the cell that perform key functions within living organisms, such as response to stimuli, transport of molecules from one location to another, DNA replication, and acting as catalysts for metabolic reactions. The key steps in the process of protein synthesis are transcription, pre-mRNA splicing and translation (shown in Figure 2.1). In transcription a gene is first copied into complementary *pre-messenger ribonucleic acid* (pre-mRNA) which then undergo *pre-mRNA splicing*. The pre-mRNA contains both coding (exons) and non-coding regions (introns). In pre-mRNA splicing the non-coding regions are removed and the resulting mature messenger RNA (mRNA) is then transported out of the nucleus of the cell into the *cytoplasm* of the cell. The cytoplasm contains essentially everything else in the cell apart from the nucleus. Here the mRNA molecule is read and translated into a protein which is a sequence of amino acids. The first nucleotide triplet of the mRNA encodes the first amino acid of the protein, the next triplet the second amino acid and so on. The rules by which the base sequence of the mRNA molecule is translated into the primary amino acid sequence of a protein are called the *genetic code*. The genetic code is universal and common for all living organisms. As a final step of protein synthesis the primary amino acid sequence of a protein is folded into a three-dimensional structure that determines the functional role of the protein [10].

An expressed gene is a gene that is transcribed into RNA. A cell in human body contains tens of thousands of potentially viable genes but all genes cannot be active at once, so cells must decide which genes to turn on and which to turn off. For instance, a bone cell turns on the genes that make it a bone cell, while a skin cell would leave those turned off. Neither of these cells need the genes that would allow cell differentiation into a neuron so these genes would be left turned off or unexpressed. The

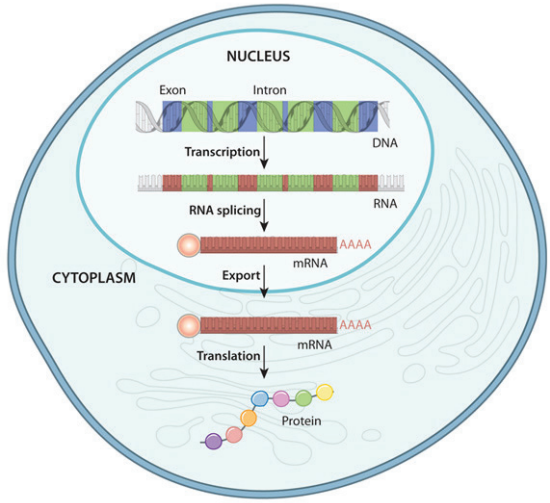


Figure 2.1. Three stages of protein synthesis: DNA is first transcribed into pre-mRNA. Next, the pre-mRNA is spliced to produce mature mRNA. The mature mRNA is finally translated into a protein using the genetic code that maps the nucleic acid triplets to amino acids. Copyright 2010 by Nature Education. Reprinted with permission.

cell types in a multicellular organism become different from one another because they synthesize and accumulate different sets of RNA and protein molecules. They generally do this without altering the sequence of their DNA. Most of the specialized cells are capable of altering their patterns of gene activation in response to extracellular cues [11]. The knowledge of which genes are expressed in a medical condition is also useful in many potential therapeutic applications. For instance, knowing which genes are expressed in cancer makes it theoretically possible to turn those genes off so that they cannot be active in the body. In this thesis the term *gene expression* is used as a synonym for transcription and most discussed work utilizes expression data samples collected from subjects under different biological conditions. The study of *transcriptomics*, also referred to as expression profiling, examines the expression levels of mRNAs in a given cell population, often using high-throughput techniques based on *DNA microarray technology*.

2.1.2 Layers of regulation

Most eukaryotic organisms, for example human beings, contain billions of individual cells. Almost all of these cells contain, within each nucleus, the entire genome for that organism. This genome contains the organism's complete hereditary information in the form of DNA, which encodes a

complete blueprint for all activities and structures within the organism.

The process by which a cell determines when and which genes it will activate is called gene regulation. It essentially allows the cell to control its function and adapt to environmental changes such as introduction of antibiotics into the environment of the cell. The gene activity is regulated at all levels of protein synthesis starting from pre-transcriptional control, to transcriptional control, to RNA processing and all the way till protein activity control [14].

To regulate genes, gene regulatory proteins need to gain access to the DNA which, with the help of packaging proteins called histones, is tightly packed into protein-DNA structures called chromatin. Chromatin occludes many DNA regulatory regions, not allowing them to regulate gene expression. At the pre-transcriptional control, chemical and structural modifications of the chromatin such as *methylation*, *acetylation*, and other histone-binding molecules affect the packing of the DNA molecule and may switch a gene on and off [15]. Certain of these modifications that regulate gene expression are believed to be heritable and constitute a major source of variation at individual and population level [16].

For most genes, transcriptional controls, controlling when and how often a given gene's DNA is transferred to mRNA, are paramount [14]. This is sensible because, of all the possible control points illustrated in Figure 2.1, only the transcriptional control ensures that no superfluous intermediates are synthesized. In transcriptional regulation, signals from the environment or from other cells activate proteins called transcriptional factors. These proteins bind to regulatory regions of a gene and increase or decrease the level of transcription.

At the post-transcriptional stage small nucleotide sequences called *micro-RNAs* can repress their target mRNA preventing protein translation [17]. Finally post-translational modifications, protein degradation and other mechanisms play a crucial role in generating heterogeneity in proteins and also help in utilizing identical proteins for different cellular functions in different cell types.

While many proteins perform their functions independently, the vast majority of proteins interact with others for proper biological activity in a cell, including cell growth, proliferation, inter-cellular communication and apoptosis. A critical test of requirement for a protein in a biological process is to inhibit its production by disrupting the corresponding coding genes [11]. The phenotypic changes are a result of coordinated activation

of multiple genes that regulate the life of an organism.

2.2 Gene expression measurement and comparison

Gene expression measurements provide an indirect view to the cellular process by recording the mRNA transcript levels in a cell population at a specific time and condition. Such global patterns of gene expression clearly show, for example, that liver cells transcribe a quite different set of genes than do white blood cells or skin cells [10]. Changes in gene expression can also be monitored during a disease process, in response to drugs or other external signals, and during development. Classical gene expression analyses compare one gene at a time by measuring how much mRNA is produced in the control treatment (e.g. a healthy subject) and how much mRNA is produced in the experimental treatment (e.g. a diseased subject). While powerful, these traditional approaches do not give a comprehensive view of the structure and activity of an organism's genome, its entire set of genes. With recent high-throughput techniques, scientists can measure activity of numerous genes at one time. This can yield a better understanding of how organisms are affected by changes in the gene expression.

2.2.1 Microarray measurement technology

There are several ways to measure the expression level of a gene within a cell and one powerful analytical tool is the microarray [12, 18]. DNA microarray technology is based on base pairing property of nucleic acid sequences where a DNA or RNA sample binds to complementary nucleotide sequence on the array. This allows to detect mRNA transcripts in the cell, thereby indicating which genes are being transcribed.

A microarray consists of DNA sequences called *probes* that are attached or synthesized in fixed positions on a solid surface (microscope slide or silicon chip). Probes are designed to uniquely match with particular mRNA sequences. To start a microarray experiment RNA is extracted from individual samples, reverse transcribed into a complementary DNA (cDNA) and labeled with a fluorescent dye. The resulting labeled transcripts are called *targets*. The labeled targets bind (*hybridize*) to the probes on the microarray with which they share sufficient sequence complementarity. The amount of the sample hybridized is used to esti-

mate the target mRNA concentration, and it is determined by measuring the intensity of light emitted by the labeled molecules with a laser scanner [10].

There are two types of microarrays that are most widely used today: *single-channel microarrays* and *dual-channel microarrays* [12]. Single-channel microarrays are hybridized with only one sample and therefore measure absolute expression levels of the mRNA sequences. In contrast, dual-channel microarrays are typically hybridized with transcripts from two samples (e.g. diseased tissue versus healthy tissue), where each sample is labeled with a fluorescent dye having a different emission wavelength. The two samples are mixed and hybridized on a single microarray that is then scanned to visualize fluorescence after excitation with a laser beam of a particular wavelength. Relative intensities may then be used in a ratio-based analysis to identify up-regulated and down-regulated genes between the two samples. Short *oligonucleotide* arrays [19] are the main sources of mRNA data in this thesis. These arrays consist of small fragments so that a transcript is not represented by one probe but by a set of them, typically $10-20$. Use of several probes for each target leads to more robust estimates of transcript activity [20].

Data preprocessing in microarray technology is a crucial initial step before data analysis is performed. The preprocessing involves a series of steps aimed at quality control (detecting irregularities in the arrays, rejection of erroneous spots), background correction (removal of signal emitted by other things than sample hybridized to probe) and normalization of the data (to correct for systematic biases due to causes such as different dye absorption, spatial heterogeneity in the chip, or others) [12]. For single-channel arrays a further step is necessary where different signals obtained from all probes representing one gene are summarized. The output of this initial process is the gene expression matrix with rows (1000-50000) representing genes and columns representing individual samples (typically from two to several hundreds). In statistical terms, the rows represent data variables (p) and columns represent individual observations (n). Manipulating such high-throughput data poses computational challenges for statistical learning that stem from the "large p small n" problem of having much fewer observations n than variables p.

2.2.2 Differential expression

Microarrays are, foremost, a tool of discovery [21]. Microarray gene expression data can be used in several different types of investigations. When several tissue samples of a certain type are hybridized in an experiment, it is natural to ask whether the samples can be grouped in homogeneous subtypes. Unsupervised clustering and classification methods are commonly used to group samples with similar expression patterns across genes or to cluster genes that follow the same expression patterns across a set of samples [22]. Computational modeling to detect survival-associated gene-expression-based biomarkers [23, 24], and modeling to detect and connect diseases to drug responses at the transcriptional level [25] are also an active area of research with implications to personalized medicine [26–28].

Often one of the first tasks in analysis of microarray data is to identify the genes that are *differentially expressed* i.e. whose expression levels change between two phenotypes. For instance, to understand the effect of a drug we may ask which genes are up-regulated (increased in expression) or down-regulated (decreased in expression) between treatment and control groups. A conventional statistical analysis method for differential expression is to examine one gene at a time, determine a p-value that the gene is differentially expressed in different phenotypes (e.g. by comparing means across two groups using student's t-tests, or with the *linear models for microarrays* [29]), and then to apply a correction (penalty) to the p-value for having tested multiple genes [12]. These methods work best when individual genes have large effects and there is a very consistent effect for subjects within each single phenotype. However, when a biological pathway is up-regulated or down-regulated, individual genes in the pathway may not show consistent, statistically significant effects in different samples. More subtle, coordinated changes in members of a *set* may be more easily detected overall across the set than detecting the change in a single member [30–32].

Gene set tests are designed to address these limitations of single gene analyses and to bring in biological knowledge in the form of pre-defined gene sets [33]. These are statistical methods which are often used to determine if predefined sets of genes are differentially expressed in different phenotypes. The tests are based on the notion that genes within the gene sets are funtionally related and, hence, will have similar expression

patterns. These expression patterns might be modest, yet by borrowing strength across gene set, there is potential for increased statistical power [34]. In addition, in comparing results on the same disease from different laboratories, one might get more reproducible results [33, 35]. The gene sets are defined based on prior biological knowledge, such as set of co-expressed genes in a previous experiment, genes in a known pathway, for instance from the KEGG pathway database [36], or from publicly available descriptions of biological processes, such as a Gene Ontology category [37].

The gene set methods can be broadly divided into two categories: *competitive* and *self-contained* tests [30, 38]. Competitive gene set tests compare genes in a test set relative to all other genes. These tests focus more on distinguishing the most important biological process from those that are less important. Self-contained tests, on the other hand, examines a set of genes against the null hypothesis that no genes are differentially expressed. The test evaluates the relevance of an individual biological process to the experiment under consideration. The self-contained null hypothesis may not always be biologically interesting in data sets where there are many differentially expressed genes, for example when comparing cancer versus normal. This is due to the self-contained null hypothesis where a gene set is considered to be differentially expressed even if only one of its genes is effectively differentially expressed.

Publications I, IV, III and VII make use of the Gene Set Enrichment Analysis (GSEA; [33, 34]), a competitive gene set test. Details of the GSEA method are discussed in Section 4.2.1.

2.3 Genomic data resources

Gene expression measurements are one of the most widely available unique data resources. These measurement collections are maintained by several public repositories, including *ArrayExpress* [3] and *Gene Expression Omnibus* (GEO) [39]. There also exist carefully controlled integrative datasets that contain thousands of genome-wide measurements of transcriptional activity across diverse conditions in a directly comparable format, such as [25, 40]. These aforementioned repositories have been used in Publications I, IV, III and VII.

In addition to gene expression, microarray-based techniques can also be used to study other functional aspects of the genome, including micro-

RNA regulation [41], alternative splicing [42], transcription factor binding [43], and different *structural variations* in the genome [44]. Structural variations in the genome play a crucial role in genetic diseases, such as cancer development and progression [45]. They are variations in the structure of a chromosome and typically each such variation affects the sequence over a length of about one kilobase to several megabases. A large category of structural variation is *copy number variations* (CNVs). The CNVs can be detected from the genome using, for example, microarray-based techniques such as array-comparative genomic hybridization (aCGH) [46]. A CNV corresponds to large regions of the genome that have been deleted or duplicated on certain chromosomes. Each deleted or duplicated region can be limited to a single gene or include a contiguous set of genes. The variations can result in having either too many or too few of the dosage-sensitive genes, which may be responsible for a substantial amount of human phenotypic variability, complex behavioral traits, and disease susceptibility [47, 48]. The Cancer Genome Atlas (TCGA; [49]) consortium provides a semi-public repository that contains copy number measurements paired with gene expression and other data types across several different cancer types. A subset of the repository is utilized in Publication V, where DNA copy number changes are integrated with transcriptional profiling data to discover potential survival-associated biomarkers for an aggressive brain tumor.

3. Probabilistic Modeling

Modern computer science allows processing of very large amounts of noisy data. A typical example of such data are medical genomic samples that are measured from subjects under diverse biological conditions and stored in different public repositories. Most publicly available biological data collections do not contain complete information about patient history, how the different pathways were active and what was the eventual diagnosis. However, it is believed that there is a hidden generative process that explains the observed data. For example, a subject with cancer will not show differential expression for random genes, instead the measured data are likely to reveal accumulated alterations of multiple genes having similar functions and belonging to pathways critical to cancer, such as cell growth pathways [50] and normal cell behavior pathways [51]. The processes underlying the observed data may not be completely identifiable but it is possible to construct a good and useful approximation for them and detect certain patterns and regularities. This is the field of machine learning and *probabilistic data modeling* [52, 53]. By probabilistic modeling of the data one can specify a set of assumptions or prior knowledge about the nature of relationships in a data collection [7]. Then by computing how the data fits the model it is possible to assess the model performance ([54]) and devise rigorous measures of relevance among different data samples [55].

This chapter discusses key principles of probabilistic modeling and provides the methodological background for subsequent chapters; Section 3.1 introduces basic concepts in probability theory, Section 3.2 presents some useful models for density estimation, while Section 3.3 discusses associated learning and inference techniques. For other broad reviews of statistical machine learning and probabilistic modeling, see [7, 52–54, 56].

3.1 Basic probability theory

Observations of complex real-world phenomena contain large amount of uncontrolled variation called *noise*. The sources of noise in biological data include experimental, measurement, reporting, annotation and data processing errors [57]. The noise and the finite size of observed datasets give rise to *uncertainty* about the phenomena underlying the observations. The calculus of uncertainty is called *Probability theory* [56].

A fundamental concept in probability theory is a *random variable* whose value is subject to variation by chance [6]. As opposed to other mathematical variables, a random variable conceptually does not have a single, fixed value; rather, in observations the variable can take on a set of possible different values, each with an associated probability. Random variables can be classified as either *discrete* (that is, taking any of a specified list of exact values) or as *continuous* (taking any numerical value in a possibly infinite interval or collection of intervals). For example, consider a discrete random variable X that represents a result of a coin toss; in this case the probabilities of the possible values are $P(X = \text{tail}) = 0.5$ and $P(X = \text{head}) = 0.5$. Notice that the probabilities for the different values k of a random variable satisfy $\sum_k P(X = k) = 1$. Although the expression $P(X = k)$ helps to avoid ambiguity, it leads to a rather cumbersome notation. Instead, in the thesis $P(X)$ is used to denote a distribution over the values of random variable X, and $P(k)$ is used to represent the distribution evaluated at a particular value k. For continuous random variables the probability over events, corresponding to intervals of values, is defined as integrals of a probability density, for example $p(X \in (a, b)) = \int_a^b p(X)dX$. With this compact notation let X be an event that a hypothetical oncogene (a gene implicated in cancer) is expressed at high levels in a person and let Y represent the event that the person has cancer. Then $p(X, Y)$ is the *joint probability* that the person has tumor and has the indicator gene over-expressed, while $p(X)$ and $p(Y)$ are the *marginal probabilities* of showing an over-expressed gene or developing cancer, respectively. The conditional probabilities are defined as follows:

$$p(X|Y) = \frac{p(X, Y)}{p(Y)}, \tag{3.1}$$

$$p(Y|X) = \frac{p(X, Y)}{p(X)}, \tag{3.2}$$

where the first conditional probability, $p(X|Y)$, is the probability of the gene being over-expressed given that the person has cancer. It is easy to

estimate by simply taking the fraction of cancer subjects that have the indicator gene over-expressed, and estimating the conditional probability with that fraction (relative frequency); such approximation can work well for large numbers of subjects because of the "law of large numbers" (see e.g., [58]). The two random variables X and Y are *independent* if $p(X|Y) = p(X)$ and $p(Y|X) = p(Y)$. In Equation (3.2) the second conditional probability, $p(Y|X)$, is the probability that a person carries cancer given that his or her indicator gene is over-expressed. The probability, though more important from a diagnosis perspective, may be difficult to estimate directly as a fraction of subjects. However, it can be expressed in terms of the earlier complementary conditional probability by substituting the joint probability from Equation (3.1) in Equation (3.2), that is,

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}. \tag{3.3}$$

Equation (3.3) is commonly known as the *Bayes' theorem* and it expresses an unknown conditional probability in terms of another complementary and easy-to-compute conditional probability [54]. In the example $p(Y|X)$ is relatively easier to estimate since the conditional probability $p(X|Y)$ and the two marginal probabilities $p(X)$ and $p(Y)$ are easily available from global statistics. Consequently, the conditional probability $p(Y|X)$ can be determined without directly estimating it.

3.2 Probabilistic modeling

A dataset $\mathcal{D}$ usually contains a set of independent observations or samples, $\mathcal{D} = \{x_i\}_{i=1}^{N}$. A common approach to model the data is to assume that it is drawn from an unknown probability distribution $P(\mathcal{D})$. This approach is usually referred to as density estimation and it can be used to summarize the data and cater for uncertainty in the observations [53]. A simple and standard method to density estimation involves choosing a specific form for the density and specifying it in terms of a parametric model $P(\mathcal{D}|\theta)$. When the data are observed, the probability of observations given a set of parameter values θ can be expressed as the *likelihood function* $\mathcal{L}(\theta; \mathcal{D})$. When each observation is assumed to be independently drawn and identically distributed (IID) the likelihood function becomes

$$\mathcal{L}(\theta; \mathcal{D}) = \prod_{i=1}^{N} P(x_i|\theta).$$

As an example consider a textual document $\mathcal{D} = \{w_i\}_{i=1}^{N}$ that contains N words, and we would like to model the document by setting probabilities for how many times a unique word w' appears in the document. The density of the observations can be modeled via a multinomial distribution over the counts: $\mathcal{L}(\theta; \mathcal{D}) = \prod_{i=1}^{N} P(w_i|\theta) = \prod_{w'=1}^{V} \theta_w^{c_{w'}}$, where $c_{w'}$ is the count of unique word w' in the data $\mathcal{D}$.

The parameters that best allow the model to represent and summarize the data can be obtained by *inference* [56]. Inference refers to estimating the unknown probabilities θ from a set of training data $\mathcal{D}$. A well-known method of estimating the model parameters is the *Maximum Likelihood estimation* (ML) which sets the parameters such that the likelihood function is maximized,

$$\widehat{\theta_{ML}} = \arg\max_{\theta} \mathcal{L}(\theta; \mathcal{D}) = \arg\max_{\theta} \prod_{i=1}^{N} P(x_i|\theta). \tag{3.4}$$

The ML estimate θ_{ML} can be obtained by setting the derivative of the likelihood to zero $\frac{\partial \mathcal{L}}{\partial \theta_w} = 0, \forall\, \theta_w \in \theta$ and solving for the values of θ_w.

In probability theory a measure of difference between two probability distributions is the *Kullback-Leibler* divergence, often abbreviated as KL-divergence [59]. The KL-divergence of distributions $q(x)$ from $p(x)$ is defined as $d_{KL}(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$, which can be interpreted as the average inefficiency (measured in bits) of assuming that the distribution is $q(x)$ when the true distribution is $p(x)$ [60]. In Equation (3.4) the likelihood function is chosen for optimization or fitting the model. This is intuitively appealing since if the chosen model $P(\mathcal{D}|\theta)$ differs from the true distribution, maximization of the likelihood corresponds to minimization of the KL divergence between the empirical distribution and the model [53]. Effectively, this results in a trained model that approximates the empirical distribution subject to the constraints of the model family. Other ways of learning a model are discussed in Section 3.3.

3.2.1 Mixture models

Classical probability distributions provide a well-justified approach to model the observed data, but in many practical situations the useful regularities in the data cannot be described with a single standard distribution. In such scenarios a superposition of multiple distributions can provide the ability to represent arbitrarily complex distributions over the data, and the overall probability density of the data can be modeled as a weighted

mixture of k components,

$$P(\mathbf{x}) = \sum_{k=1}^{K} \pi_k P_k(\mathbf{x}|\theta_k). \tag{3.5}$$

Such models, termed mixture models, can contain different distributions for each component and each component distribution $P_k(x|\theta_k)$ has its own parameters θ_k (e.g. mean and variance for a component with normal distribution) [53]. The parameters π_k are called the *mixing coefficients* or weights; they are non-negative and sum to one ($\sum_{k=1}^{K} \pi_k = 1$) in order to be valid probabilities. In practice, the mixing coefficients are often unknown and can be estimated from the observed data by considering them as standard model parameters fitted with a ML estimate. Publication VII utilizes the mixture model formulation to propose a mixture of background models (detailed in Section 6.3).

3.2.2 Latent variable models

A way to describe more complex probability spaces and model hidden variables that generate the observed data is to introduce latent variables $\mathbf{z} = \{z_1, z_2, ...\}$, where each variable z_k describes a simple distribution [53, 56]. The latent variables are not directly observed and provide a flexible way to express dynamic dependencies between other variables.

A latent variable model is defined by specifying the joint distribution over the latent variables $\mathbf{z} \in \mathbb{R}^K$ and the observations $\mathbf{x} \in \mathbb{R}^D$. Considering again the mixture model from Equation (3.5), we can formulate the same model by introducing a K-dimensional latent variable, $\mathbf{z} = \{z_1, z_2, ...z_K\}$, where $z_k = 1$ if the observation belongs to the k^{th} mixture and all other element of $\mathbf{z}$ are equal to 0. The marginal distribution over $\mathbf{z}$ is specified using a multinomial distribution from k categories with prior probabilities π_k, such that $P(z_k = 1) = \pi_k$. Then the prior distribution over the latent variables $\mathbf{z}$ can be specified as

$$P(\mathbf{z}) = \prod_{k=1}^{K} \pi_i^{z_k}.$$

The joint distribution is decomposed as $P(\mathbf{x}, \mathbf{z}) = P(\mathbf{z})P(\mathbf{x}|\mathbf{z})$ where $P(\mathbf{x}|\mathbf{z})$ is a conditional distribution which expresses the uncertainty in the observations given the mixture component that generated it:

$$P(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} P(\mathbf{x}|\theta_k)^{z_k}.$$

Depending on the nature of the observed data θ_k would correspond to parameters of a specific distribution, e.g., for continuous variables it would represent mean and variance of a Gaussian distribution (as used in Publication V to model shared effects between gene expression and copy number variations) or for multivariate count data it could represent probabilities of observing a particular count value in a multinomial distribution (as used to model latent components in Publication I, III, and VI).

The joint distribution is given by $P(\mathbf{z})P(\mathbf{x}|\mathbf{z})$ and the *marginal distribution* of the observation is obtained by summing out the joint distribution over all possible values of the other variable, that is,

$$P(\mathbf{x}) = \sum_{\mathbf{z}} P(\mathbf{z})P(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} P(z_k = 1)P(x|z_k = 1).$$

Here $\mathbf{z}$ is marginalized out. The marginal distribution is equivalent to the earlier formulation in Equation (3.5). The original data $\mathbf{x}$ can now be expressed in terms of a smaller number of latent variables $\mathbf{z}$ that can be obtained by using the Bayes rule:

$$P(z_k|\mathbf{x}) = \frac{P(\mathbf{x}|z_k = 1)P(z_k = 1)}{\sum_{k=1}^{K} P(z_k = 1)P(\mathbf{x}|z_k = 1)}.$$

If there are several observations $\{\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_N\}$, then for each observed data point $\mathbf{x}_i$ there is a corresponding latent variable $\mathbf{z}_i$. Let $\mathbf{X} \in \mathbb{R}^{N\times D}$ represent the entire data containing N observations and D features and let $\mathbf{Z} \in \mathbb{R}^{N\times D}$ denote the corresponding latent variables with columns $\mathbf{z}_k$, then the log-likelihood function of the entire data is given by

$$P(\mathbf{X}|\theta) = \log \sum_{\mathbf{Z}} P(\mathbf{X}, \mathbf{Z}|\theta).$$

An important observation is that the summation over latent variables appears inside the logarithm and therefore even if the joint distribution $P(\mathbf{X}, \mathbf{Z}|\theta)$ has an analytical solution, the marginal distribution typically does not.

3.2.3 Bayesian modeling

In Bayesian data analysis the uncertainty in the unknown parameter values (or different models) is quantified before making inferences from the data. The uncertainty is described in terms of probability distributions [54]. This is in contrast to the traditional point estimation approach where a particular parameter value is learned instead of a distribution for the parameter. In Bayesian framework one defines a prior probability

distribution for possible values of the unknown parameters. These prior probabilities can be based on earlier observations, knowledge given by an expert of application domain or can be so-called *uninformative* priors that try to make the least amount of assumptions about which parameter values are likely before seeing the data. After specifying the priors, one constructs a posterior distribution of the unknown parameter values θ conditioned on observed data $\mathbf{X}$

$$P(\theta|\mathbf{X}) = \frac{P(\mathbf{X}|\theta)P(\theta)}{P(\mathbf{X})}.$$

The prior can be seen as a penalty term and it favors solutions that match with the prior assumptions; in particular, many priors give more probability to simpler models. Such regularization properties are useful when the data are scarce and there is high uncertainty in the parameter estimates. The prior predictive density $P(\mathbf{X})$ is a normalization constant which is independent of the parameters θ and can often be ignored during model fitting. In Bayesian modeling one is typically interested in deriving an estimate for the posterior distribution over different models rather than one fixed model, as is the case of point-estimates. If a point-estimate is needed, however, it can be taken by seeking the maximum value of the posterior distribution $P(\theta|\mathbf{X})$, which is usually referred to as the *Maximum-a-posteriori* estimate (MAP):

$$\widehat{\theta}_{MAP} = \arg\max_{\theta} \frac{P(\mathbf{X}|\theta)P(\theta)}{\sum_{\theta} P(\mathbf{X}|\theta)P(\theta)} = \arg\max_{\theta} \prod_{i=1}^{N} P(\mathbf{x}_i|\theta)P(\theta). \tag{3.6}$$

Unlike in ML estimation, in MAP inference of a single parameter, the prior can act as a regularization. The regularization ensures that events that have not occurred in the data so far do not necessarily have zero probabilities but a value depending on the prior. This property of the prior can help prevent over-fitting of learned models. Nevertheless, the posterior distribution can have many peaks and estimating just the highest peak might be misleading. The Bayesian viewpoint is to use all possible models to draw inferences (or to evaluate the predictions in a prediction task) and weight them by their respective posterior probabilities. This means inferences will be affected by regions of the posterior distribution where the probability mass is large rather than only the highest value of the probability density. Since the evaluation of the posterior distribution typically involves integration of complicated functions, it is rare that a closed form or analytical solution is available. Therefore, the usual way to learn and evaluate the Bayesian models is either by Markov Chain Monte Carlo

(MCMC) sampling or approximation of the posterior distribution by variational inference [53, 54, 56]. The methods in Publications I, II, III, VI and the background models in Publication VII are based on Bayesian latent variable models.

3.2.4 Nonparametric modeling

The finite mixture models or classical latent variable models require a pre-specified model structure where the number and distributional shape of generative processes is known before any data analysis. This is problematic in many practical tasks: for instance, in mixture modeling the number of mixture components and in network reconstruction the structure of the network are not known before the analysis. In such scenarios, a typical solution is the *model selection* over a finite set of candidate models [54]. In model selection each model is evaluated using validation data based on an evaluation criterion. Modern nonparametric Bayesian models provide an alternative and principled approaches to learn the model structure from data [61]. These models are based on *nonparametric priors* (discussed in Section 6.2.2) that allow the number of mixture components to grow in order to accommodate the complexity of data. Publication VI introduces a new nonparametric Bayesian hierarchical model which is suitable to relate sets of datasets.

3.3 Learning and inference

Learning a model is the process of updating probabilities of outcomes based upon the relationships in the model and the evidence collected from the observed data. It focuses on learning the model parameters θ. In statistical machine learning inference refers to estimating the posterior distribution of the latent variables. Various optimization schemes are available to learn statistical models; however, there are several potential challenges including, a) the danger of finding only poor local optima, b) computational complexity due to limited resources, c) un-identifiability arising from a complex model structure, and d) ultimately the uncertainty remaining after the inference stemming from lack of sufficient data. This section focuses on learning procedures that are central to the thesis: Expectation Maximization (EM; [62]) and approximate inference [54, 56].

3.3.1 Expectation Maximization

EM is an algorithm for maximizing a likelihood function for probabilistic latent variable models. In these models there is no simple analytical form for ML estimates of parameters because the likelihood function $p(\mathbf{X}|\theta)$ has a complicated expression due to marginalization over the latent variables $\mathbf{Z}$ in the complete-data likelihood $p(\mathbf{X}, \mathbf{Z}|\theta)$, that is,

$$p(\mathbf{X}|\theta) = \int_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta) d\mathbf{Z}.$$

Effectively, there is a chicken-and-egg problem: to solve for the model parameters in an analytical fashion one needs to know the distribution of the latent variables, but the distribution of the latent variables $p(\mathbf{Z}|\mathbf{X}, \theta)$ is a function of model parameters. EM tries to get around this by iterating between estimation of the posterior of the latent variables and optimization of the model parameters. If the current estimate of the model parameters is denoted by θ_{old}, then a pair of successive *expectation* (E) and *maximization* (M) steps give rise to a revised estimate θ_{new}. In the E-step the algorithm evaluates the expectation of the complete-data log-likelihood over the posterior density of the latent variables, $p(\mathbf{Z}|\mathbf{X}, \theta_{\text{old}})$, keeping θ_{old} fixed,

$$\mathcal{Q}(\theta, \theta_{\text{old}}) = \int_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta_{\text{old}}) \log p(\mathbf{X}, \mathbf{Z}|\theta) d\mathbf{Z}.$$

In the subsequent M-step this posterior distribution is used to find a revised point estimate for the model parameters θ_{new} by maximizing the function $\mathcal{Q}$ so that

$$\theta_{\text{new}} = \arg\max_{\theta} \mathcal{Q}(\theta, \theta_{\text{old}}).$$

The E and M steps are repeated until convergence of either the parameter values or the log-likelihood. The EM algorithm can be understood as optimizing a lower bound on the log-likelihood [53].

The EM algorithm is particularly useful when it is possible to derive closed-form updates for both the E and M-steps. The algorithm can be used to find MAP solutions for the models by incorporating prior information about the parameters $p(\theta)$. In this case the E step remains the same as in the maximum likelihood case, whereas in the M step the expression to be maximized is given by $\mathcal{Q}(\theta, \theta_{\text{old}}) + \log p(\theta)$. This essentially avoids singularities (for some parameter values, which can yield zero probabilities for some future data) and over-fitting by focusing the modeling on particular features in the data, as in the mixture of unigrams model in

Publication VII, and the regularized dependency modeling framework of Publication V, respectively.

3.3.2 Approximate inference and Gibbs sampling

A central challenge in Bayesian modeling is to evaluate the posterior distribution over the model parameters θ and the latent variables z. The dimensionality of the latent space is often too high to work with the posterior distribution directly. In addition, the distribution may have highly complex forms for which expectations are not analytically tractable. For discrete variables, the marginalization needed to evaluate posterior probabilities involves summing over all exponentially many possible configurations of the latent variables, which is a prohibitively expensive operation, while for the case of continuous variables, the necessary integration may not have a closed-form analytical solution, and the dimensionality of the space may prohibit numerical integration [53]. In such situations, *approximate inference* makes it possible to learn a model. Approximate inference approaches fall into two broad classes; deterministic or stochastic approximations. Deterministic schemes are based on analytical approximations to the posterior distribution, for example by assuming that the posterior distribution factorizes in a particular way [63]. As such they are less likely to generate the exact result, but some of these schemes scale well to large applications. Stochastic techniques have the property that given infinite computational resources, they can generate exact results, and the approximation arises from the use of finite computational time [54]. Here *Gibbs sampling* is described, which has enabled wide-spread use of Bayesian methods across several domains [64].

Gibbs sampling works by successively simulating observations that are approximated from a joint posterior distribution $p(\mathbf{z}|\mathbf{X}, \theta)$ without requiring to directly sample from the joint distribution. By simulating sufficiently many (independent) samples $\mathbf{z}^1, \mathbf{z}^2, ...\mathbf{z}^T$ the mean, variance or any other characteristics of a function $f(\mathbf{z})$ can be evaluated to the desired degree of accuracy [65]. The simulated samples follow the true posterior distributions and can be used to compute the population quantities; for example those samples can be used to evaluate expectation of some function $f(\mathbf{z})$ as

$$\mathbb{E}_{p(\mathbf{z}|\mathbf{X},\theta)}\left[f(\mathbf{z})\right] = \int f(\mathbf{z})p(\mathbf{z}|\mathbf{X},\theta)d\mathbf{z} = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{M} f(\mathbf{z}^{(t)}).$$

A Gibbs sampler has been found useful in many multidimensional

problems. Suppose $\mathbf{z}$ is divided into d components or sub-vectors $\mathbf{z} = (z_1, z_2, ...z_d)$. To sample from the posterior $p(\mathbf{z}|\mathbf{X}, \theta)$, a Gibbs sampler is defined in terms of components of $\mathbf{z}$. At each iteration the sampler cycles through the components and draws each component from its conditional posterior distribution given all other components at their current values:

$$z_j^t \sim p(z_j|\mathbf{z}_{-j}^{t-1}, \mathbf{X}, \theta),$$

where $\mathbf{z}_{-j}^{t-1} = (z_1^t, ..., z_{j-1}^t, z_{j+1}^{t-1}, ..., z_d^{t-1})$. There are thus d steps in iteration t. In each iteration, an ordering of the d components is chosen and, in turn, each z_j^t is sampled. The samples $\mathbf{z}^{(t)}$ are drawn sequentially with the distribution of the sampled draws depending on the last value drawn; hence the draws form a Markov chain, where the approximate distributions at each step are improved in the simulation, in the sense of converging to the target distribution [54]. In practice, successive samples from the Gibbs sampler are strongly dependent and in order to obtain independent samples from the desired posterior distribution $p(\mathbf{z}|\mathbf{X}, \theta)$ one must discard the initial samples as part of the *burn in* period before the sampler reaches its stationary distribution and use *thinning* to obtain lagged samples. The particular appeal of Gibbs sampler stems from the most common scenarios where even though the joint posterior density of all variables is analytically intractable to compute, it is possible to easily sample from most or all conditional posterior distributions of the parameters.

An extension of Gibbs sampler used in conjugate models is the collapsed Gibbs sampler, which marginalizes out any nuisance model variables ψ that are not of direct interest, so that $p(\mathbf{z}|\mathbf{X}, \theta) = \int p(\mathbf{z}, \psi|\mathbf{X}, \theta) d\psi$ [54]. These variables ψ can be later estimated using the obtained Gibbs samples. The reason for marginalization of model parameters is that when we do not need to sample extra parameters, effectively only a subspace is sampled where information is updated sooner and so the Markov chain converges faster to the stationary distribution. Collapsed Gibbs sampling was used in all publications except Publications IV and V.

4. Model-based retrieval for data samples

The availability of thousands of expression studies in public repositories makes it increasingly challenging to notice good results among non-relevant results. At the same time, these repositories give us the opportunity to develop retrieval and exploration methods that use the gene expression data from the collection to deliver biologically meaningful results. This chapter introduces model-based solutions for the scenario where a researcher has a new sample or a set of potentially interesting genes and would like to find earlier data samples that are most relevant for the new sample. The chapter starts with the classical approach to the retrieval problem and the related motivation of the data-driven search. The next two sections describe specific contributions towards two model-based data-driven retrieval solutions; the first solution is useful when a researcher has a sample of interest and would like to search for earlier samples that are most relevant to the sample of interest (Publications I, III), the second solution is suitable for the case when the researcher, instead, has a few genes or gene sets of interest and would like to infer interactions among them, and utilize them to find the most relevant earlier data samples (Publication II).

4.1 Motivation and Related work

Most traditional search engines provide the user with a non-data-driven search facility [3, 4], where the user issues a query as text either directly by typing one or more keywords or by selecting an ontology term of interest from a pre-defined ontology. The basic content-based search engines work by computing a distance function for the input query text against the textual descriptions of background data. This approach not only requires carefully written descriptions with controlled vocabularies and standard-

ized practices [1] but also limits the potential findings to existing knowledge. Usually the researcher who is interested in searching the public databases is looking for data that could complement his/her existing measurements or to investigate with the aim of discovering something new. Data-driven approaches, that let the measured data speak for themselves, provide potentially better methods where the inferred relevant data corresponds to statistical similarity in the actual measurement data. The next two sections summarize contributions of the thesis to the model-based data-driven retrieval methods where a researcher can position her own measurement data into the context of earlier biology.

4.2 Retrieval of relevant samples given a query sample

Given a query gene expression profile, Publications I and III provide probabilistic models that inherently yield model-based similarity measures between gene expression profiles. These models have been designed to retrieve data relevant to the user query in the sense that the retrieved profiles exhibit similar patterns of expression levels with the aim that the retrieval results are interpretable. Both existing approaches and the two proposed ones can be seen as instances of a general purpose retrieval framework that can be divided into four components. In the first component, each gene expression study is decomposed into meaningful comparisons between biological conditions and each comparison is represented in terms of differential expression for genes or gene sets. In the second component, biologically meaningful patterns of expression are extracted using appropriate probabilistic modeling methods. In the third component, a relevance measure quantifies similarity between any two profiles and can be used to retrieve a ranked list of most relevant results given the input query. In the fourth component, a model selection procedure is used to select an appropriate model among a set of candidate models. In the following subsections each of the four main parts are described.

4.2.1 Study decomposition and representation

A background database contains microarray datasets or studies submitted by researchers. Each study is based on an experimental design and can contain multiple samples. In the first step of the proposed methods the experimental design is decomposed into a set of comparisons between

pairs of conditions, such that it minimizes the effect of confounding factors and increases the interpretability of results. In the ArrayExpress database [3] an experimental design of a study is translated into different experimental factors e.g. “disease state”, “compound” or “tissue”. A given sample in a study has been measured in a condition having a specific value for each experimental factor. For instance, a sample may have the annotation “disease state = normal” and “tissue = heart”. For every study a list of comparisons is derived such that within each comparison all experimental factors share the same value, except for a single factor which has either of the two possible values. For example, a given set of samples may share the annotation “tissue = heart” and “gender = male” but may have the two possible annotations for “disease state = muscular dystrophy” or “disease state = normal”. This results in a comparison between muscular dystrophy and normal samples in the context of “gender = male”, tissue = “heart” and compound = “none”. Alternative methods that compare a sample against the average of all samples within a study, e.g. the Module Maps method [32], yield comparisons that depend on the conditions used to compute the average and therefore contain an additional layer of study-specific bias.

After decomposing each experiment into a set of comparisons between biological conditions, the differential expression patterns for each comparison can be represented as differential expression profile across genes or a set of pathways. Concretely, GSEA is used to extract the differential expression of a set of manually curated pathways obtained from the MSigDB collection [33]. For each comparison, genes are sorted with respect to their differential expression levels and a running sum is computed over the sorted list; this running sum, known as the enrichment score (ES), increases when a gene belongs to the gene set and decreases otherwise; the final statistic is the maximum of this sum. For each sample the ES is normalized by dividing it by the mean of random ES’s computed by permuting the phenotype labels of the samples. The top scoring gene sets are selected according to this normalized score. Finally, to represent the comparison, each gene set is associated with an integer value that corresponds to the number of genes in the gene set that are found before the running score reached its maximum (called the Leading edge subset). This effectively yields a bag-of-words representation for each sample in a dataset. The use of gene set level tests rather than gene level tests is not only due to the fact that procedures used to test for differential ex-

pression of gene sets are observed to be more robust across studies [33], but also because gene sets allow re-using existing biological knowledge of pathways.

4.2.2 Probabilistic Topic models

To model the background collection of comparisons, each represented by the GSEA output, the proposed pipeline uses probabilistic latent variable mixture models. The models are used to infer biologically meaningful co-activations between patterns of differential expression and their generative nature provides a basis for a sensible relevance measure between a given query and each background comparison. In Publications I and III, the latent variable mixture models come from the *topic model* family. In topic models the latent variables represent multiple components or topics in a comparison and each comparison can have a mixture of components. Topic models have been successfully used in textual information retrieval [66, 67]. They are unsupervised probabilistic models that provide useful descriptive statistics for analyzing and understanding the latent structure from count data. In count data, each object is represented as a vector whose elements contain counts of how many times a particular event occurred in the object, such as text from documents assumed to be in a bag-of-words representation, where for each document the numbers of occurrences of each word are counted. The latent structure is captured by a fixed number of latent mixture components and a distribution over the components, for each document, that is most likely to have generated the observed data of a document (details below). Publication I utilizes one of the most simple topic models, namely, the Latent Dirichlet allocation (LDA; [66]) where components aim to capture co-occurrence patterns among gene sets. In Publication III an extended model is presented that captures co-occurrence among gene sets coupled with co-occurrence among genes. The intuition behind both models is to infer components that are groups of gene sets expressed together in a similar fashion across different comparisons. Since the models infer a latent structure (i.e. per comparison distribution over the inferred components and per component distribution over gene sets) it is possible to interpret two potentially relevant profiles by examining their respective latent structures.

In Publication I, the classical LDA method is used to model the background collection of comparisons $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, .., \mathbf{x}_D]$. The LDA models each comparison $\mathbf{x}_d$ as a mixture over K components. A component k is the

central concept in this type of modeling and each component is built as a multinomial distribution over the gene sets that are often co-activated together across the comparisons. For each gene set, the level of activity of the gene set in a sample is represented as a count of activations. Each comparison is generated as a set of activations of the various gene sets. Within each comparison, each activation i of a gene set is generated by first picking a component index $z_{d,i}$ using the comparison-to-component distribution θ_d, that is, $z_{d,i} \sim \text{Multinomial}(\theta_d)$. Given the component index, the index $x_{d,i}$ of the gene set to be activated is generated using the component-to-gene-set distribution ψ_k, that is, $x_{d,i} \sim \text{Multinomial}(\psi_k)$. Lastly, the count of activations of the chosen gene set is increased by one. This procedure is repeated to generate more activations within the comparison, until the desired total number of activations has been reached. The probability of generating gene sets activations $\mathbf{x}_d = \{x_{d,i}\}_{i=1}^{N_d}$ for a single comparison, given ψ and θ_d, is

$$p(\mathbf{x}_d|\psi, \theta_d) = \prod_{i=1}^{N_d} \sum_{z_{d,i}=1}^{K} p(z_{d,i}|\theta_d) p(x_{d,i}|z_{d,i}, \psi) \tag{4.1}$$

where $z_{d,i}$ is marginalized out. Conjugate Dirichlet priors are placed over both the distribution over components $\theta_d \sim \text{Dirichlet}(\alpha)$ and the distribution over gene sets $\psi_k \sim \text{Dirichlet}(\beta)$. A plate diagram is presented in Figure 4.1. The complete-data likelihood for a single comparison d can be specified as:

$$p(\mathbf{x}_d, \mathbf{z}_d, \psi, \theta_d|\alpha, \beta) = \prod_{i=1}^{N_d} p(x_{d,i}|\psi_{z_{d,i}}) p(z_{d,i}|\theta_d) p(\theta_d|\alpha) p(\psi|\beta). \tag{4.2}$$

Inferring the latent parameters of the generating distributions in the model from a set of observations is not possible in closed form and therefore approximate inference using a collapsed Gibbs sampler is used to infer the latent parameters.

In Publication III the classical LDA is extended to 1. model the activity of gene sets as well as specific genes within gene sets and 2. model correlation between components via so called modules. To accomplish the first extension, an alternative representation is used where each comparison contains two pieces of information for every gene set: binary activation of the gene set and for each gene set a binary vector specifying if a gene belongs to the leading edge subset of the gene set. To accomplish the second extension, a hierarchical component structure is developed where each comparison i has a distribution over so-called modules θ_i and each

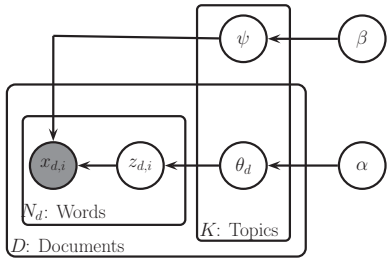


Figure 4.1. Plate diagram of Latent Dirichlet allocation model. Rectangles indicate sets of variables, with cardinality of the set marked in the bottom left corner. Gray nodes correspond to observed data.

module in turn has a distribution over components (or submodules) η_m. A module represents a combination of components and aims at capturing high-level biological phenomenon. A gene set s is generated by first picking a module index $u_{i,g} = m$ and a component index $v_{u,g} = k$ (from the corresponding module's distribution η_m), then the gene set is activated with probability $\phi_{s,k}$ given the selected component and finally each gene g within the leading edge subset of the gene set is activated with probability $\psi_{k,g}$. Conjugate priors are used to integrate out the model parameters, and a collapsed Gibbs sampler is used to perform inference for latent variables $\mathbf{u}$ and $\mathbf{v}$.

4.2.3 Probabilistic measure of relevance

To address the problem of retrieving the most meaningful and relevant samples, one needs to specify a similarity measure between a given query comparison and each comparison from the background collection. While classical correlation measures, such as Spearman and Pearson correlation, can be used to measure similarity between gene (or gene set) activity in two comparisons, a natural measure is to utilize the model structure and calculate how probable it is to generate a query comparison given the model parameter fitted for a background comparison [55]. Using the model structure has the intuition that if two comparisons have similar expression patterns then the model fitted on one should be able to generate the other with high probability. Formally, this corresponds to the following expression:

$$\begin{aligned} \text{rel}(q, d) &= P(q|d, \text{collection}) \\ &= \int_{\Theta} P(\mathbf{x}_q|\Theta_q = \theta_d) P(\Theta|\mathbf{X}) d\Theta \end{aligned} \tag{4.3}$$

where $P(\Theta|\mathbf{X})$ is the posterior distribution of model parameters fitted on the background collection $\mathbf{X}$, $\mathbf{x}_q$ is the data of the query comparison, and $P(\mathbf{x}_q|\Theta_q = \theta_d)$ is the predictive likelihood of the query comparison given the parameters fitted on the background comparison. When Gibbs sampling is used to infer the posterior distribution of model parameters, the integral in the relevance measure is approximated by a mean over samples. In the classical LDA of Publication I, θ_d corresponds to the distribution of mixture components in the earlier comparison while in Publication III, it corresponds to the distribution of mixture modules in a background comparison. As the similarity measure in Equation (4.3) is not specific to a model family, it can be used to compute relevance based on any Bayesian latent variable model.

4.2.4 Model selection and evaluation

Model selection in topic models refers to estimating the optimal number of components. It becomes an important step especially in an exploratory analysis where a countless number of candidate models (with different number of components) could have produced the observed data. A common model selection approach is to have some measure of model performance that is used to compare the candidate models learned on different numbers of unknown latent components. Two model selection methods are used in Publication III. The first method compares the retrieval performance of the model under a varying numbers of total components. The second method estimates and compares the average predictive likelihood of unseen held-out data samples given some training data [66]. The next paragraph discusses two methods that are used to estimate the likelihood.

In classical LDA the predictive likelihood can be expressed as

$$P(X^{\text{test}}|X^{\text{train}}) = \int d\psi d\alpha P(X^{\text{test}}|\psi,\alpha)P(\psi,\alpha|X^{\text{train}}), \tag{4.4}$$

where ψ is the component-to-gene-set distribution and α is the prior for comparison-to-component distribution. The predictive likelihood is computationally feasible if one is willing to approximate the integral with a point estimate for ψ and α. In the study an MCMC sampler is used to marginalize out the component assignments associated to training data and to infer the point estimates ($\widehat{\psi}$ and $\widehat{\alpha}$). Considering each held-out comparison $\mathbf{x}_d$ independent, the first term in Equation 4.4 factorizes as follows:

$$P(X^{\text{test}}|\widehat{\psi},\widehat{\alpha}) = \prod_d P(\mathbf{x}_d|\widehat{\psi},\widehat{\alpha}). \tag{4.5}$$

This term can be interpreted as a normalization constant relating the posterior distribution of a component assignment to its joint distribution with the data by the Bayes rule:

$$P(\mathbf{z}|\mathbf{x}, \widehat{\psi}, \widehat{\alpha}) = \frac{P(\mathbf{z}, \mathbf{x}|\widehat{\psi}, \widehat{\alpha})}{P(\mathbf{x}|\widehat{\psi}, \widehat{\alpha})}, \tag{4.6}$$

where the current held-out comparison is represented as $\mathbf{x}$ and its latent components as $\mathbf{z}$. The subscript d is omitted because each held-out comparison can be evaluated separately, since the component assignments for one held-out comparison are independent of the component assignments for all other held-out comparisons. Several methods can be used to evaluate the normalization constant [68]. Publication III uses two alternative strategies; in the first an importance sampler is designed by setting the proposal distribution over the posterior of $\mathbf{z}$ in a way that yields a Harmonic mean estimator (HM [69]):

$$P(\mathbf{x}|\widehat{\psi}, \widehat{\alpha}) \approx \frac{1}{\frac{1}{S}\sum_s \frac{1}{P(\mathbf{x}|\mathbf{z}^s, \widehat{\psi})}}, \tag{4.7}$$

where $\mathbf{z}^{(s)} \sim P(\mathbf{z}|\mathbf{x}, \widehat{\psi}, \widehat{\alpha})$ and $\{\mathbf{z}^s\}_{s=1}^S$ are S samples taken from a Gibbs sampler after a burn-in period. The HM estimator is widely used due to its ease of implementation and relatively low computational costs. However, it has been criticized for its misleadingly low empirical variance [70]. An alternative better approach to estimate the normalization constant is via the Annealed importance sampler [68, 70]. It is a variant of simple importance sampling defined on a higher-dimensional state space where auxiliary variables are introduced in order to make the proposal distribution closer to the target distribution, so that

$$P_s(\mathbf{z}) \propto P(\mathbf{x}|\mathbf{z}, \psi)^{\tau_s} P(\mathbf{z}|\alpha). \tag{4.8}$$

The proposal distribution is built over an extended space $\mathbf{Z} = \{\mathbf{z}^{(s)}\}_{s=1}^S$ by first sampling from the tractable prior $P_0(\mathbf{z})$ and then, through a series of auxiliary variables, $0 < \tau_1, \tau_2, ..., \tau_S = 1$ moving the sample through intermediate distributions towards the posterior $P_S(\mathbf{z})$. In the analysis no major difference in the result was found between the model selection using the Annealed importance sampler and the Harmonic mean estimator.

4.2.5 Results

Both models are applied on data collected from the ArrayExpress repository. In Publication I around 800 comparisons are collected that corre-

spond to 288 different studies, while in Publication III an extended collection of 6925 comparisons is derived from 1082 different studies corresponding to three different species. The methods are evaluated based on their retrieval performance, qualitative assessment of modeled components and selected retrieval case studies. The retrieval performance evaluation is restricted to case vs control comparisons as they are easier to systematically assess.

In Publication I average precision is used to compare the model performance against a random base-line. The retrieval results show that in 20 out of 27 cancer vs normal comparisons the LDA model performed significantly better that the random base-line.

In Publication III the evaluation method uses a controlled vocabulary known as the Experimental Factor Ontology (EFO; [71]). EFO systematically characterizes the existing factor values and represents relationships between their values to describe biological conditions investigated in the ArrayExpress studies. A mapping for the non-control conditions to ontology terms is obtained for 219 interpretable comparisons, where the ground truth between two comparisons is based on the shared path between the corresponding terms in the EFO. Since this method yields a non-binary relevance a graded relevance measure, namely *Normalized Discounted Cumulative Grain* (NDCG) is used to measure the ranking quality. The NDCG measure quantifies how much the user gains for a query when a background comparison with a particular relevance is found at a particular rank in the ranked list of retrieval results. The retrieval results reveal comparable performance of the proposed method compared to the classical LDA and other existing methods.

Both studies evaluate the inferred components by interpreting the respective top gene sets. The inferred components model functionally coherent differential expression patterns and explain a wide range of biological processes, such as cell cycle, apoptosis, glycolysis, DNA replication and respiration etc. Several retrieval case studies also show that both models found meaningful existing and potentially new connections between different comparisons. In Publication III, the model suggested a previously unknown connection between Malignant Pleural Mesothelioma (*MPM*), which was a query comparison and Single-minded homolog 2, short isoform (*SIM2s*) transcription factor, which was the third most relevant result. This connection was followed up with a RT-PCR experiment on an independent set of mesothelioma samples. The experiment

validated the computationally predicted connection between *MPM* and *SIM2*, which leads to a hypothesis that *SIM2s* may have a role in *MPM* via the estrogen signaling network.

4.3 Retrieval of relevant samples given a set of genes

While Publications I and III compare expression profiles on a global scale, an important feature is to focus the search of earlier samples based on relevant regulatory relationships among user-defined genes-of-interest. There are three main steps needed to achieve this; first, a suitable model for reverse engineering a regulatory network among the genes must be constructed, second, since there exist several potentially useful reverse engineering models, a quantitative evaluation and understanding of the relative merits and shortcomings of the different models must be achieved, and third, a rigorous measure of relevance between a query sample and a background sample that utilizes the regulatory relationships among the user-defined genes must be created. In the following subsections the three main parts are described; the first two correspond to relevant contributions from Publication II while the third briefly summarizes a rigorous model-based similarity measure published in an earlier study [8].

4.3.1 Network reconstruction approaches

The challenge of identifying regulation networks from functional genomics data has resulted in development of a number of statistical and machine learning methods. In particular, Publication II compared four statistical methods for the recovery of network structure; 1. Graphical Gaussian Models (GGMs), 2. Linear regression with Least Absolute Shrinkage and Selection Operator (LASSO), 3. Sparse Bayesian Regression (SBR) and 4. Bayesian Networks (BN). In a regulatory or interaction network each node represents a variable of interest (e.g., gene) and edges among nodes represent strength of interaction.

Graphical Gaussian methods

Gaussian graphical models (GGMs) are undirected graphical networks that are used to infer conditional independences among a set of variables (or nodes of the network) under the assumption of a multivariate Gaussian distribution of the data. A GGM can be constructed by estimating

partial correlation coefficients among the set of variables. A partial correlation describes the correlation between two nodes conditional on all the other nodes in the network. From the theory of normal distribution it is related to the inverse of the covariance matrix [72], and therefore an important step in learning the model is the estimation of the covariance matrix and its inverse. In many cases of network reconstruction applications and specially in molecular biology, the number of observations is smaller than the number of variables and therefore the covariance matrix becomes singular. Publication II employs an existing shrinkage-based regularization approach that was found superior to other alternatives in [73]. The shrinkage approach replaces the empirical estimate of the covariance matrix by a weighted mixture of an empirical covariance estimate and a non-singular regularization matrix. The regularization matrix shrinks the off-diagonal entries to zero and leaves diagonal entries (variances) intact. The weight parameter is estimated analytically by minimizing the expected deviation of the inferred covariance matrix from the true covariance matrix (see [73] for details).

Linear regression

While partial correlation is one sensible approach to predict interactions, another alternative paradigm is the linear regression model. Classic linear regression takes as input multiple observations for both response $\mathbf{y}_q = \{y_1, y_2, ..., y_N\}$ and the predictor variables $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_R\}$, and predicts the value of a response variable with a weighted sum of predictor variables so that

$$\widehat{\mathbf{y}}_q = \sum_{r=1}^{R} w_{qr}\mathbf{x}_r, \tag{4.9}$$

where $\widehat{\mathbf{y}}_q$ is the predicted value of the response variable $\mathbf{y}_q$, and the regression parameters w_{qr} represent the strength of interaction between a predictor and the response variable[1]. The goal is to produce a weight vector $\mathbf{w}_q \in \mathcal{R}^R$ where the element r corresponds to the influence of a predictor variable. To obtain the weight vector typically the squared error between the the predictor and observed value of the response is minimized:

$$\widehat{\mathbf{w}}_q = \arg\min_{\mathbf{w}_q} \|\widehat{\mathbf{y}}_q - \mathbf{y}_q\|^2. \tag{4.10}$$

[1]Typically in non-regularized regression (Equation 4.9) a bias term (w_{q0}) is introduced by simply augmenting the data with a constant for the bias term, that is, $\mathbf{X} = \{1, \mathbf{x}_1, \mathbf{x}_2, ...\}$. In the regularized version, instead of the bias term, the data ($\mathbf{x}_r$ and $\mathbf{y}_q$) is standardized.

Minimizing the squared loss function corresponds to maximum likelihood estimation under a Gaussian model for the observations drawn independently from a normally distributed isotropic noise distribution [53] where the likelihood function becomes

$$p(\mathbf{y}_q \mid \mathbf{w}_q, \mathbf{X}) = \prod_{n=1}^{N} \mathcal{N}(y_n \mid \mathbf{w}_q^T \mathbf{x}_n, \sigma^2). \tag{4.11}$$

In practice this approach is usually susceptible to over-fitting, which calls for suitable regularization. The standard method of ridge regression penalizes the L2-norm of the weights, where the error function of Equation (4.10) is replaced by

$$\widehat{\mathbf{w}}_q = \arg\min_{\mathbf{w}_q} \left(\|\widehat{\mathbf{y}}_q - \mathbf{y}_q\|^2 + \lambda \sum_r w_{qr}^2 \right). \tag{4.12}$$

This can be interpreted as Bayesian MAP estimate under a zero-mean Gaussian prior on the weights with an isotropic covariance matrix. An alternative to L2 is the L1-norm that contains a stronger regularization term and yields more sparse results. It is commonly referred to as the *Least absolute shrinkage and selection operator* (LASSO; [74]). The LASSO cost function is

$$\widehat{\mathbf{w}}_q = \arg\min_{\mathbf{w}_q} \left(\|\widehat{\mathbf{y}}_q - \mathbf{y}_q\|^2 + \lambda \sum_r |w_{qr}| \right) \tag{4.13}$$

which can be interpreted as a Bayesian MAP estimate under a Laplacian prior on $\mathbf{w_q}$ [75].

Sparse Bayesian regression

The Sparse Bayesian regression model (SBR; [76]) is simply a Bayesian MAP estimation of Equation (4.10) under an Automatic Relevance Determination prior (ARD; [77]) on the interaction weights so that

$$p(\mathbf{w}_q \mid \lambda) = \prod_r \mathcal{N}\left(w_{qr} \mid 0, \lambda_r^{-1}\right). \tag{4.14}$$

In this prior the hyperparameters λ are optimized by maximizing the marginal likelihood. The reason for sparsity of the SBR approach is due to the hierarchical nature of the prior; each hyperparameter λ_r has an uninformative Gamma prior with shape and inverse scale parameters set to zero, which effectively leads to an improper prior. Integrating the hyperparameter out leads to a prior that is clearly sparse: $p(w_{gr}) \propto 1/w_{gr}$.

Bayesian Networks

A Bayesian network (BN) is a probabilistic graphical model that indicates how different random variables of interest interact. Each random variable

$\mathbf{x}_r$ is represented by a node r in the network. The model is defined by a directed acyclic graphical structure $\mathcal{H}$ where edges among the nodes are associated with conditional probabilities with parameters $\mathbf{q}$. If $\mathrm{pa}[r]$ defines parents of a node r and $\{\mathbf{x}\}_{\mathbf{pa}[r]}$ represents the set of random variables associated with $\mathrm{pa}[r]$, then due to the acyclicity, the joint distribution of all the random variables can be factorized into a product of lower-complexity conditional probabilities defined by the graphical structure:

$$p(\mathbf{x}_1, ...\mathbf{x}_R) = \prod_{r=1}^{R} p\left(\mathbf{x}_r \mid \{\mathbf{x}\}_{\mathbf{pa}[r]}\right). \tag{4.15}$$

The objective of Bayesian networks is to find a model structure $\mathcal{H}$ that best explains the data $\mathcal{D}$, that is, to sample structures from the posterior $p(\mathcal{D} \mid \mathcal{H})$. The posterior involves the marginal likelihood $p(\mathcal{D} \mid \mathcal{H})$ which averages the probability of data over all possible parameter assignments q,

$$p(\mathcal{D} \mid \mathcal{H}) = \int p(\mathcal{D}|\mathbf{q}, \mathcal{H}) p(\mathbf{q}|\mathcal{H}) d\mathbf{q}. \tag{4.16}$$

As the species data in the Publication II (discussed later) has been discretized, the marginal likelihood was computed under the assumption of a multinomial distribution with a Dirichlet prior, which results in a closed form solution for the likelihood [78]. Direct sampling from the posterior is analytically intractable and is therefore approximated by MCMC sampling [6, 79]. A structure MCMC method, proposed in [80], was used in the study; the method constructs a chain of network structures by starting with an initial graph and at each step either creates, deletes or inverts an edge. To constrain the search space for network structures a restriction of at most three parents was imposed. This method, commonly adapted in other studies [81], incorporates the prior knowledge that interaction networks are usually sparse. For detailed description of BNs see for instance [6, 78].

4.3.2 Evaluation of network reconstruction approaches

A crucial requirement to evaluate the different reconstruction methods is to test the ability of the models to recover the true network structure. The study uses simulated and real-world ecological datasets. The models are first evaluated on simulated food-webs where the true network structure is known precisely. The best performing methods from the simulated data evaluation are further tested to infer interactions among

39 bird species of European warblers. Since the true real-world network is not known for these species, the edges inferred from the methods are compared against those reported or expected from literature.

In order to simulate test datasets, first a set of 10 different network structures are simulated from a niche model that takes as input the number of species and the network density, then the abundance of the species is generated from a population model which takes as input the growth rate, species-specific demographic and environmental effect and effect of competition for common resources. The model generated 10 datasets. The input parameters are picked empirically so that they are close to the actual values in real food-webs and lead to stable simulation where population levels reach a steady state after a short burn-in phase.

Each network reconstruction method compared in the study leads to a matrix of scores associated with edges in a network. These scores are different in nature: partial correlation coefficients for GGMs, regression coefficients for LASSO and SBR, and marginal posterior probabilities for BNs. All three scores define a ranking of the edges. This ranking is used to define a receiver operator characteristic (ROC) curve by varying the threshold on the scores. The ROC curve is summarized as area under the curve (AUC), with a larger score indicating overall better performance, and as True-positive rate at 5% false-positive rate (TPFP5), which highlights performance at a low false-positive rate. The TPFP5 and AUC scores for simulated data indicated superior performance for the BN and the LASSO models. For the real-world data the performance of BN and LASSO was comparable.

4.3.3 Model-based similarity measure

To address the problem of querying an existing database of microarray measurements with a list of genes and to identify what experiments might be relevant based on differential activity of particular cellular processes, a predictive likelihood based Fisher similarity measure is adapted for retrieval in [8].

The study uses the LASSO regression model to learn a regulatory network among the genes in the query list. The interactions of each gene j in the list of target genes T are modeled as $\mathbf{x}_j = \mathbf{X}_{-j}\mathbf{w}_j + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is the noise term independent of any relevant biological condition, $\mathbf{x}_j = \{x_j^{(1)}, x_j^{(2)}, ..., x_j^{(n)}\}$ represents the observations for the interesting gene, $\mathbf{X}_{-j}$ contains the data for remaining genes, and $\mathbf{w}_j$ are the

corresponding regression coefficients. The model for the entire set of interesting genes is approximated by the so-called pseudo-likelihood [82]; the pseudo-likelihood regarding a data sample $\mathbf{x}^{(i)}$ is given by

$$p(\mathbf{x}_T^{(i)} \mid \mathbf{x}_{-T}^{(i)}, \mathbf{w}) = \prod_{j \in T} p(x_j^{(i)} \mid \mathbf{x}_{-j}^{(i)}, \mathbf{w_j}) \tag{4.17}$$

where $p(x_j^{(i)} \mid \mathbf{x}_{-j}^{(i)}, \mathbf{w_j}) = \mathcal{N}\left(x_j^{(i)} \mid \mathbf{w}_j^T \mathbf{x}_{-j}^{(i)}, \sigma^2\right)$.

Using the approximated model the study computes the Fisher score for each data point $s_{\hat{\mathbf{w}}}(\mathbf{x}^{(i)})$. The Fisher scores are defined by concatenating the partial derivatives of the log-likelihood with respect to the model parameters for a data sample [83]. The score indicates the direction in which to update the parameter estimates $\hat{\mathbf{w}}$ in order to maximize the log-likelihood of the sample starting from $\hat{\mathbf{w}}$. The inner product of the Fisher scores $K_{\hat{\mathbf{w}}}(\mathbf{x}^{(i_q)}, \mathbf{x}^{(i_k)})$, also known as the simple Fisher kernel, is then used to compute the relevance of each background sample $\mathbf{x}^{(i_k)}$ given an input query $\mathbf{x}^{(i_q)}$.

If $\mathcal{D}$ is the dataset from which the model is learned, then adding two new data points produces $\mathcal{D} + \mathbf{x}^{(i_q)}$ and $\mathcal{D} + \mathbf{x}^{(i_k)}$. The inner product in the Fisher kernel can be seen as updated parameters $\hat{\mathbf{w}}^{new\,\mathbf{x}^{(i_q)}}$ and $\hat{\mathbf{w}}^{new\,\mathbf{x}^{(i_k)}}$ that can be derived as a parameter update for the extended datasets by gradient ascent [8]. The score indicates the strength and sign of gene relationship. The study shows better retrieval performance of the simple Fisher kernel measure compared to other alternatives based on Euclidean and Pearson correlation.

4.4 Discussion

Publications I, II and III demonstrate that, given a query sample, even simple model-based probabilistic methods are able to retrieve biologically meaningful samples with a reasonably high accuracy. The proposed methods point out relationships between samples in the form of retrieval results and allow interpretation of retrieval results by modeling underlying biological processes. A careful study of the processes revealed many known results in Publications I and III and led to a novel biological finding for a Mesothelioma cancer in Publication III.

The proposed methods can be extended in several directions. For instance, an interesting direction is to consider nonparametric extensions of topic models that do not require to pre-specify the number of components. Additionally, it is important to adapt the search engine to other commonly

available background data repositories such as collections of background datasets where each dataset contains a set of samples and multiple collections of repositories where each repository contains measurements that are paired with the corresponding samples in the other repository. The next chapters describe solutions suitable for such scenarios.

5. Multi-view retrieval with biological samples as queries

Methods discussed in Chapter 4 were suitable for a single-view repository. This chapter presents the contributions to multi-view retrieval of gene expression data paired with another data type. It starts with a brief motivation, followed by a description of specific contributions, namely retrieval of relevant biological samples from paired genomic measurements (Publication IV) and retrieval of survival-associated genomic regions that are dependent among multiple sources of genomic measurements (Publication V).

5.1 Motivation and Related work

Cancers are complex diseases where cellular responses to a disease type or drug treatment are characterized by multivariate genome-wide changes at several layers of regulation [84, 85]. Therefore in cancer studies, it is becoming increasingly common to profile measurements from multiple genomic views where each view provides a complementary source of information to the underlying responses or mechanisms. For instance, to better characterize cellular responses to different cancers, the Cancer Genome Atlas (TCGA repository; [49]) provides not only gene expression measurements but also corresponding copy number variations, methylation data and micro-RNA measurements as multiple views of cellular responses to different cancer types. The rapid growth of such multi-view repositories requires new tools that are able to retrieve key variables and samples and increase our understanding of the underlying cellular processes.

Most related solutions are designed for single-view repositories, for instance, the *genome-wide association analysis* searches the genome for features with small variations that occur more frequently in people with a particular disease than in people without the disease [86]. Similarly, most

methods to search for relevant samples, including the ones discussed in Chapter 4, are specifically tailored for single-view repositories. The data integration approach adapted in Publication IV and V seek maximal dependence between two data sources. Other related integrative approaches that have been recently applied in functional genomics are kernel methods [87] that operate on similarity matrices and can model nonlinear feature spaces, asymmetric integration of one data source to support the analysis of another (primary) data source, and simultaneous non-negative matrix factorization [88].

5.2 Retrieval of relevant samples using paired measurements

The study in Publication IV investigates the following research question: *Given paired background data samples from a large repository, would it be possible to exploit their shared patterns to enhance the accuracy of retrieval of background samples that are most relevant to a given query?* The data repository contains biological and chemical effects of several drugs. The biological effects of the drugs are measured by gene expression measurements (referred to as the biological space) while the chemical properties are obtained as binary descriptors for each chemical (referred to as the chemical space). The following subsections summarize 1) the data representation for the paired measurements, 2) the computational model which captures the shared information between the biological and the chemical space, 3) the retrieval of relevant results based on a relevance measure that uses the shared information, and 4) evaluation of the results that allow system-level understanding of drug actions.

5.2.1 Representation of paired measurements

Biological responses of drugs are obtained from the Connectivity Map study [25]. For each drug, the study contains microarray measurements before and after drug treatment on three different cancer cell lines. A classical case-control design is used and differential expression is computed for each drug molecule. Since not each drug is used in every cell line, the cell line with the strongest effect is selected for each drug. The resulting data consisted of gene expression profiles of 1159 different drugs. Like in earlier studies, to bring in prior knowledge of biological processes and to reduce the dimensionality of the data, GSEA is performed for the cu-

rated 1321 gene sets from the C2 collection of MSigDB [33]. The chemical space is defined by Volsurf descriptors [89]. These descriptors are based on 3-D molecular fields and capture both structural similarities, such as shape, as well as general chemical features, such as hydrophobicity and lipophilicity properties. In total 76 different chemical descriptors were collected for each of the 1159 different drugs.

5.2.2 Data fusion using CCA model

The paired measurements for the drugs can be viewed as two background databases containing datasets $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$ and $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_N]$, where each pair ($\mathbf{x}_n \in \mathcal{R}^{D_1}, \mathbf{y}_n \in \mathcal{R}^{D_2}$) contains a two-view profile for a drug n. A regularized version of Canonical correlation analysis (CCA; [90]) is used to analyze the data. The CCA model decomposes the variation in each data source[1] into source-specific and shared components. It is an unsupervised latent variable model where the within-source variation is assumed irrelevant, sometimes called "noise", and only the shared effects are considered relevant. This is a sensible assumption in case of drug action mechanisms because the biological or chemical space considered alone consists of noisy measurements of drug functional similarity and the characteristics of the "noise" are not known.

While ordinary correlation characterizes the association strength between two paired scalar observations, CCA assumes paired vectorial values, and generalizes correlation to multidimensional sources (views). To capture the shared information between the two sets of sources, CCA searches for linear combination weights, here called basis vectors, for the two sources ($\mathbf{w}_s$ and $\mathbf{v}_s$; for the gene set activation values and for chemical descriptors, respectively) in a way that maximizes the correlation between the linear projections of the views onto these basis vectors, $\mathrm{cor}(\mathbf{X}\mathbf{w}_s, \mathbf{Y}\mathbf{v}_s)$. The variance of the projections is normalized which makes the magnitude of the *canonical correlation* τ_s bounded between $[-1, 1]$. The linear projections onto the basis vectors ($\mathbf{X}\mathbf{w}_s$ and $\mathbf{Y}\mathbf{v}_s$) are also called *canonical variates* or CCA components. Using the Pearson correlation as an estimator, the function to be maximized becomes:

$$\begin{aligned} \tau_s &= \underset{\mathbf{w}_s, \mathbf{v}_s}{\arg\max}\, \mathrm{cor}(\mathbf{X}\mathbf{w}_s, \mathbf{Y}\mathbf{v}_s) \\ &= \underset{\mathbf{w}_s, \mathbf{v}_s}{\arg\max}\, \frac{\mathbf{w}_s^T \mathbf{C}_{\mathbf{xy}} \mathbf{v}_s}{\sqrt{\mathbf{w}_s^T \mathbf{C}_{\mathbf{xx}} \mathbf{w}_s} \sqrt{\mathbf{v}_s^T \mathbf{C}_{\mathbf{yy}} \mathbf{v}_s}}, \end{aligned} \tag{5.1}$$

[1] Here sources mean the biological and chemical views.

where $\mathbf{C_{xx}}$ and $\mathbf{C_{yy}}$ are the within-sources covariance matrices of X and Y respectively and $\mathbf{C_{xy}}$ is the between-sources covariance matrix. Multiple pairs of basis vectors can be obtained iteratively. The first pair $(\mathbf{w}_1, \mathbf{v}_1)$ is optimized such that it has the largest canonical correlation (τ_1), the next pair $(\mathbf{w}_2, \mathbf{v}_2)$ is optimized to have the largest correlation with the constraint that it is uncorrelated with the previously found linear combination. Taking partial derivatives of Equation (5.1) with respect to basis vectors and normalizing them, the CCA optimization reduces to a generalized eigenvalue problem [91], where the analytical solution for the basis vectors is given by solving the eigenvalue equations

$$\begin{aligned} \mathbf{C_{xx}^{-1}C_{xy}C_{yy}^{-1}C_{yx}}\hat{\mathbf{w}}_s &= \tau_s^2\hat{\mathbf{w}}_s, \\ \mathbf{C_{yy}^{-1}C_{yx}C_{xx}^{-1}C_{xy}}\hat{\mathbf{v}}_s &= \tau_s^2\hat{\mathbf{v}}_s. \end{aligned} \tag{5.2}$$

The CCA solution has two useful properties: the result is invariant to linear transformations of the data, and the solution for any fixed number of components maximizes mutual information between the projections for Gaussian data [92].

Classical CCA cannot be applied directly to high-dimensional data settings due to unreliable inverses of the sample covariance matrices. This happens when the individual variables are highly correlated and the covariance matrices are ill-conditioned. The situation is commonly encountered with biological data. Two commonly used approaches to deal with both problems are 1) regularization [93] and 2) Bayesian modeling [94]. In the study a regularized approach is used, where the empirical covariance matrices $\mathbf{C_{xx}}$ and $\mathbf{C_{yy}}$ are replaced by regularized estimators defined by $\mathbf{C_{xx}} + \lambda_1\mathbf{I}$ and $\mathbf{C_{yy}} + \lambda_2\mathbf{I}$, respectively. The regularization parameters λ_1 and λ_2 are estimated in a cross validation fashion by maximizing the average retrieval performance.

5.2.3 Retrieval using CCA latent space

To test the performance of the CCA components in extracting functionally similar drugs and in combining potentially relevant statistical dependencies between the two views, the CCA-based retrieval performance is evaluated against the baseline retrieval using each single-view separately. The projection of the original data onto the CCA components is used to provide a low-dimensional vector representation for each drug molecule in each view. The projected views are concatenated for each drug and a simple pairwise Pearson's correlation is used to rank the different drugs given

a query drug.

5.2.4 Results

For quantifying the retrieval performance, each drug chemical is taken as a query and the average precision of retrieval results of most similar drugs is computed. The retrieval performance is compared against a gold standard that represents functional similarity of different chemicals based on their known protein targets and Anatomical Therapeutic Chemical (ATC) codes. As baseline comparison methods, results for single-view retrieval are evaluated for three different representations of the drugs: 1. gene expression, 2. expression of gene sets and 3. chemical descriptors. The retrieval results show that the combined space formed by the CCA components performed significantly better than any of the three spaces considered separately. Within single-view retrieval, the chemical descriptor space clearly performed better than retrieval based on the biological space (activities of genes or gene sets), indicating that the chemical space is more informative than this particular selection of genes for evaluating the functional similarity of the drug molecules. The retrieval performance using genes versus using gene sets was similar which indicates that the information loss in using gene sets, due to the smaller amount of features, is compensated by prior knowledge of which genes form biologically meaningful sets.

Next the CCA components are used to analyze complex relationships between chemical structure of drug molecules and their genome-wide responses in the cells. A detailed interpretation for the top ten most correlated components led to several sensible and potentially useful hypotheses of drug response mechanisms. For instance, three subcomponents shared the same or similar chemotherapeutic and DNA damaging drugs, while their top chemical and biological characteristics revealed two different DNA damage response mechanisms, namely mitotic arrest response due to hydrophobic and size related features and a reparative response driven by hydrogen bonding and hydrophilic features (details in the study section “Components 2B & 10A - functionally similar but gene-wise different responses” and section “Components 3/3A - A cell stress component”, respectively).

5.3 Survival analysis for multi-view components

In recent decades cancer genomics has focused on the discovery of genetic mutations and chromosomal changes associated to a cancer phenotype. Though a single mutation may relate to a particular phenotype, it is the combination of many different molecular mechanisms that disrupt cellular pathways and characterize cancer [85]. Given multi-view samples, Publication V presents an effective pipeline that addresses the problem of retrieving multi-view regions of the genome that effectively stratify data samples (patients) into low and high survival groups. The following sections summarize key steps of the pipeline: first relevant clinical and patient specific covariates are collected from a multi-view repository (Section 5.3.1); next, a Bayesian variant is used to model multi-view regions in order to be able to incorporate suitable priors for constraining dependencies between multi-view sources and avoid the over-fitting problem of the classical CCA (Section 5.3.2), after this the regions are used to stratify patients into groups and a survival association analysis is performed to identify potentially interesting regions (Section 5.3.3), and lastly the interesting regions are evaluated against existing literature (Section 5.3.3).

5.3.1 Representation of paired data samples

As a multi-view dataset three data types; namely, gene expression, DNA copy number changes, and methylation pre-treatment measurements were collected for the available 250 Glioblastoma Multiforme (GBM) subjects from the TCGA repository in Publication V. GBM is one of the most aggressive malignant brain tumors where affected patients have a uniformly poor prognosis with a median survival time of only 15 months over the past 25 years [95]. These tumors are now well characterized at genome and transcriptome levels and several studies have demonstrated that the combination of these two molecular levels may be advantageous to characterize robust signatures that are clinically relevant for GBM [96, 97].

In Publication V a chromosomal continuous data source (either copy number changes or methylation) is considered as one view and gene expression is taken as the second. This results in two studies: a) search for dependencies between copy number and gene expression, and b) search for dependencies between methylation and gene expression. The probes for each dataset were matched resulting in 3480 genes for the gene expression copy number pair and 2530 genes for the gene expression methyla-

tion pair. In addition to the molecular profiles, subject-specific clinical information such as age, gender and race was also collected from the TCGA database.

5.3.2 Data fusion using similarity-constrained CCA

Prior biological knowledge can help in modeling potential dependencies between chromosomal gains or losses and gene expression of the associated genes. Copy number or methylation changes in a particular chromosomal region are captured by multiple probes, and this is also visible in the expression of the corresponding genes in the affected region. The copy number gain and loss are likely to be positively correlated with the expression levels of the affected genes, and the gain of methylation is likely to be negatively correlated to gene expression.

A recently developed constrained version of Bayesian CCA [98] is adapted in the study to encode this prior knowledge by enforcing constraints on the projection vectors $\mathbf{w}_s$ and $\mathbf{v}_s$. The constrained Bayesian CCA model couples the projection vectors with a transformation matrix $\mathbf{T}$: $\mathbf{v}_s = \mathbf{T}\mathbf{w}_s$. The Bayesian CCA provides a flexible approach to incorporate suitable constraints on the projection matrices and deal with the uncertainty in the data and model parameters. In the Bayesian formulation, the two data sources are assumed to be generated by a shared Gaussian latent variable $\mathbf{z}$ and a normally distributed dataset-specific noise with zero mean and covariance $\mathbf{\Psi}$. The model is formally defined as

$$\begin{aligned} \mathbf{x}_n &\sim \mathcal{N}(\mathbf{W}\mathbf{z}_n, \mathbf{\Psi}_x) \\ \mathbf{y}_n &\sim \mathcal{N}(\mathbf{V}\mathbf{z}_n, \mathbf{\Psi}_y), \end{aligned} \tag{5.3}$$

where the individual samples $\mathbf{x}_n$ and $\mathbf{y}_n$ are assumed to stem from a shared latent variable $\mathbf{z}_n \in \mathcal{R}^{R\times 1}$ and view-specific effects. The manifestation of $\mathbf{z}_n$ in each data source can be different and is parameterized by the projection matrices $\mathbf{W} \in \mathcal{R}^{D_1\times R}$ and $\mathbf{V} \in \mathcal{R}^{D_2\times R}$. Assuming a standard Gaussian model for the latent variable, $\mathbf{z}_n \sim \mathcal{N}(0, \mathbf{I})$, the correlation maximization projections of the classical CCA can be retrieved from the ML solution of the model [99, 100]. The model likelihood is given by

$$p(\mathbf{X}, \mathbf{Y}, \mathbf{V}, \mathbf{W}, \mathbf{\Psi}) = \int p(\mathbf{X}, \mathbf{Y}|\mathbf{V}, \mathbf{W}, \mathbf{\Psi}) p(\mathbf{V}|\mathbf{W}) p(\mathbf{W}) p(\mathbf{\Psi}), \tag{5.4}$$

where $\mathbf{\Psi}$ is a block diagonal matrix consisting of $\mathbf{\Psi}_x$ and $\mathbf{\Psi}_y$ as the blocks. The conditional probability $p(\mathbf{V} \mid \mathbf{W})$ encodes the relationship between

the transformation matrices for the shared latent variable. It is reparameterized with a transformation matrix $\mathbf{T}$ such that $\mathbf{V} = \mathbf{TW}$; assuming $\mathbf{W}^T\mathbf{W}$ is invertible $\mathbf{T} = \mathbf{V}(\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T$.

Publication V uses the truncated normal distribution $p(\mathbf{T}) = \mathcal{N}_+(\|(\mathbf{T}-\mathbf{I})\| \mid 0, \sigma^2\mathbf{I})$ as a prior on $\mathbf{T}$. The prior is used to make the model focus on searching for dependencies which combine the signal across adjacent genes within a particular chromosomal region. It can be plugged into $p(\mathbf{V}|\mathbf{W})$. There are two extremes for the prior; in the unconstrained form $\sigma \to \infty$, the model reduces to the traditional CCA, while setting $\sigma \to 0$ yields identical shared components derived from both data sources. The variance parameter can be used to tune the trade-off between the two extremes. The prior favors positive correlations between two data sources which is sensible for paired samples of gene expression and copy number changes [98, 101]. For the case of the gene expression and methylation pair, the relationship is inverse because down-regulation of a gene can be due to hyper-methylation and similarly up-regulation of a gene can be due to hypo-methylation. In Publication V the inverse relationship is encoded via the prior $p(\mathbf{T}) = \mathcal{N}_+(\|(\mathbf{T}+\mathbf{I})\| \mid 0, \sigma^2\mathbf{I})$. Uninformative priors are assumed for the model parameters $\mathbf{W}$ and $\mathbf{\Psi}$, and these were estimated using an EM algorithm.

Following [98], to find dependent regions a chromosomal region is defined via a window that is centered at a gene and spans across ten neighboring genes within the chromosomal arm. The window is slid across all chromosomal arms and a dependency score and each sample's contribution towards the score for each region is calculated. The dependency score is computed as a ratio of the strength of the shared signal versus the marginal effects, computed as $\mathrm{Tr}(\mathbf{WW}^T)/\mathrm{Tr}(\mathbf{\Psi})$, where Tr denotes matrix trace. A high score reveals a correlating expression and corresponding chromosomal change; high-scoring regions with q-value < 0.05 are selected for further analysis. Only one-dimensional latent variables are considered in the study. For each significantly dependent region, sample-wise contribution scores, as manifested in the latent variable z_n, are ordered and three groups are formed based on the 10th percentile, the 90th percentile and the rest. The same analysis is repeated for the pair of the gene expression and methylation datasets using the inverse prior.

5.3.3 Survival association analysis

In order to quantify the survival association of the significantly dependent regions, two patient groups are formed for each region, based on extreme values of the dependency score, and their survival curves are compared to check for any significant difference.

The TCGA survival time data are *right-censored* which implies that the survival age of subjects is partially known. This can happen if some subjects are alive and others withdraw or their information is lost during monitoring before the final outcome is observed. In this context the final outcome variable is when the patient expires. In the study the survival functions of different groups are compared by basic survival association analysis techniques [102]. The survival association methods estimate the outcome variable of interest, namely the time until an event occurs. The analysis captures the probability that a system will survive beyond a specified time. A system in this context is the group of subjects that contribute most to a significantly dependent region. There are two main components in a survival analysis: estimation of the survival function given censored data and comparison of the functions for multiple groups. The survival function $S(t)$ is the probability that an individual survives longer than time t. In the study the classical Kaplan-Meier (KM; [102, 103]) estimator for the survival function is used:

$$\hat{S}(t_{j-1}) = \hat{S}(t_j)p(T > t_j | T \geq t_j), \tag{5.5}$$

where T is a random variable denoting time of death. Equation 5.5 evaluates the probability of surviving past the previous event time $t(j-1)$, multiplied by the conditional probability of surviving past the current time $t(j)$, given survival to at least time $t(j)$. The estimator allows to draw KM survival curves for each group. A standard log-rank test is used to compute significance for the differences [102].

As the KM analysis does not model the effect of covariates, the significance levels can be biased due to any external confounder. In order to check for the bias a Fisher contingency table analysis is performed where one of the groupings is induced by a quantile clustering on the sample-wise contribution scores from the model and the second grouping is formed from any of the three external clinical factors considered separately. Two clinical factors, race and gender, are transformed into binary variables while the third variable, age, is discretized to four values.

5.3.4 Results - survival associated dependent regions

The dependency analysis of Publication V resulted in 281 significantly dependent regions between the gene expression and copy number datasets and 313 regions between the gene expression and methylation datasets. The histograms for the patient contribution scores to the dependent regions followed a bell shape centered at zero, with a few patients that contribute most to the dependency score. These scores are used to stratify the patients into groups and to compare the corresponding KM survival curves. Using a strict cut-off $(q < 0.05)$ on the survival test score, the gene expression and copy number datasets identified three chromosomal regions in the chromosome 10 that is known to be closely related to the GBM [104, 105]. Similarly, the methylation and gene expression dataset revealed a single region that is recognized to have both tumor suppressive and promoting properties depending on different tumor types [106, 107]. Overall the pipeline found biologically sensible regions from a multi-view repository that were predictive of patient survival.

5.4 Discussion

In this chapter both studies used variants of canonical correlation analysis, a subspace learning algorithm where the input views are assumed to have been generated from a shared latent subspace. Additionally, the views may contain an independent but unknown type of noise. The first study (Publication IV) investigated the potential of exploiting latent components modeled by a simple multi-view model in improving the performance of correctly identifying relevant samples given a query sample, while the second (Publication V) utilized a constrained version of the method that helps analyze and extract survival-associated relevant multi-view features by searching for biologically sensible dependencies between the features of the data.

The two studies identify relevant data by modeling hidden relationships among data features with a flexible multi-view model. The results from Publication IV indicate that integrated analysis of both the chemical and biological dataset is more informative than either dataset considered alone in predicting drug similarities as measured by comparison to ground-truth. The qualitative results allow system-level understanding of drug actions, which is of extreme importance given their complexity. The

complexity stems not only because the treatment drugs can often bind to and interact with multiple targets, but also from the fact that the diversity of biological responses to diseases at the cellular level is immense. The case study on GBM, in Publication V, reveals that the constrained version of the CCA model indeed finds multi-view regions that are known to be predictive of patient survival.

The proposed multi-view method searches for dependencies between different functional layers at the transcriptome and genome levels, which makes it possible to discover mechanisms and interactions that are not seen in the individual measurement sources. The assumption of CCA, that shared variation is interesting, is also useful in other multi-platform measurements that are rapidly becoming common in cancer studies. The results highlight the need for advanced algorithms to identify genomic regions or transcript profiles that play a key role in cancer progression and drug resistance.

6. Multi-task learning and retrieval of datasets

The methods discussed in the earlier chapters were suitable for modeling and retrieval of relevant samples. This chapter presents the contributions to two new models that extend the approaches described in the earlier chapters to model and relate a collection of datasets. A summarized review of existing work and related motivation is presented in Section 6.1. The first proposed model is a nonparametric multi-task method discussed in Section 6.2 while the second approach is a scalable and rapidly computable model-based dataset retrieval engine discussed in Section 6.3. The two models correspond to Publications VI and VII, respectively.

6.1 Motivation and Related work

A typical setting in molecular biology is to assay several variables (p) with small sample sizes (n). For instance, high-throughput technologies measure many genes for a single sample, rather than many samples for a single variable. Therefore, a key challenge in current studies is how to make trustworthy models based on few samples when the number of studied variables is large [108, 109].

A commonly used solution to the "large p small n" problem is to combine statistical evidence across related datasets [110, 111]. *Multi-task learning* provides a suitable class of approaches for such solutions [112]. In multi-task learning several estimation tasks (or datasets) are pursued together assuming properties which can be shared across datasets. The objective of multi-task learning is to boost performance of a new task by transferring domain knowledge from previously observed tasks or to improve learning performance of each individual task. A related method is *meta-analysis*, where several related studies are combined to enhance statistical power in order to obtain more accurate inference on target vari-

ables [113]. The meta-analysis methods are not simple, as the user who wants to find datasets that are combinable with her own data must resort to searches in free text or in controlled vocabularies that require much downstream data curation [1].

Other works that try to relate datasets typically utilize pairwise similarities between datasets, where the simplest method are based on correlation between vectors that represent datasets. For example, a recent work [114] uses within-dataset gene-gene pairwise correlation. This representation is not ideal for relating datasets as it requires a large number of samples to sensibly estimate gene correlation matrices and furthermore makes the dataset representation bulkier than the original data. Other existing alternatives require specific case-control designs [115], expert curated training data [116] or carefully chosen keywords [4].

The next two sections summarize a multi-task model and a more rapidly computable "combination model" to relate a new dataset to background datasets from earlier studies.

6.2 Multi-task topic model for transfer learning

Learning a model for a single dataset can be called a *task*. To gain more information about a dataset of user-interest (also called the task-of-interest), *transfer learning* methods transfer knowledge from earlier tasks to a new one, and *multi-task learning* methods learn several tasks together from their respective datasets utilizing their shared relationships. For example, the data of these related tasks may be genomic datasets from microarray repositories, or textual articles from other tracks in a conference. The work in Publication VI proposes a new multi-task model that implicitly represents similarity across datasets by sparse sharing of latent topics. The model improves performance on modeling topics underlying the dataset-of-interest by transferring domain knowledge from previously observed datasets. It extends the single-task LDA topic model of Publication I in three ways:

1. It builds a hierarchical model which is able to model a collection of datasets.

2. The model does not require one to pre-specify the number of available topics.

3. Unlike other multi-task topic models, it decouples topic sharing from topic strength, which makes sharing of low-strength topics easier.

These extensions are described in the following subsections.

6.2.1 Hierarchical Multi-task topic model

Probabilistic topic models are suitable for inferring latent components from count data, such as texts in a bag-of-words representation where individual topics represent co-occurrences of words [66, 67] or gene set activation counts from microarray samples where topics represent biologically meaningful co-activations between patterns of differential expression (as described in Section 4.2.2).

To model sharing of information among multiple tasks, the topics are allowed to be shared across the tasks by extending the hierarchical structure of the classical single-task LDA model. The generative process for the counts (of words within documents, or gene sets within microarray samples) is similar to LDA except for the generation of the document-level topic distribution. In the multi-task model this distribution is made specific to each dataset by adding an additional layer of task-level parameters that allow the model to sample the document-level topic distribution from the task-level parameters. The task-level parameters specify the task-level distribution over topics. To model sharing of topics across different tasks these task-level parameters are further sampled from a shared set of hyperparameters that control the overall strength and prevalence of topics across the entire set of datasets. A plate diagram of an existing multi-task LDA model is shown in Figure 6.1 and briefly discussed in the following section. Standard statistical techniques can be used to infer the set of topics that are responsible for generating a collection of documents.

6.2.2 Nonparametric priors

Nonparametric Bayesian variants of the LDA model are appealing since they allow easy inference of how many topics are active in a collection, instead of specifying a prior upper-limit. In earlier work, Teh et al., (2006) have proposed a Hierarchical Dirichlet allocation (HDPLDA; [117]) model where a Dirichlet process prior is used for topics. In the same study the authors further extend the single-task HDPLDA model to multiple groups of documents which is here denoted as MT-HDPLDA.

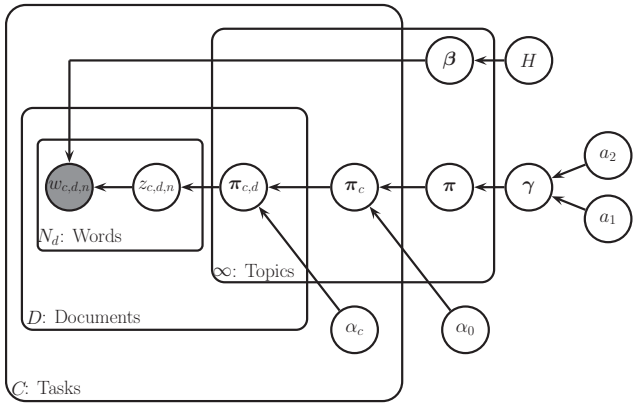


Figure 6.1. Plate diagram for multi-task LDA model; a nonparametric topic model for datasets (MT-HDPLDA). Rectangles indicate sets of variables, with cardinality of the set marked in the bottom left corner. Gray nodes correspond to observed data, count of a unique word in a document. In each dataset (task) the overall topic distribution is controlled by a dataset-specific Dirichlet process with task-level parameters π_c and α_c (see text for description of other variables). Figure adapted from Publication VI.

In the MT-HDPLDA model (Figure 4.3), topics for a document ($z_{c,d,n}$) are drawn from a Dirichlet process (DP with parameters $\pi_{c,d}$ and α_c), which in turn is drawn from a dataset-level DP (with parameters π_c and α_0), which can in turn be drawn from an overall DP across datasets (with parameters π and base measure H). The topmost DP in the hierarchy determines which topics are active overall and their strengths; lower-level DPs choose among their parent-level active topics, varying their strengths by a stick-breaking construction to yield differing topic distributions at each branch of the hierarchy. When inferring topics from data, the topmost DP can activate new topics as well as change their strength; the Hierarchical DP can thus infer the number of topics from data. Since sharing is done by the topic strength hierarchy, with the stick-breaking construction the strongest topics (which generate many words overall) are most likely to survive in several branches of the hierarchy and thus be shared across datasets; this can make the model a bad fit for multi-task problems with low-strength shared topics (topics discussed in many document collections but not at great length).

Nonparametric prior for low-strength shared topics

Unlike MT-HDPLDA model, the proposed multi-task model in Publication VI uses an Indian Buffet Process (IBP; [118]) based spike-and-slab prior that controls the sharing of topics across and within data collections. The IBP is a nonparametric prior over binary matrices that allows

potentially infinite number of active topics or features K. The distribution of IBP is sampled from a *stick-breaking construction;* [119]. Metaphorically the stick-breaking construction can be understood as follows: start off with a stick of length 1, then at each iteration $k = 1, 2, ...$ break off a piece at a point $v^{(k)}$ relative to the current length of the stick $\pi^{(k-1)}$, then record the length $\pi^{(k)}$ of the stick that was just broken off and recursively repeat the process on the broken-off piece. The construction produces a decreasing sequence of latent probabilities $\pi^{(k)}$ that can be used as a prior over unbounded binary matrices having a finite number of rows and an infinite number of columns, with a finite number of 1-valued elements in each row.

In the proposed model, the IBP controls presence of topics across the different tasks where rows of the IBP matrix represent different tasks and columns represent topics. To draw a topic for a new task, the IBP chooses one of the existing topics according to how many tasks they are already present in, or activates a new topic, hence the number of active topics is inferred from data. The IBP by itself did not provide enough sparsity as its parameters are learned from few observations (one row per task of a binary matrix); the proposed model therefore contains an additional sparsity masking step that turns off some components in each task. The strengths of the remaining active topics are drawn from a Gamma distribution within each task; from this task-specific topic prior, the remaining generation proceeds as in the classical LDA, drawing document-specific topic distributions and then the words for each document. The combination of the Gamma-distributed topic strengths and the IBP can be seen as an infinite spike-and-slab prior, where the IBP and the additional masking generates the spikes (probability of non-active topics) and the Gamma distribution acts as a slab (strength for active topics). The use of the independent topic strength variable avoids the restriction imposed by the DP construction which makes it easier to model weak shared topics in a data collection.

To infer the model from observations it is possible to directly integrate out the nuisance model parameters, the posterior for the rest of the variables is then sampled using a combination of collapsed Gibbs sampling and the Metropolis-Hastings algorithm.

6.2.3 Comparative performance evaluation

The model performance is compared on simulated and real world textual count data that is a standard type of data in topic modeling research. In this setting each dataset in a collection represents a learning task and contains a group of documents, each document contains a collection of word occurrences that are assumed to arise from the underlying different topics discussed in the document. Documents that belong to the same dataset tend to share topics more than documents that belong to other datasets. The model's ability to perform transfer learning is compared against the state-of-the-art MT-HDPLDA model. For this the standard predictive log-likelihood on held-out test documents from the interesting task (dataset) is used as a measure to benchmark the model's performance.

Two related simulated experiments are designed. The first experiment considers a continuum of problem domains where each represents a different multi-task learning problem; intermediate continuum points contain weak shared topics where the proposed model performs better while the extreme points on the continuum contains strong shared topics where the alternative MT-HDPLDA is superior. The second set of experiments evaluates the model performance under a varying number of total tasks by considering the different domains in the continuum. The proposed model performs better at continuum points where weak shared topics are likely to exist, and there are not very many tasks to learn the models from. Both models increase their performance as more tasks become available and this behavior is consistent even in the domain where all topics are relatively strongly present in their respective tasks.

To illustrate the topic model that the proposed method learns, a collection of NIPS conference articles is considered; in total the collection contains more than a thousand documents that are grouped into five different groups; e.g. Neuroscience, Algorithms, Learning theory etc. Top words from the strongest two topics for each task represent the topics that are expected to be discussed in the corresponding documents of the group. A comparative performance evaluation based on predictive likelihood for NIPS collection and another 20-Newsgroup data collection reveals that the proposed model is superior when the number of available documents for the task-of-interest is low; a case where transfer learning

is most wanted.

6.2.4 Discussion

The Bayesian multi-task models provide principled approaches to model datasets. The proposed nonparametric multi-task method models the topics in a dataset-of-interest by transferring knowledge from previously observed topic modeling tasks from earlier datasets. The method extends the classical LDA with a nonparametric prior that inherently performs model selection to explore different numbers of topics. The benefit of the multi-task method is that the modeled latent components are shared across datasets. These latent components can be analyzed (e.g. along the lines of Publication III or IV) to interpret relevance between two datasets. Evaluation on a well-annotated collection of microarray experiments provides a natural extension of the study.

6.3 Efficient combination of models

Multi-task learning is a form of global analysis that builds a single unified model of all the data. As the number of datasets keep increasing and the amount of quantitative biological knowledge keeps accumulating, the complexity of the task of building an accurate unified model becomes increasingly prohibitive [120]. However, since the "large p small n" problem anyway requires taking properly into account both the uncertainty in the data and the existing biological knowledge, it makes sense to assume that in the future researchers will increasingly develop their hypotheses in terms of (probabilistic) models of their own data. Publication VII presents a feasibility study for the future scenario where a large number of experiments are modeled beforehand and the models are stored in public repositories analogously to how their data are currently stored in public repositories. The following subsections describe 1. the proposed *combination model* that is used as a dataset retrieval engine, 2. representation of the background datasets in terms of their respective base models and lastly 3. a quantitative and qualitative evaluation of the model using an annotated collection of experiments as a case study.

6.3.1 Combination model

The model proposed in Publication VII is a probabilistic mixture of models that assumes that the biological activity in the query dataset can be approximately explained by a model of potential active biological effects represented as a combination of background models of the earlier datasets. Each earlier dataset s_j is represented by a background model M^{s_j} (also referred to as a *base model*) that can be any model selected by the author of the earlier dataset as long as it allows to compute the predictive likelihood of the query dataset, denoted as $P(\mathbf{X}|M^{s_j})$.

The combination model is defined for the samples $\mathbf{x}_i^q$ in the query dataset as a simple mixture of base distributions $P(\mathbf{x}_i^q|M^{s_j})$, where each distribution (for each dataset s_j) is associated with a mixture proportion or weight θ_j. The resulting predictive likelihood, given by the whole combination model to a query dataset, becomes

$$P(\{\mathbf{x}_i^q\}_{i=1}^{N_q}; \boldsymbol{\Theta}^q) = \prod_{i=1}^{N_q} \Big[\Big(\sum_{j=1}^{N_S} \theta_j^q P(\mathbf{x}_i^q|M^{s_j}) \Big) + \theta_{N_S+1}^q P(\mathbf{x}_i^q|\boldsymbol{\Psi}) \Big] \tag{6.1}$$

where N_q is the total number of samples in the query dataset and $\boldsymbol{\Psi}$ is a noise model. The mixture weights are constrained to be non-negative and sum to one, that is, $\sum_{j=1}^{N_S+1} \theta_j^q = 1$. Base models that explain a large proportion of the query (i.e., ones with large θ_j) are ranked higher in the retrieval results and the mixture weight values are directly used as a proxy for similarity of background datasets to the query.

The estimation of combination weights turns out to be a constrained concave optimization problem for which both projected gradient optimization and the Frank-Wolfe algorithm [121] can be used. Constraints in the former can be imposed after each gradient update where the resulting weight vector are projected onto the canonical simplex C using an efficient algorithm that minimizes the squared Euclidean distance between the new point $\boldsymbol{\Theta}^q$ and the original point $\boldsymbol{\Theta}_0^q$ [122]. Since the resulting cost function is strictly concave and globally smooth the optimization enjoys fast convergence and the computation time remains linear in the number of background datasets.

6.3.2 Dataset representation as a base model

The base models are assumed to be probabilistic generative models that are able to capture both prior and data-driven knowledge deemed necessary by the author of the data. In particular, the study considers two

model types (topic models and mixtures of unigrams) that have been used in earlier studies to model gene expression data [123–125]. Both model variants use counts obtained from the gene set enrichment analysis.

The unsupervised LDA [66, 67] or mixture of unigrams [126] is considered as a base model (which-ever models the data better) and is trained for every dataset. Effectively, each dataset is represented by a probability distribution over components (also called topics) which are shared across all samples but with a different degree of activation in each. In LDA each sample may be produced by multiple topics while in mixture of unigrams each sample is assumed to stem from a single component. Standard inference techniques, Gibbs sampling for LDA [66, 127] and EM for mixture of unigrams [126] are used to estimate the model parameters (per sample distribution over topics and per topic distribution over gene sets) and hyperparameters that control the prior probability of each topic. Given a new query sample the predictive likelihood is computed using the base model which was learned on the background dataset using the empirical likelihood based scheme [128].

6.3.3 Model performance and the inferred network

Quantitative Performance

The combination model in Publication VII is evaluated on a large annotated collection of microarray experiments that is a subset of the ArrayExpress repository [40]. The collection contains 206 datasets that in total have 5372 microarray samples which are consistently annotated with a tissue type and disease name. All datasets are modeled by either the Latent Dirichlet allocation or the mixture of unigrams model. The retrieval performance of the combination model is compared against the retrieval based on the keywords, by evaluating the retrieval results for each query dataset in the collection. The quality of retrieved results is measured by the standard precision-recall curve which reveals good and consistently better performance of the proposed model compared to the keyword-based search.

Relationships among datasets

In the combination model, each query dataset is represented as a combination of earlier datasets, encoded as a weight vector whose dimen-

sionality equals the number of datasets in the repository. A single non-zero weight value represents an edge between a query and a background dataset and is used as a proxy for relevance of the background dataset to the query. In order to interpret and visualize these relationships, a non-linear projection scheme is used that preserves the inter-point distance between any two datasets in the original space. In particular, *Markov clustering* [129] is used over the matrix of combination weights between all query datasets and their respective earlier background datasets, followed by non-linear projection with a variant of weighted Multi-Dimensional Scaling [130].

The clustering is mainly explained by tissue types, where three main clusters dominate: 1. solid normal and neoplastic tissue, 2. cell lines and 3. hematopoietic tissue (the clusters are visualized in Figure 2 of the study). Fine grained structures within clusters are biologically sensible; the solid normal tissue cluster forms a subnetwork of closely connected skeletal and heart muscle datasets, the hematopoietic cluster contains a small sub-cluster of Myeloma and Leukemia that is separated from the Mononuclear cells, and the cell line cluster mostly contains non-cancerous cell lines while cancer cell lines are placed as outliers mostly connected with the disease that they profile.

Comparison to citation information

The data-driven network produced by the combination model is compared with citation data. For that, several statistics about each dataset's respective publication are extracted from Pubmed and Web of Science, such as direct and indirect citations, impact factor of publication venue, total number of citations, and h-index of the last author. Interestingly, two of the datasets found to have high data-driven citations (i.e. having high out-degrees as found by the combination model) were associated to inconsistent publication entries in the public repositories; a scatter plot of the number of citations against normalized weighted out-degrees revealed an extreme off-diagonal position for these datasets, which upon inspection led to the inconsistent publication entry. The corrected publication information increases the correlation between the citation counts and the data-driven out-degree measure. A systematic enrichment test over extreme values of the weighted out-degrees and the corresponding citations indicates a significant bias for citations towards high impact factor publication venues and h-index of the last author. Systematic analysis of densely

connected datasets (cliques of experiments) reveals a breast cancer and a leukocyte clique that are shared between the data-driven network and the citation network. There are also cliques that are only visible in the data-driven network; these sets reveal biologically meaningful relationships, some of which are not easily visible from annotations, for instance, strong connections among cells in a T-cell related clique that capture different developmental stages such as Thymocytes and T-cells.

6.3.4 Discussion

The proposed *combination model* in Publication VII is a novel general-purpose model that is both scalable and rapidly computable. It is able to decompose a given query set into effects explained by earlier datasets. The method can model both multi-view and single-view data collections as long as suitable models exist for the background datasets. Evaluation on a larger data collection and a corresponding comparison of the citation patterns and the data-driven network provide an interesting future direction of the study.

7. Summary and conclusions

In this thesis new methods are proposed that utilize existing knowledge in the form of measurement data taken from earlier studies. This form of prior knowledge is quite complex, partly relevant, heterogeneous and noisy given that laboratories around the globe have different procedures to take patient samples. Specifically, the research problems can be abstracted as *what can be done with the available large repositories towards cumulatively building knowledge from data in molecular biology*. The thesis arrives at an answer: a *modeling-driven data retrieval engine*, which researchers can use to position their measurement data into the context of earlier biology. It is argued that using the available background information from hundreds of different situations or conditions, it is potentially feasible to both complement the existing scarce data and to focus the analysis on relevant variables.

The thesis considers three different scenarios for the background biological measurements and proposes novel retrieval engines for each:

1. a collection of earlier samples, where each sample is a microarray measurement (Chapter 4).

2. a collection of paired samples, where each sample is represented by more than one data type, for instance, a paired profile of copy number changes and an associated gene expression measurement (Chapter 5).

3. a collection of datasets, where each dataset contains multiple samples and corresponds to an experiment (Chapter 6).

The thesis considers two cases for each; for scenario 1 it considers retrieval of relevant results at both the global genomic scale and at a local

scale focused on relationships among a set of user specified genes; for scenario 2 it considers retrieval of relevant samples and retrieval of key survival-associated regions from multi-view profiles by modeling hidden relationships among data features; and for scenario 3 it considers a non-parametric method for relating datasets that have weak hidden processes and a more scalable and rapidly computable retrieval model for datasets when earlier datasets are already modeled with a probabilistic generative model that allows computing predictive likelihoods.

The main findings from Chapter 4 are a) the model-based retrieval of transcriptomic samples at a global scale that is able to find biologically meaningful relevant samples given a query sample and b) an evaluation of several potential underlying network reconstruction models that can be used to focus the search of earlier samples based on relevant regulatory relationships among user-defined genes-of-interest. The latter comparative evaluation of the reconstruction approaches finds superior performance for the LASSO regression and the Bayesian network models on both simulated and real-world data collections.

The results from Chapter 5 indicate that the added benefit of modeling shared patterns from a paired data source increases the retrieval performance of identifying relevant drug profiles. In particular, the underlying canonical correlation analysis was able to extend our understanding of drug action mechanisms by modeling biologically meaningful shared patterns between gene expression responses and corresponding chemical descriptors for a large collection of annotated drug molecules. The second study in the chapter presents an effective pipeline to search for survival associated multi-view chromosomal regions that are dependent among paired data sources, such as gene expression paired with copy number or methylation patterns.

While the studies in Chapter 4 and Chapter 5 present models suitable for a collection of data sample, Chapter 6 presents two schemes to relate collection of datasets. The first proposed scheme is a unified Bayesian nonparametric multi-task model where the number of latent components are inferred automatically and the components are allowed to be shared across datasets. Evaluation on simulated and real-world textual data collections reveal superior performance of the proposed model over the state-of-the-art method when the number of available samples within the dataset-of-interest is low; a case where relating datasets is most needed. The second proposed scheme is a mixture of models, which

is both rapidly computable and scalable. Results on an annotated gene expression database indicate that the modeled data-driven relationships between datasets match well with citations between the corresponding research articles, and even found mistakes in the database annotations.

The recent growth in the development of genomic sequencing capability has led to an exponential growth in the amount of publicly available sequence data [120]. This poses a fundamental challenge of scalability for computational modeling. There are at least two complementary approaches to address the scalability challenge; first efficient pre-processing schemes are required to compress the large raw data, and second the methods need to be parallelized where the computation would need to move to the data rather than moving the data to the computation. In addition to adapting the proposed algorithms in the thesis, intelligent preprocessing that helps filter out unwanted background data, for example by utilizing the standardized downstream ontologies or by restricting background data with suitable hashing functions, provides a promising direction for the future.

Bibliography

[1] J. Rung and A. Brazma, "Reuse of public genome-wide gene expression data," *Nature Reviews Genetics*, vol. 14, pp. 89–99, 2012.

[2] O. G. Troyanskaya, "Putting microarrays in a context: integrated analysis of diverse biological data," *Briefings in bioinformatics*, vol. 6, pp. 34–43, 2005.

[3] H. Parkinson *et al.*, "ArrayExpress update — an archive of functional genomics experiments to the atlas of gene expression," *Nucleic Acids Research*, vol. 37, pp. D868–D872, 2009.

[4] Y. Zhu *et al.*, "GEOmetadb: Powerful alternative search engine for the gene expression omnibus," *Bioinformatics*, vol. 24, pp. 2798–2800, 2008.

[5] L. Hunter *et al.*, "GEST: A gene expression search tool based on a novel bayesian similarity metric," *Bioinformatics*, vol. 17, pp. S115–S122, 2001.

[6] D. Husmeier, R. Dybowski, and S. Roberts, *Probabilistic Modeling in Bioinformatics and Medical Informatics*. Springer, New York, USA, 2005.

[7] D. Koller and N. Friedman, *Probabilistic graphical models: Principles and techniques*. MIT Press, Cambridge, Massachusetts, USA, 2009.

[8] E. Georgii, J. Salojärvi, M. Brosché, J. Kangasjärvi, and S. Kaski, "Targeted retrieval of gene expression measurements using regulatory models," *Bioinformatics*, vol. 28, pp. 2349–2356, 2012.

[9] C. Huttenhower and O. Hofmann, "A quick guide to large-scale genomic data mining," *PLoS Computational Biology*, vol. 6, p. e1000779, 2010.

[10] H. Lodish *et al.*, *Molecular Cell Biology*. Wiley Online Library, New York, USA, 2000.

[11] B. Alberts *et al.*, *Molecular Biology of the Cell*. Garland Science, New York, USA, 4th ed., 2002.

[12] R. M. Simon, *Design and analysis of DNA microarray investigations*. Springer, New York, USA, 2003.

[13] F. Crick *et al.*, "Central dogma of molecular biology," *Nature*, vol. 227, pp. 561–563, 1970.

[14] D. Latchman, *Gene regulation*. Taylor & Francis, Abingdon, UK, 2007.

[15] A. Saha, J. Wittmeyer, and B. R. Cairns, "Chromatin remodelling: the industrial revolution of DNA around histones," *Nature Reviews Molecular Cell Biology*, vol. 7, pp. 437–447, 2006.

[16] L. J. Johnson and P. J. Tricker, "Epigenomic plasticity within populations: its evolutionary significance and potential," *Heredity*, vol. 105, pp. 113–121, 2010.

[17] K. Chen and N. Rajewsky, "The evolution of gene regulation by transcription factors and microRNAs," *Nature Reviews Genetics*, vol. 8, pp. 93–103, 2007.

[18] J. Sobek *et al.*, "Microarray technology as a universal tool for high-throughput analysis of biological systems," *Combinatorial Chemistry & High-throughput Screening*, vol. 9, pp. 365–380, 2006.

[19] D. Lockhart, "Expression monitoring by hybridization to high-density oligonucleotide arrays," *Nature Biotechnology*, vol. 14, pp. 1675–80, 1996.

[20] R. J. Lipshutz, S. P. Fodor, T. R. Gingeras, and D. J. Lockhart, "High density synthetic oligonucleotide arrays," *Nature Genetics*, vol. 21, pp. 20–24, 1999.

[21] E. Wit and J. McClure, *Statistics for Microarrays: Design, Analysis and Inference*. Wiley, Sussex, England, 2004.

[22] 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, USA*, vol. 95, pp. 14863–14868, 1998.

[23] M. J. Van De Vijver *et al.*, "A gene-expression signature as a predictor of survival in breast cancer," *New England Journal of Medicine*, vol. 347, pp. 1999–2009, 2002.

[24] G. V. Glinsky *et al.*, "Gene expression profiling predicts clinical outcome of prostate cancer," *Journal of Clinical Investigation*, vol. 113, pp. 913–923, 2004.

[25] J. Lamb *et al.*, "The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease," *Science*, vol. 313, pp. 1929–1935, 2006.

[26] J. Corander, T. Aittokallio, S. Ripatti, and S. Kaski, "The rocky road to personalized medicine: computational and statistical challenges," *Personalized Medicine*, vol. 9, pp. 109–114, 2012.

[27] M. A. Hamburg and F. S. Collins, "The path to personalized medicine," *New England Journal of Medicine*, vol. 363, pp. 301–304, 2010.

[28] G. S. Ginsburg and J. J. McCarthy, "Personalized medicine: revolutionizing drug discovery and patient care," *Trends in Biotechnology*, vol. 19, pp. 491–496, 2001.

[29] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, 2004.

[30] J. J. Goeman and P. Bühlmann, "Analyzing gene expression in terms of gene sets: methodological issues," *Bioinformatics*, vol. 23, pp. 980–987, 2007.

[31] D. Nam and S. Y. Kim, "Gene-set approach for expression pattern analysis," *Briefings in Bioinformatics*, vol. 9, pp. 189–197, 2008.

[32] E. Segal *et al.*, "A module map showing conditional activity of expression modules in cancer," *Nature Genetics*, vol. 36, pp. 1090–1098, 2004.

[33] A. Subramanian *et al.*, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences, USA*, vol. 102, pp. 15545–15550, 2005.

[34] V. K. Mootha *et al.*, "PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes," *Nature Genetics*, vol. 34, pp. 267–273, 2003.

[35] G. Abraham, A. Kowalczyk, S. Loi, I. Haviv, and J. Zobel, "Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context," *BMC Bioinformatics*, vol. 11, p. 277, 2010.

[36] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, pp. 27–30, 2000.

[37] M. Ashburner *et al.*, "Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nature Genetics*, vol. 25, pp. 25–9, 2000.

[38] L. Tian *et al.*, "Discovering statistically significant pathways in expression profiling studies," *Proceedings of the National Academy of Sciences, USA*, vol. 102, pp. 13544–13549, 2005.

[39] T. Barrett *et al.*, "NCBI GEO: Archive for high-throughput functional genomic data," *Nucleic Acids Research*, vol. 37, pp. D885–D890, 2009.

[40] M. Lukk *et al.*, "A global map of human gene expression," *Nature Biotechnology*, vol. 28, pp. 322–324, 2010.

[41] J. Shingara *et al.*, "An optimized isolation and labeling platform for accurate microRNA expression profiling," *RNA*, vol. 11, pp. 1461–1470, 2005.

[42] J. G. Hacia *et al.*, "Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays," *Nature Genetics*, vol. 22, pp. 164–167, 1999.

[43] A. S. Weinmann, P. S. Yan, M. J. Oberley, T. H.-M. Huang, and P. J. Farnham, "Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis," *Genes & Development*, vol. 16, pp. 235–244, 2002.

[44] G. Moran *et al.*, "Comparative genomics using Candida albicans DNA microarrays reveals absence and divergence of virulence-associated genes in Candida dubliniensis," *Microbiology*, vol. 150, pp. 3363–3382, 2004.

[45] J. R. Pollack *et al.*, "Genome-wide analysis of DNA copy-number changes using cDNA microarrays," *Nature Genetics*, vol. 23, pp. 41–46, 1999.

[46] N. P. Carter, "Methods and strategies for analyzing copy number variation using DNA microarrays," *Nature Genetics*, vol. 39, pp. S16–S21, 2007.

[47] R. Redon *et al.*, "Global variation in copy number in the human genome," *Nature*, vol. 444, pp. 444–454, 2006.

[48] J. L. Freeman *et al.*, "Copy number variation: new insights in genome diversity," *Genome Research*, vol. 16, pp. 949–961, 2006.

[49] R. Verhaak *et al.*, "Cancer Genome Atlas Research Network Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1," *Cancer Cell*, vol. 17, pp. 98–110, 2010.

[50] J. Downward, "Targeting RAS signalling pathways in cancer therapy," *Nature Reviews Cancer*, vol. 3, pp. 11–22, 2003.

[51] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer," *Cell*, vol. 100, pp. 57–70, 2000.

[52] E. Alpaydin, *Introduction to Machine Learning*. MIT Press, Cambridge, Massachusetts and London, England, 2004.

[53] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, New York, USA, 2006.

[54] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science, Boca Raton, Florida, USA, 2nd ed., 2003.

[55] W. Buntine *et al.*, "A scalable topic-based open source search engine," in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 228–234, 2004.

[56] M. Seeger, "Bayesian Modelling in Machine Learning: A Tutorial Review," *(Technical Report 161462), Departament of Computer Science, EPFL, Lausanne*, 2009.

[57] V. Brusic, J. S. Wilkins, C. A. Stanyon, and J. Zeleznikow, "Data learning: understanding biological data," in *Proceedings of the AAAI workshop on knowledge sharing across biological and medical knowledge based systems*, pp. 12–19, 1998.

[58] M. Loéve, *Probability Theory I*. Springer Verlag, New York, USA, 4th ed., 1977.

[59] S. Kullback, *Information theory and statistics*. John Wiley and Sons, New York, USA, 1959.

[60] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley series, New York, USA, 1991.

[61] P. Orbanz and Y. W. Teh, "Bayesian nonparametric models," in *Encyclopedia of Machine Learning*, pp. 81–89, Springer, 2010.

[62] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[63] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, pp. 183–233, 1999.

[64] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, pp. 721–741, 1984.

[65] G. Casella and E. I. George, "Explaining the Gibbs sampler," *The American Statistician*, vol. 46, pp. 167–174, 1992.

[66] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences, USA*, vol. 101, pp. 5228–5235, 2004.

[67] D. M. Blei *et al.*, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[68] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, "Evaluation methods for topic models," in *Proceedings of the Annual International Conference on Machine Learning*, pp. 1105–1112, 2009.

[69] M. A. Newton and A. E. Raftery, "Approximate Bayesian inference with the weighted likelihood bootstrap," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 56, pp. 3–48, 1994.

[70] R. M. Neal, "Annealed importance sampling," *Statistics and Computing*, vol. 11, pp. 125–139, 2001.

[71] J. Malone *et al.*, "Modeling sample variables with an experimental factor ontology," *Bioinformatics*, vol. 26, pp. 1112–1118, 2010.

[72] D. Edwards, *Introduction to Graphical Modelling*. Springer, New York, USA, 2000.

[73] J. Schäfer and K. Strimmer, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, p. 32, 2005.

[74] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, pp. 267–288, 1996.

[75] P. M. Williams, "Bayesian regularization and pruning using a Laplace prior," *Neural Computation*, vol. 7, pp. 117–143, 1995.

[76] S. Rogers and M. Girolami, "A Bayesian regression approach to the inference of regulatory networks from gene expression data," *Bioinformatics*, vol. 21, pp. 3131–3137, 2005.

[77] D. J. MacKay, "Bayesian interpolation," *Neural Computation*, vol. 4, pp. 415–447, 1992.

[78] D. Heckerman, D. Geiger, and D. M. Chickering, "Learning Bayesian networks: The combination of knowledge and statistical data," *Machine Learning*, vol. 20, pp. 197–243, 1995.

[79] D. Madigan, J. York, and D. Allard, "Bayesian graphical models for discrete data," *International Statistical Review*, pp. 215–232, 1995.

[80] M. Grzegorczyk and D. Husmeier, "Improving the structure MCMC sampler for Bayesian networks by introducing a new edge reversal move," *Machine Learning*, vol. 71, pp. 265–305, 2008.

[81] N. Friedman and D. Koller, "Being Bayesian about network structure. A bayesian approach to structure discovery in Bayesian networks," *Machine Learning*, vol. 50, pp. 95–125, 2003.

[82] J. Besag, "Statistical analysis of non-lattice data," *Journal of the Royal Statistical Society. Series D (The Statistician)*, vol. 24, pp. 179–195, 1975.

[83] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis*. Cambridge University Press, New York, USA, 2004.

[84] C. J. Sherr, "Cancer cell cycles," *Science*, vol. 274, pp. 1672–1677, 1996.

[85] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *Cell*, vol. 144, pp. 646–674, 2011.

[86] J. N. Hirschhorn and M. J. Daly, "Genome-wide association studies for common diseases and complex traits," *Nature Reviews Genetics*, vol. 6, pp. 95–108, 2005.

[87] B. Scholkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, Cambridge, USA, 2001.

[88] C. M. Lee *et al.*, "Simultaneous non-negative matrix factorization for multiple large scale gene expression datasets in toxicology," *PLoS One*, vol. 7, 2012.

[89] G. Cruciani, M. Pastor, and W. Guba, "VolSurf: a new tool for the pharmacokinetic optimization of lead compounds," *European Journal of Pharmaceutical Sciences*, vol. 11, pp. S29–S39, 2000.

[90] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, pp. 321–377, 1936.

[91] M. Borga, H. Knutsson, and T. Landelinus, "Learning Canonical Correlations," in *Proceedings of the Scandinavian Conference on Image Analysis*, vol. 1, pp. 1–8, 1997.

[92] F. R. Bach and M. I. Jordan, "Kernel independent component analysis," *Journal of Machine Learning Research*, vol. 3, pp. 1–48, 2003.

[93] H. D. Vinod, "Canonical ridge and econometrics of joint production," *Journal of Econometrics*, vol. 4, pp. 147–166, 1976.

[94] A. Klami and S. Kaski, "Local dependent components," in *Proceedings of the International Conference on Machine Learning*, pp. 425–432, 2007.

[95] J. T. Huse and E. C. Holland, "Targeting brain cancer: advances in the molecular pathology of malignant glioma and medulloblastoma," *Nature Reviews Cancer*, vol. 10, pp. 319–331, 2010.

[96] J. M. Nigro *et al.*, "Integrated array-comparative genomic hybridization and expression array profiles identify clinically relevant molecular subtypes of glioblastoma," *Cancer Research*, vol. 65, pp. 1678–1686, 2005.

[97] K. Ovaska *et al.*, "Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme," *Genome Medicine*, vol. 2:65, 2010.

[98] L. Lahti, S. Myllykangas, S. Knuutila, and S. Kaski, "Dependency detection with similarity constraints," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6, 2009.

[99] F. R. Bach and M. I. Jordan, "A probabilistic interpretation of Canonical correlation analysis," *(Technical Report 688), Department of Statistics, University of California, Berkeley*, 2005.

[100] C. Archambeau, N. Delannay, and M. Verleysen, "Robust probabilistic projections," in *Proceedings of the International Conference on Machine Learning*, pp. 33–40, 2006.

[101] R. Louhimo, T. Lepikhova, O. Monni, and S. Hautaniemi, "Comparative analysis of algorithms for integration of copy number and expression data," *Nature Methods*, vol. 9, pp. 351–355, 2012.

[102] D. Kleinbaum and M. Klein, *Survival analysis: A self-learning approach*. Springer, New York, USA, 2005.

[103] M. Bradburn, T. Clark, S. Love, and D. Altman, "Survival analysis part II: Multivariate data analysis: an introduction to concepts and methods," *British Journal of Cancer*, vol. 89, pp. 431–436, 2003.

[104] B. Rasheed, G. N. Fuller, A. H. Friedman, D. D. Bigner, and S. H. Bigner, "Loss of heterozygosity for 10q loci in human gliomas," *Genes, Chromosomes and Cancer*, vol. 5, pp. 75–82, 1992.

[105] K. Tokiyoshi, T. Yoshimine, M. Maruno, A. Muhammad, and T. Hayakawa, "Accumulation of allelic losses on chromosome 10 in human gliomas at recurrence," *Clinical Molecular Pathology*, vol. 49, pp. M218–M222, 1996.

[106] J. A. Tynan, F. Wen, W. J. Muller, and R. G. Oshima, "ETS2-dependent microenvironmental support of mouse mammary tumors," *Oncogene*, vol. 24, pp. 6870–6876, 2005.

[107] T. E. Sussan, F. L. Annan Yang, M. C. Ostrowski, and R. H. Reeves, "Trisomy represses ApcMin-mediated tumours in mouse models of Down's syndrome," *Nature*, vol. 451, pp. 73–75, 2008.

[108] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour, "Microarray data analysis: from disarray to consolidation and consensus," *Nature Reviews Genetics*, vol. 7, pp. 55–65, 2006.

[109] R. Clarke *et al.*, "The properties of high-dimensional data spaces: implications for exploring gene and protein expression data," *Nature Reviews Cancer*, vol. 8, pp. 37–49, 2008.

[110] H. Huang *et al.*, "Bayesian approach to transforming public gene expression repositories into disease diagnosis databases," *Proceedings of the National Academy of Sciences, USA*, vol. 107, pp. 6823–6828, 2010.

[111] A. Tanay, I. Steinfeld, M. Kupiec, and R. Shamir, "Integrative analysis of genome-wide experiments in the context of a large high-throughput data compendium," *Molecular Systems Biology*, vol. 1, pp. E1–E10, 2005.

[112] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, pp. 41–75, 1997.

[113] G. Tseng, D. Ghosh, and E. Feingold, "Comprehensive literature review and statistical considerations for microarray meta-analysis," *Nucleic Acids Research*, vol. 40, pp. 3785–3799, 2012.

[114] J. Russ and M. Futschik, "Comparison and consolidation of microarray data sets of human tissue expression," *BMC Genomics*, vol. 11:305, 2010.

[115] S. Suthram *et al.*, "Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets," *PLoS Computational Biology*, vol. 6:e1000662, 2010.

[116] C. Huttenhower and O. Troyanskaya, "Assessing the functional structure of genomic data," *Bioinformatics*, vol. 24, pp. i330–8, 2008.

[117] Y. W. Teh *et al.*, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, pp. 1566–1581, 2006.

[118] T. Griffiths and Z. Ghahramani, "Infinite latent feature models and the Indian buffet process," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 475–482, 2006.

[119] Y. W. Teh, "Stick-breaking construction for the Indian buffet process," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 1–10, 2007.

[120] B. Berger, J. Peng, and M. Singh, "Computational solutions for omics data," *Nature Reviews Genetics*, vol. 14, pp. 333–346, 2013.

[121] K. Clarkson, "Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm," in *Proceedings of the ACM-SIAM symposium on Discrete algorithms*, pp. 922–931, 2008.

[122] Y. Chen and X. Ye, "Projection onto a simplex," *arXiv preprint*, vol. arXiv:1101.6081, 2011.

[123] G. Gerber *et al.*, "Automated discovery of functional generality of human gene expression programs," *PLoS Computational Biology*, vol. 3:e148, 2007.

[124] J. Engreitz *et al.*, "Content-based microarray search using differential expression profiles," *BMC Bioinformatics*, vol. 11:603, 2010.

[125] J. Caldas *et al.*, "Data-driven information retrieval in heterogeneous collections of transcriptomics data links *SIM2s* to malignant pleural mesothelioma," *Bioinformatics*, vol. 28, pp. i246–i253, 2012.

[126] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Machine Learning*, vol. 39, pp. 103–134, 2000.

[127] M. Thomas, "Estimating a Dirichlet distribution (2000)." `http://research.microsoft.com/~minka/papers/dirichlet`, 2003.

[128] W. Li and A. McCallum, "Pachinko allocation: DAG-structured mixture models of topic correlations," in *Proceedings of the International Conference on Machine Learning*, pp. 577–584, 2006.

[129] S. van Dongen, *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, 2000.

[130] J. Sammon, "A nonlinear mapping for data structure analysis," *IEEE Transactions on Computers*, vol. 18, pp. 401–409, 1969.

Aalto-DD 108/2014

ISBN 978-952-60-5780-4
ISBN 978-952-60-5781-1 (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