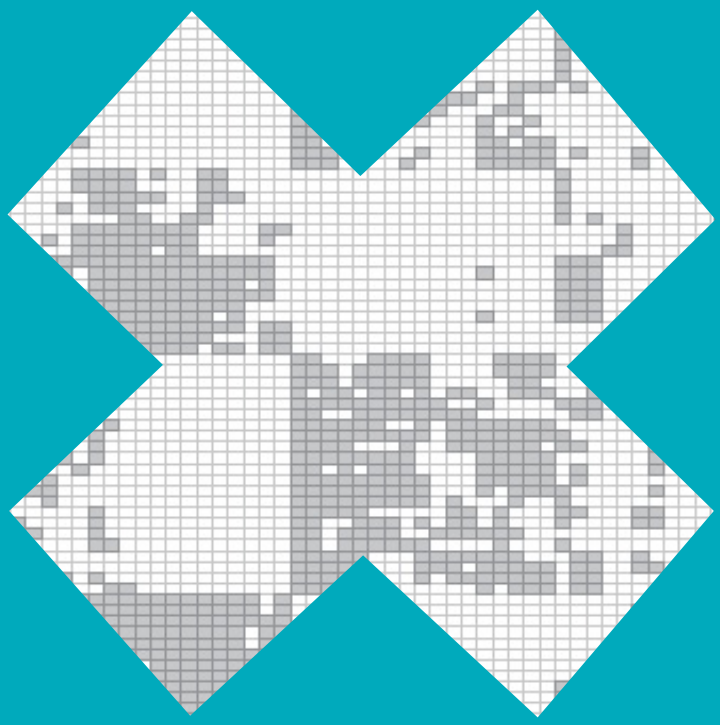
Department of Information and Computer Science

Graphical Models for Biclustering and Information Retrieval in Gene Expression Data

José Caldas



Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 33/2012

Graphical Models for Biclustering and Information Retrieval in Gene Expression Data

José Caldas

Doctoral dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the School of Science for public examination and debate in Auditorium T2 at the Aalto University School of Science (Espoo, Finland) on the 20th of April 2012 at noon.

Aalto University
School of Science
Department of Information and Computer Science

Supervisor
Samuel Kaski, D.Sc. (Tech.)

Instructor
Leo Lahti, D.Sc. (Tech.)

Preliminary examiners
Guido Sanguinetti, D.Phil.
Sampsa Hautaniemi, D. Sc. (Tech.)

Opponent
Eric Xing, PhD, PhD

Aalto University publication series
DOCTORAL DISSERTATIONS 33/2012

ISBN 978-952-60-4558-0 (printed)
ISBN 978-952-60-4559-7 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Unigrafia Oy
Helsinki 2012

Finland

The dissertation can be read at http://lib.tkk.fi/Diss/

441 697
Printed matter

Aalto University

Abstract

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

Author
José Caldas

Name of the doctoral dissertation
Graphical Models for Biclustering and Information Retrieval in Gene Expression Data

Publisher School of Science

Unit Department of Information and Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 33/2012

Field of research Bioinformatics

Manuscript submitted 28 October 2011 **Manuscript revised** 9 January 2012

Date of the defence 20 April 2012 **Language** English

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

Abstract

The cell coordinates its biological response to the environment partly via the selective synthesis of thousands of unique RNA and protein molecules. Understanding the molecular biology of the cell is thus essential to the advancement of areas such as health care, agriculture, and energy production, but requires the ability to simultaneously acquire information about thousands of molecules in a sample. Recent high-throughput measurement technologies address this concern. While being useful, they generate a high volume of data and bring in methodological challenges, effectively shifting the bottleneck in molecular biology research from data acquisition to data analysis. In particular, an important challenge is the genome-wide analysis of how RNA is transcribed under different conditions, organisms, and tissues, a process known as gene expression.

When developing computational methods for biological data analysis tasks, probabilistic frameworks constitute promising approaches due to their flexibility, soundness, and ability to handle noisy data. In this thesis, the contributions are in the development of probabilistic methods for two relevant tasks in genome-wide gene expression analysis, namely biclustering and information retrieval.

Biclustering concerns the simultaneous grouping of objects, e.g., genes, and conditions. The first contribution is the development of a Bayesian extension to an existing biclustering model. The second contribution is a novel probabilistic method that allows deriving a hierarchical organization of microarrays in a gene expression data set and at the same time indicate the genes that characterize the hierarchy. Finally, the third contribution is a general probabilistic biclustering framework that easily lends itself to different data types and model assumptions.

Information retrieval in gene expression data is needed because of the increasing amount of available data stored in public databases. Two probabilistic methods for information retrieval are proposed. The models are used in a series of biological case studies that show how the proposed approaches have the potential to accelerate biological research by jointly analyzing data from different studies. In particular, several connections between biological conditions found by the models either correspond to existing biological knowledge or were used in a confirmatory follow-up study to obtain novel biological findings.

Keywords Probabilistic Modelling, Bayesian Network, Biclustering, Information Retrieval, Transcriptomics

ISBN (printed) 978-952-60-4558-0 **ISBN (pdf)** 978-952-60-4559-7

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

Location of publisher Espoo **Location of printing** Helsinki **Year** 2012

Pages 158 **The dissertation can be read at** http://lib.tkk.fi/Diss/

Preface

This thesis describes my research work as a doctoral student in the Statistical Machine Learning and Bioinformatics (MI) Group of the Department of Information and Computer Science (ICS), Aalto University School of Science. My work was funded by the Portuguese Science and Technology Foundation (FCT), the Finnish Funding Agency for Technology and Innovation (TEKES), and the Pattern Analysis, Statistical Modelling and Computational Learning Network of Excellence (PASCAL 2 EU Network of Excellence). I am also a member of the Helsinki Institute for Information Technology HIIT.

I would like to thank my supervisor, Prof. Samuel Kaski, for granting me the opportunity to study machine learning and bioinformatics, for enthusiastically guiding my research work, and for the flexibility shown at all levels. I am also grateful to my thesis reviewers, Dr. Sampsa Hautaniemi and Dr. Guido Sanguinetti, for their useful remarks. Finally, I would like to thank my former M.Sc. supervisors, Profs. Arlindo Oliveira and Ana Teresa Freitas, who introduced me to the world of bioinformatics.

I am indebted to all past and present MI group members for the friendly and stimulating workplace environment. I am particularly thankful to Dr. Janne Sinkkonen and Dr. Arto Klami, for their enlightening perspectives on machine learning, to Dr. Leo Lahti, for sharing his knowledge on biology and bioinformatics, and to M.Sc. Melih Kandemir, for the many discussions on all sorts of topics. I would also like to show my gratitude to Prof. Erkki Oja, who warmly welcomed me to the ICS department, to Prof. Harri Lähdesmäki, who introduced me to new topics in computational biology, and to Dr. Ricardo Vigário and M.Sc. Nicolau Gonçalves, who often made me feel at home away from home. On many occasions, I benefited from Dr. Miki Sirola's expertise on computer systems and Ms. Leila Koivisto's expertise on administrative issues, both of whom I am ex-

tremely grateful to. I am also deeply thankful to Ms. Ritva Siukkola from the Center for International Mobility (CIMO), who so kindly allowed me to live in two of CIMO's excellent apartments.

While most of my research work was performed in Helsinki, I also visited the Brazma group at the European Bioinformatics Institute (EBI) in 2007 and 2008. I would like to express my gratitude to Dr. Alvis Brazma, for welcoming me into his group and for our fruitful collaboration. In particular, I had the privilege and pleasure of collaborating during most of my doctoral studies with Dr. Nils Gehlenborg, from whom I learned so much.

I was lucky to have had the chance of collaborating with biologists who, in a way, brought our methods to life, by using them as support for their wet lab experiments. I would therefore like to thank my collaborators and co-authors Dr. Eeva Kettunen and Prof. Sakari Knuutila for their insight, openness, enthusiasm, and patience during our interdisciplinary collaboration.

It would not have been possible to write this thesis without the support of those dearest to me. I would like to thank Guilherme for our lifelong friendship and the many adventures we shared. I would also like to thank Raquel for her love, support, and understanding, which are ever so important to me. Finally, I would like to thank my maternal family for their unflinching love, protection, and support during both good and bad times. This thesis is dedicated to the memory of my parents, António José and Maria Fernanda, to the memory of my maternal grandparents, Fernando and Leopoldina, and to my maternal family.

Lisbon, Portugal, March 13, 2012,

José Caldas

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 and Samuel Kaski. Bayesian biclustering with the plaid model. In *Proceedings of the 2008 IEEE International Workshop on Machine Learning for Signal Processing XVIII*, José Príncipe, Deniz Erdogmus, and Tulay Adali (editors), pages 291–296, IEEE, Piscataway, N.J., October 2008.

II 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 (ISMB/ECCB 2009 Conference Proceedings), June 2009.

III José Caldas and Samuel Kaski. Hierarchical generative biclustering for microRNA expression analysis. *Journal of Computational Biology*, 18(3):251–261 (RECOMB 2010 Special Issue), March 2011.

IV José Caldas and Samuel Kaski. A mixture-of-experts approach to biclustering. *Submitted to a journal*, 10 pages, 2011.

V 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, January 2012.

Author's Contribution

Publication I: "Bayesian biclustering with the plaid model"

Derived and implemented a Bayesian formulation for the plaid model; ran and partly designed the experiments; co-wrote the manuscript.

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

Co-designed and implemented all components of the proposed framework, except for the visualization methods; designed and ran the experiments; co-wrote the manuscript.

Publication III: "Hierarchical generative biclustering for microRNA expression analysis"

Designed and implemented the probabilistic hierarchical biclustering framework; designed and ran the experiments; co-wrote the manuscript.

Publication IV: "A mixture-of-experts approach to biclustering"

Designed and implemented the mixture-of-experts framework; designed and ran the experiments; co-wrote the manuscript.

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

Partly designed the information retrieval framework; partly designed the experiments; suggested the follow-up study for the connection between mesothelioma and *SIM2s*; co-wrote the manuscript.

List of Abbreviations and Symbols

In this thesis, boldface symbols represent multidimensional entities, while normal font symbols represent scalar entities. Uppercase symbols are used to distinguish random variables from their specific instantiations and also to distinguish between matrices (uppercase) and vectors (lowercase).

A	Adenine
ANOVA	Analysis of Variance
BN	Bayesian Network
C	Cytosine
C_T	Threshold Cycle
cDNA	Complementary DNA
CRP	Chinese Restaurant Process
d-separation	Directed Separation
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide
Dom	Domain
DP	Dirichlet Process
EFO	Experimental Factor Ontology
EM	Expectation-Maximization
EMT	Epithelial-Mesenchymal Transition
ER	Estrogen Receptor
ES	Enrichment Score
E-Step	Expectation Step
G	Guanine
GEM	Griffiths, Engen, and McCloskey
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
HDP	Hierarchical Dirichlet Process

IBP	Indian Buffet Process
IR	Information Retrieval
LDA	Latent Dirichlet Allocation
M-Step	Maximization Step
MAP	Maximum *a posteriori*
MCMC	Markov Chain Monte Carlo
miRNA	Micro RNA
ML	Maximum Likelihood
MPM	Malignant Pleural Mesothelioma
mRNA	Messenger RNA
MSigDB	Molecular Signatures Database
nCRP	Nested CRP
NDCG	Normalized Discounted Cumulative Gain
RNA	Ribonucleic acid
RNAi	RNA interference
RT-PCR	Real-Time Polymerase Chain Reaction
SIM2s	*Single-minded homolog 2*, short isoform
T	Thymine
tRNA	Transfer RNA
X, Y	Scalar random variables
x, y	Scalar data samples
$\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$	Multidimensional random variables, matrices
$\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$	Multidimensional data samples, vectors
$\boldsymbol{\Theta}$	Multidimensional model parameter
$\boldsymbol{\theta}$	Multidimensional model parameter instantiation
$p(X)$	Probability density function of X
$P(X)$	Probability distribution of X
$P(X, Y)$	Joint probability distribution X and Y
$P(Y\|X)$	Conditional probability distribution Y given X
$\mathrm{E}_{p(\cdot)}[\cdot]$	Expectation over a probability density
$\mathrm{D}_{\mathrm{KL}}(Q\|\|P)$	Kullback-Leibler Divergence between Q and P
$\mathrm{H}[\cdot]$	Entropy functional
$\exp\{\cdot\}$	Exponential function
$\boldsymbol{I}$	Identity Matrix
$\pi(i)$	Parent of node i in a graph
$\mathrm{Bernoulli}(p)$	Bernoulli distribution with parameter p
$\mathrm{Beta}(\alpha_1, \alpha_2)$	Beta distribution with parameters α_1 and α_2
$\mathrm{Dir}(\alpha_1, \ldots, \alpha_k)$	Dirichlet distribution with parameters $\alpha_1, \ldots, \alpha_k$

$\mathrm{DP}(\alpha_0, G_0)$	Dirichlet Process with concentration α_0 and base measure G_0
$\mathrm{GEM}(\alpha_0)$	GEM distribution with parameter α_0
$\mathrm{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
$\mathrm{Poisson}(\lambda)$	Poisson distribution with parameter λ

1. Introduction

Recent advances in biological high-throughput measurement technologies have led to several new challenges in data analysis. Technologies such as DNA microarrays hold great potential for uncovering the laws governing the cell at the molecular level, but generate large amounts of noisy high-dimensional data. Additionally, the samples and biological conditions underlying the data correspond to highly complex and heterogeneous phenotypes, *e.g.*, tumor specimens [54].

One of the chief tasks in molecular biology is to understand the phenomenon of transcription by which DNA is used to synthesize RNA transcripts. DNA microarrays have been effectively used for this purpose, but several computational tasks related to the analysis of gene expression DNA microarray data still lack satisfactory solutions. Among those tasks are biclustering and information retrieval, which are the concern of this thesis.

The proposed methods for handling biclustering and information retrieval problems belong to the graphical model family. Graphical models such as Bayesian networks [110] are probabilistic frameworks for modelling partially observable, stochastic systems with noisy measurements [119]. The observed data is assumed to be the product of a probabilistic *generative process* involving unobserved, hidden *latent variables* that convey the assumptions about the underlying system. Reverse engineering the latent variables corresponds to a well-defined *posterior inference* problem that typically yields efficient computational procedures. Graphical models are currently being successfully used as general problem-solving tools involving noisy data stemming from natural processes, but remain an open research topic. In general, graphical modelling can be seen as an inductive formalism, providing a departure from classical, deductive computer science frameworks [14]. The proposed methods make use of recent

advances in graphical modelling, such as topic models [15] and stochastic processes on combinatorial structures [113], in order to design models for biclustering and information retrieval problems.

Biclustering is an unsupervised learning task in which the aim is to analyze an input data matrix in order to find submatrices with desirable properties. Biclustering can be seen as a generalization of clustering; it is useful in gene expression analysis because it allows finding groups of genes which are related in a subset of the studied conditions. In this thesis, three novel probabilistic biclustering models are proposed. The first model is a Bayesian extension of the plaid model [82]. The second model is based on a stochastic process on trees known as the nested Chinese restaurant process [14] and allows obtaining a hierarchy of conditions, associating with each part of the hierarchy the genes that are co-expressed in the corresponding conditions. The third model is a general biclustering framework akin to the mixture-of-experts model family [10, 68]. This model also makes use of a nonparametric Bayesian prior, in this case the Indian buffet process [45], and is easily adaptable to varying assumptions and input data types, therefore providing a flexible and general framework for performing biclustering.

Information retrieval is a well-known task in the natural language and Internet domains, and critically relies on the availability of large amounts of data. As the use of DNA microarrays for measuring gene expression has become ubiquitous, thousands of studies have been deposited in databases such as ArrayExpress [108], which brings in the challenge of retrieving relevant microarrays to a given query. In this thesis, two related methods are proposed for querying with and retrieving biological conditions corresponding to subsets of microarrays. The proposed frameworks are based on differential expression, with studies being converted into sets of comparisons between conditions. Then, latent variable mixture models are coupled with a model-based relevance measure, in order to both infer patterns of differential expression and perform information retrieval. The proposed methods are competitive with existing approaches and the results led to a follow-up study in which it was found that the gene *SIM2s* is under-expressed in malignant pleural mesothelioma patients, which constitutes a novel biological finding based on an information retrieval search engine.

This thesis is organized as follows: The first two chapters provide brief background material on molecular biology, transcriptomics, and graphi-

cal modelling. The remaining two chapters describe the contributions on biclustering and information retrieval, respectively.

2. Molecular Biology and Transcriptomics

This chapter provides basic background material on molecular biology, gene expression measurement technologies, and computational analysis of gene expression data.

2.1 Basic Aspects of Molecular Cell Biology

The genetic material contained in a cell is known as its *genome*. The genome has the information required for the cell to function within its environment. While in some viruses the genome is made of ribonucleic acid (RNA), in all known organisms and most viruses the genome is composed of deoxyribonucleic acid (DNA), which is identically present in every cell of a given organism. In organisms known as eukaryotes, most DNA is stored in an enclosed cellular compartment known as the *nucleus*. Nuclear DNA is organized as a set of linear DNA molecules known as *chromosomes* [74].

The DNA molecule forms a double helix structure with a sugar-phosphate backbone, containing two strands of nitrogenous bases of which there are four types: Adenine (A), thymine (T), cytosine (C), and guanine (G). Bases in opposing strands form hydrogen bonds with the complementary base pairing rules A–T and G–C.

Within a DNA molecule there is a number of functional units known as *genes*. Genes are *transcribed* into complementary RNA sequences, which then undergo a maturation process. The resulting mature RNA may be an end-product, in the case of non-coding RNAs such as transfer RNAs (tRNAs) or microRNAs (miRNAs), or it may be *translated* into a protein. The latter type of RNA is known as *messenger RNA* (mRNA). Proteins possess a large number of roles, from enzymatic activity to DNA repair or transport of other molecules across the cell membrane.

The process of transcription followed by translation, shown in Figure 2.1, is the basic means by which the information encoded in DNA elicits a cellular response to the environment. Understanding the principles that govern transcription and translation is thus essential for understanding the biology of organisms and improve healthcare. The process of going from DNA to RNA and protein is known as *gene expression*, although the same designation is usually applied to refer only to the process of transcription. In this thesis, "gene expression" is used as a synonym for transcription. The study of transcriptional data is known as *transcriptomics*.

The regulation of transcription and translation is a multilayered process. In order to fit within the nucleus, DNA binds to histone proteins, forming a structure known as chromatin; this packaging process makes some genes less prone to transcription than others, with chromatin remodelling modulating transcription [120]. Molecules such as transcription factor proteins may bind specific DNA regions, altering the transcription rate of one or more genes [37]. Regulatory RNAs such as miRNAs may post-transcriptionally repress or degrade mRNA transcripts [22]. Finally, proteins may undergo post-translational modifications such as phosphorylation or addition of chemical groups [93].

The system of regulatory interactions constitutes a series of modular, complex networks [71], which drive biological processes such as development [85] or metastasis [83]. In general, genes work not as isolated units, but rather in the context of larger networks and pathways [89].

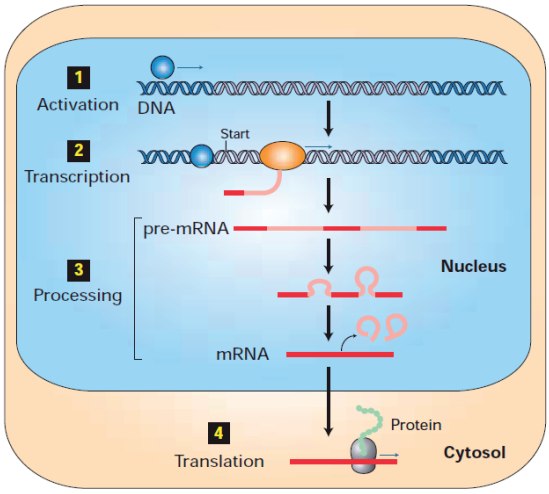


Figure 2.1. Transcription and translation in an eukaryotic cell. Figure adapted from Lodish *et al*. [89].

2.2 Gene Expression Measurement Technologies

Biological processes in the cell are not only complex but involve a large number of unique molecules. For instance, the number of genes in the human genome is estimated at 20,000–25,000 [127]. Recent high-throughput technologies that produce data about thousands of molecules in parallel are thus extremely important in the molecular biology field.

Several types of high-throughput assays exist, for instance to measure protein-DNA interactions [65] or to detect single-nucleotide polymorphisms [150]. In this thesis, the focus is on gene expression data generated via DNA microarrays. Real-time polymerase chain reaction (RT-PCR) assays are also briefly described, as they were used in Publication V to validate an *in silico* computational prediction.

While high-throughput sequencing technologies are emerging [151], DNA microarrays are the current standard for high-throughput mRNA expression profiling. Gene expression microarray data sets are routinely deposited in databases such as ArrayExpress [108] or the Gene Expression Omnibus [6]. A DNA microarray consists of a substrate on which single-stranded DNA sequences or *probes* are deposited [33]. Probes are designed so that they are complementary to known transcripts. On *single-channel* microarrays, the mRNA from the sample of interest is reverse transcribed into fluorescent dye-labelled complementary DNA (cDNA), which is then also deposited on the microarray. The cDNA sequences that hybridize to complementary probes are retained in the microarray. The resulting light intensity of each spot is a measure of the deposited cDNA. In *two-channel* microarrays, two samples are simultaneously deposited on the microarray, with the fluorescent dye used to label each sample having a different emission wavelength. The light intensities in each spot are then informative of the relative transcript abundance between the two deposited samples.

Pre-processing of microarray data involves a series of steps aimed at ignoring erroneous spots, correcting for background intensities and systematic bias, and normalizing the expression data within and across microarrays. For microarray platforms in which a set of probes is used to detect the signal from any given gene, a further aggregation step is necessary to obtain a single numerical expression value per gene [42]. An important concern is how probes map to transcripts. It has been shown that, because probe designs are based on genomic annotations that be-

come outdated, probes are often measuring different transcripts from the ones they were originally meant to measure [29]. This has been used to show that relabelling probes according to updated annotations yields significantly different differential expression results [29].

RT-PCR is another technique for measuring gene expression [152]. After reverse transcribing sample mRNA into cDNA, PCR consists of a number of *cycles* in which the mixture is first heated to induce strand separation and then cooled. Two short DNA *primers* of opposing polarity, DNA polymerase, and deoxynucleotides (dNTPs) are present in excess. The two primers represent the start and end subsequences of a larger sequence which is to be amplified. The use of the two primers as templates for DNA replication in the cooling phase leads to the sequence of interest being ideally doubled every cycle. There is thus an exponential increase in the number of copies of the desired sequence, until dNTPs are depleted. In RT-PCR, the amplification data is collected throughout the PCR process, typically by use of dyes that become fluorescent in the presence of double-stranded DNA [74]. This process generates a *threshold cycle* (C_T), which is the cycle at which the fluorescence signal surpasses a given detection threshold [122]. In Publication V, fold-change values between two conditions from an RT-PCR assay are computed using the comparative C_t method, which is standard practice [122].

RT-PCR measurements are often taken to confirm a subset of results obtained via microarrays in gene expression profiling studies. While both approaches have inherent pitfalls, it has been shown that the correlation between microarray and RT-PCR measurements is significantly high among differentially expressed transcripts [101].

2.3 Differential Expression

Gene expression data is computationally analyzed in a multitude of tasks [3]. For instance, classification methods are commonly used to find the molecular basis that distinguishes between phenotypes [48], and unsupervised learning methods such as hierarchical clustering are already a standard part of transcription profiling studies [34]. Computational inference of gene expression-based biomarkers [147] and connecting diseases to drug responses at the transcriptional level [79] is also an active area of research with implications for personalized medicine [121, 132, 146].

One of the main paradigms in gene expression analysis is to look for *differentially expressed genes* between two phenotypes of interest, for instance healthy vs. diseased tissue or treatment vs. control. Differential expression tests in a high-throughput context bring in the additional challenge of controlling for multiple hypothesis testing, which is achieved via family-wise error rate correction methods such as Bonferroni correction [33] or alternatively via false discovery rate correction [129].

Classical hypothesis testing methods such as Student's t-test [99] are routinely used to detect differentially expressed genes [64]. However, lists of differentially expressed genes derived from independent studies often have a low overlap despite being enriched for common functional categories [116, 123].

Recently, the field has shifted towards more robust methods that directly test for the differential expression of a set of related genes [1, 46, 88, 103, 131, 141]. Gene sets may correspond for instance to known pathways or groups of miRNA or transcription factor targets, and can be obtained from databases such as the Molecular Signatures Database (MSigDB) [131]. Gene set differential expression tests are based on the notion that genes act in tandem within a larger context. This suggests the hypothesis that differences in gene expression patterns between biological conditions typically consist of a coherent accumulation of small changes in the expression of related genes, rather than large changes in the expression of a few isolated genes [100].

Gene set tests can be divided in two broad categories: Competitive and self-contained tests [46, 141]. A competitive test asserts if the genes in a given gene set tend to appear among the most differentially expressed genes. The main issue with competitive tests is that they do not directly test if a gene set is up or down-expressed. Self-contained tests, on the other hand, test against the null hypothesis that no gene in the gene set is differentially expressed. A caveat concerning self-contained tests is that a gene set is considered to be differentially expressed even if only one of its genes is effectively differentially expressed.

Publications II and V make use of Gene Set Enrichment Analysis (GSEA) [100, 131], a well-known competitive gene set test. GSEA is described in detail in Subsection 5.2.2.

3. Graphical Models

The aim of this chapter is to succinctly introduce some of the most important notions in probabilistic modelling, as well as the tools that were used in the publications discussed in this thesis. The chapter starts with a brief, practical description of some of the basic properties of probability distributions that are directly used in probabilistic modelling, avoiding a detailed measure-theoretic treatment (*e.g.*, [19]). Then, several aspects of graphical modelling are described, namely the structure of Bayesian networks, general techniques for inference and learning, recent advances in nonparametric Bayesian statistics that allow defining flexible priors on combinatorial structures, and relevance measures that can be used for performing model-based information retrieval.

3.1 Basic Properties of Probability Distributions

The *probability distribution* of a *random variable* X formalizes the notion of uncertainty or subjective belief among multiple outcomes [73]. Assuming X takes on a finite or countable set of values, the probability of a given outcome x is designated by $P(X = x)$, where P is a non-negative function and $\sum_x P(X = x) = 1$. For instance, if X represents the outcome of a fair coin toss, then $P(X = \text{heads}) = P(X = \text{tails}) = 0.5$. When X has a continuous support, the notion of probability distribution has to be adjusted due to $P(X = x)$ being zero for any given value x. In that case, a non-negative *probability density function* p is used instead, with the property that $\int p(x)\mathrm{d}x = 1$. The probability of an event is then defined as

$$P(a \leq X \leq b) = \int_a^b p(x)\mathrm{d}x. \tag{3.1}$$

A probability distribution can also be defined over a collection of random variables. The *joint distribution* of two random variables X and Y

is defined as $P(X, Y)$, with any particular assignment having the probability $P(X = x, Y = y)$. The *marginal probability* of X is defined as $P(X = x) = \sum_y P(X = x, Y = y)$.[1] Two random variables are *independent* when, for any x and y, $P(X = x, Y = y) = P(X = x)P(Y = y)$. Given the notions of independence and marginal probability, the *conditional distribution* of Y given X is defined as

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

This yields an equivalent definition of independence, whereby X and Y are independent when $P(Y|X) = P(Y)$ and vice-versa. The notion of independence can be generalized to conditional independence: X and Y are *conditionally independent* given Z when $P(X, Y|Z) = P(X|Z)P(Y|Z)$. From (3.2) it is straightforward to derive *Bayes' rule*, which states that

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

When having a set of random variables, their joint distribution can be factorized into a product of conditional distributions according to the *chain rule*,

$$P(X_1, \ldots, X_n) = P(X_1)P(X_2|X_1) \ldots P(X_n|X_1, \ldots, X_{n-1}), \tag{3.4}$$

with the above factorization being valid for any permutation of the random variables. Finally, the *expectation* of a function of the random variable $f(X)$ over $P(X)$ is defined as

$$\mathrm{E}\left[f(X)\right] = \sum_{x \in \mathrm{Dom(X)}} P(X = x)f(x), \tag{3.5}$$

where $\mathrm{Dom}(X)$ is the domain of X.

When defining conditional probability distributions, it is common to make use of conjugate distributions for analytical tractability purposes. A distribution $P(X)$ is *conjugate* to $P(Y|X)$ when $P(X|Y) \propto P(X, Y)$ is from the same parametric family as $P(X)$ [8, 40].

3.2 Graphical models

Graphical models [9, 73, 81, 110] are a general framework for the probabilistic modelling of a given system. In a graphical model, system components are modelled as random variables, represented as nodes in a graph,

[1] For continuous random variables, summation is replaced by integration.

where edges in that graph determine the conditional dependency structure of the model. The graph can be directed, undirected, or combine both types of edges [20]. In the publications featured in this thesis, the focus is solely on graphical models with directed acyclic graphs, also known as Bayesian networks (BNs).

3.2.1 General Structure

A BN specifies a directed acyclic graph [27] that defines the conditional independence structure of the corresponding model. Formally, let X_i designate the random variable corresponding to node i with parents $\pi(i)$. The joint probability of the set $\boldsymbol{X}$ of n random variables in the model is given by

$$P(\boldsymbol{X}) = \prod_{i=1}^{n} P(X_i|X_{\pi(i)}). \tag{3.6}$$

This decomposition does not depend on any specific parametric assumptions about each conditional probability distribution $P(X_i|X_{\pi(i)})$. An example of a BN is shown in Figure 3.1 [15], where the joint probability of the model variables factorizes as

$$P(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}) = P(\boldsymbol{\alpha})P(\boldsymbol{\beta}) \prod_{i=1}^{M} P(\boldsymbol{\theta}_i|\boldsymbol{\alpha}) \prod_{j=1}^{N} P(z_{ij}|\boldsymbol{\theta}_i)P(w_{ij}|z_{ij}, \boldsymbol{\beta}). \tag{3.7}$$

The consequences of the joint probability decomposition (3.6) can be fully

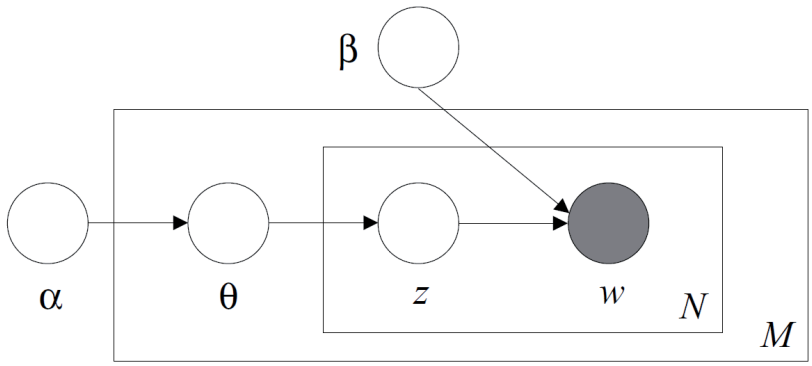


Figure 3.1. Example of a BN, in this case the Latent Dirichlet Allocation (LDA) model [15]. The grey node corresponds to observed data. A compact representation of sets of variables is achieved via the use of nested rectangles or *plates*. The outer plate indicates that there exist M variables θ_i. The inner plate indicates that, for each $i \in \{1, \ldots, M\}$, there exist N variables z_{ij} and w_{ij}. The directed arrow between the nodes in the inner plate indicates a paired dependency between each z_{ij} and the corresponding w_{ij}, while the directed arrow between θ and z indicates a dependency between each θ_i and the corresponding set of variables z_{ij}. Figure taken from Blei *et al*. [15].

understood via the concept of *directed separation* (d-separation). Intuitively, two sets of random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ are said to be d-separated given a third set of variables $\boldsymbol{Z}$ if there can be no information flow from any node $X \in \boldsymbol{X}$ to any node $Y \in \boldsymbol{Y}$ given the observed nodes $\boldsymbol{Z}$ [9]. Formally, this amounts to verifying that every undirected path between $X \in \boldsymbol{X}$ and $Y \in \boldsymbol{Y}$ obeys one of the following properties[2]:

$$X \rightarrow Z \rightarrow Y \,, Z \in \boldsymbol{Z}, \tag{3.8}$$

$$Y \rightarrow Z \rightarrow X \,, Z \in \boldsymbol{Z}, \tag{3.9}$$

$$X \leftarrow Z \rightarrow Y \,, Z \in \boldsymbol{Z}, \tag{3.10}$$

$$X \rightarrow W \leftarrow Y \,, W \notin Z, \forall_{V \in \text{descendants}(W)} V \notin Z. \tag{3.11}$$

The first two conditions above state that a chain of nodes between $X \in \boldsymbol{X}$ and $Y \in \boldsymbol{Y}$ must be blocked by an observed variable $Z \in \boldsymbol{Z}$. The third condition states that when X and Y have a common ancestor node, it must be an observed variable in $\boldsymbol{Z}$, so that X and Y are rendered conditionally independent. Finally, the fourth condition states that when X and Y have a common descendant W, neither W nor any of its descendants are observed, otherwise any observation X may yield information about Y due to the observation of some of their shared descendants. If an undirected path fails the above test, it is said to be *active*, *i.e.*, it allows information to pass between $\boldsymbol{X}$ and $\boldsymbol{Y}$ given $\boldsymbol{Z}$. Therefore, $\boldsymbol{X}$ and $\boldsymbol{Y}$ are d-separated given $\boldsymbol{Z}$ iff there are no active undirected paths between $\boldsymbol{X}$ and $\boldsymbol{Y}$ given the observed nodes $\boldsymbol{Z}$.

It can be shown that the set of d-separations in a BN is equivalent to the set of conditional independence statements defined by almost all probability distributions that factorize in the same manner as specified by the BN [73]. The concept of d-separation leads to the notion of a *Markov blanket*. The Markov blanket of X_i is the minimal set of variables that, when observed, render X_i independent from the remaining variables in the model. This minimal set of variables consists of the parents of X_i, the children of X_i, and the co-parents of the children of X_i, as can be shown either analytically [9] or using the notion of d-separation.

[2]In a directed graph, an undirected path or *trail* between X_i and X_j consists of an acyclic sequence of nodes $(X_i, \ldots, X_j)$, where for each consecutive pair X_i, X_{i+1} there is either a directed edge $X_i \rightarrow X_{i+1}$ or a directed edge $X_{i+1} \rightarrow X_i$.

3.2.2 Inference of Posterior Distributions

In a BN, some of the random variables are clamped to specific values that correspond to a given data set, being known as the *observed* variables, while the remaining random variables are known as *latent* or *hidden* variables [9]. Conceptually, a BN specifies a statistical process that is assumed to give rise to a data set; as such, BNs are commonly referred to as *generative models*. A central task is then to compute the *posterior* distribution $P(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{\Theta})$ of a given set of variables $\boldsymbol{X}$ given another set of variables $\boldsymbol{Y}$, where $\boldsymbol{\Theta}$ is a set of parameters.

The computational complexity of posterior inference depends on the graph structure of the BN that incorporates the assignment $\boldsymbol{Y} = \boldsymbol{y}$. For trees and polytrees [27] the sum-product algorithm can be applied [75], while arbitrary graphs require the more general junction tree algorithm [69]. While the sum-product algorithm has a linear time complexity in the number of nodes, the junction tree algorithm is exponential in the size of the largest clique in the undirected graph obtained by first converting all directed edges in the original BN to undirected edges, then "moralizing" the graph by connecting all nodes that share a common child node, and finally triangulating the graph by connecting all non-neighboring nodes in loops with four or more nodes [67]. Obtaining a triangulated graph with the smallest maximal clique is an NP-hard problem [69]; in fact, exact inference in BNs is NP-hard [73].

Approximate inference methods attempt to provide computationally efficient alternatives to exact inference methods. While several successful approximate inference approaches exist, for instance expectation-propagation [97] or belief propagation [154], here the focus is on the ones used in the papers featured in this thesis, namely Gibbs sampling and mean field variational inference. Both of these frameworks yield well-performing and relatively straightforward inference algorithms, and are currently the most popular approaches for inference in graphical models.

Gibbs Sampling

Gibbs sampling is a Markov chain Monte Carlo (MCMC) approach for obtaining samples from the posterior distribution $P(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{\Theta})$ [41, 118]. Those samples can then be used to compute unbiased estimates of expec-

tations as

$$\mathrm{E}_{\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\Theta}}[f] = \int f(\boldsymbol{X}) P(\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\Theta}) \mathrm{d}\boldsymbol{X} \approx \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{x}_i), \tag{3.12}$$

where $\boldsymbol{x}_i$ is the i-th sample and n is the number of obtained samples [73]. MCMC methods are named as such because the generated samples form a Markov chain that converges to the desired posterior distribution $P(\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\Theta})$ [9]. Gibbs sampling works by considering each variable X_i in turn and sampling from the corresponding conditional posterior distribution $P(X_i|\boldsymbol{X}_{-i},\boldsymbol{Y},\boldsymbol{\Theta})$. The proof that a Gibbs sampler converges to the desired posterior distribution relies on viewing the Gibbs sampler as a particular instance of the Metropolis-Hastings framework [9].

The appeal of the Gibbs sampler as a general-purpose inference tool is that while the global posterior distribution $P(\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\Theta})$ is usually intractable, the conditional distributions $P(X_i|\boldsymbol{X}_{-i},\boldsymbol{Y},\boldsymbol{\Theta})$ are typically easy to sample from in BNs. In practice, the initial samples obtained by the Gibbs sampler are discarded as part of the *burn-in* period, before the sampler reaches its stationary distribution. Also, because successive samples from the Gibbs sampler are correlated, in order to obtain two independent samples from the desired posterior distribution one must use lagged samples.

A variation of the Gibbs sampler which is typically used in conjugate mixture models is the collapsed Gibbs sampler [87]. This approach consists of applying a Gibbs sampler to a posterior probability distribution in which some of the model variables have been integrated out,

$$P(\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{\Theta}) = \int P(\boldsymbol{X},\boldsymbol{U}|\boldsymbol{Y},\boldsymbol{\Theta}) \mathrm{d}\boldsymbol{U}, \tag{3.13}$$

where the integrated-out variables $\boldsymbol{U}$ are known as the *nuisance* variables. These variables may effectively correspond to variables which are not of interest for the subsequent analysis, or they can alternatively be integrated out and later estimated using the obtained Gibbs samples [51]. The advantage of collapsed Gibbs samplers with respect to the classical Gibbs sampler is that they typically converge faster to a local mode, although the integration induces correlations between the sampled variables, which in turn may deter the sampler from moving between different modes and may also induce a computationally more complex sampling process. Collapsed Gibbs samplers were used in all publications except for Publication IV.

Mean Field Variational Inference

Variational inference methods [67, 149] are methodologically different from MCMC methods. In variational inference, one considers a family of *variational posterior distributions* with a simpler parametric form than the original, intractable posterior. The variational posterior is indexed by a set of *variational parameters*, and the aim is to find the configuration of parameters that brings the variational posterior distribution "closest" to the intractable posterior. While variational inference methods generally lack the same theoretical guarantees of convergence as MCMC methods, their deterministic approximation setup yields optimization problems that are faster to solve and whose convergence is easier to monitor.

Formally, the setup consists of approximating $P(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{\Theta})$ by a variational distribution $Q(\boldsymbol{X}; \boldsymbol{\lambda})$ parameterized by the variational parameters $\boldsymbol{\lambda}$. Here, the approximation is obtained by minimizing the Kullback-Leibler divergence $\mathrm{D}_{\mathrm{KL}}(Q||P_{\mathrm{post}})$, where P_{post} designates the posterior distribution $P(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{\Theta})$. However, it is possible to minimize other α-divergence measures, yielding alternative approximate inference schemes [98]. Minimizing $\mathrm{D}_{\mathrm{KL}}(Q||P_{\mathrm{post}})$ is equivalent to optimizing a lower bound on the marginal log-probability of the observed nodes $\boldsymbol{Y}$, as can be seen in the following derivation:

$$\begin{aligned}
\log P(\boldsymbol{Y}|\boldsymbol{\Theta}) &= \log \int P(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{\Theta}) \mathrm{d}\boldsymbol{X} \\
&= \log \int P(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{\Theta}) \frac{Q(\boldsymbol{X})}{Q(\boldsymbol{X})} \mathrm{d}\boldsymbol{X} \\
&\geq \int \left(\log P(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{\Theta}) - \log Q(\boldsymbol{X})\right) Q(\boldsymbol{X}) \mathrm{d}\boldsymbol{X} \\
&= \mathrm{H}\left[Q\right] + \mathbf{E}_Q\left[\log P(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{\Theta})\right], \qquad (3.14)
\end{aligned}$$

where the inequality is obtained by making use of Jensen's inequality [18] and H designates the entropy functional. It is straightforward to observe that the difference between $\log P(\boldsymbol{Y}|\boldsymbol{\Theta})$ and the lower bound is precisely $\mathrm{D}_{\mathrm{KL}}(Q||P_{\mathrm{post}})$. Thus, minimizing $\mathrm{D}_{\mathrm{KL}}(Q||P_{\mathrm{post}})$ is equivalent to maximizing the lower bound, because they sum to the constant term $\log P(\boldsymbol{Y}|\boldsymbol{\Theta})$.

A main challenge is how to choose a variational distribution that both yields a feasible optimization problem and leads to a satisfactory solution. The *mean field* approach consists of choosing a fully factorized distribution

$$Q(\boldsymbol{X}; \boldsymbol{\lambda}) = \prod_i Q_i(X_i; \lambda_i). \qquad (3.15)$$

It can be shown that the resulting problem yields an iterative optimization procedure, where each λ_i is optimized in turn, given the remaining

$\boldsymbol{\lambda}_{j \neq i}$, but requiring only the parameters corresponding to the variables that belong to the Markov blanket of X_i [11]. In certain classes of models that include log-normalization terms, such as models with logistic or multinomial regression terms, the resulting optimization problem is still infeasible. In those cases, an additional lower bound on the problematic terms must be created, indexed by an additional set of variational parameters [10, 13, 62]. This was done in Publication IV for a multinomial regression term, using the same approach as Blei and Lafferty [13].

3.2.3 Point Estimation and the Expectation-Maximization Algorithm

When the aim is to obtain a point estimate for $\boldsymbol{\Theta}$, the standard approach is to compute the $\boldsymbol{\Theta}$ that maximizes either the log-probability of the observed nodes $\log P(\boldsymbol{Y}|\boldsymbol{\Theta})$ or the log-posterior probability of $\boldsymbol{\Theta}$,

$$\log P(\boldsymbol{\Theta}|\boldsymbol{Y}) = \log P(\boldsymbol{\Theta}) + \log P(\boldsymbol{Y}|\boldsymbol{\Theta}) - \log P(\boldsymbol{Y}). \tag{3.16}$$

The $\boldsymbol{\Theta}$ that results from maximizing each of these criteria is known as the maximum likelihood (ML) solution or the maximum *a posteriori* (MAP) solution, respectively. For succinctness, only the ML estimation case is described.

Obtaining a ML solution is usually a hard optimization task that contains multiple local optima, due to the non-trivial variable dependency structure stipulated in a given BN. Assuming that the objective function is continuous and differentiable with respect to $\boldsymbol{\Theta}$, gradient-based optimization techniques such as gradient descent or quasi-Newton methods can be used [4]. However, the most common approach is to apply a general iterative scheme known as the Expectation-Maximization (EM) algorithm [30].

In the EM literature, the log-likelihood of $\boldsymbol{\Theta}$, $\log P(\boldsymbol{Y}|\boldsymbol{\Theta})$, is known as the incomplete data log-likelihood, due to the unobserved status of the variables $\boldsymbol{X}$, which are here designated as $\boldsymbol{Z}$ to use standard notation. The EM algorithm alternates between *expectation* and *maximization* steps (E-step and M-step, respectively). The E-step consists of computing the expectation of $\log P(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{\Theta})$ over the posterior distribution of $\boldsymbol{Z}$ given $\boldsymbol{Y}$ and the current estimate of $\boldsymbol{\Theta}$, designated as $\boldsymbol{\theta}_{\text{old}}$. The log-joint probability $\log P(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{\Theta})$ is known as the complete data log-likelihood. This

expectation is computed as a function f of $\boldsymbol{\theta}$,[3]

$$f(\boldsymbol{\theta}; \boldsymbol{\theta}_{\mathbf{old}}) = \mathrm{E}_{\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{\Theta}=\boldsymbol{\theta}_{\mathbf{old}}} \left[\log P(\boldsymbol{Z}, \boldsymbol{Y}|\boldsymbol{\Theta} = \boldsymbol{\theta})\right]. \tag{3.17}$$

In the M-step, a new estimate of θ is obtained by maximizing f. It can be shown that each round of the EM algorithm increases the log-likelihood of $\boldsymbol{\Theta}$ [9].

The EM algorithm can be understood in terms of optimizing a lower bound on the log-likelihood of $\boldsymbol{\Theta}$ [104]. Using the inequality in (3.14),

$$\log P(\boldsymbol{Y}|\boldsymbol{\Theta}) \geq \mathrm{E}_Q \left[\log P(\boldsymbol{Y}, \boldsymbol{Z}|\boldsymbol{\Theta})\right] + \mathrm{H}\left[Q\right], \tag{3.18}$$

where Q is any distribution over $\boldsymbol{Z}$. The difference between the left-hand side and the right-hand side is $\mathrm{D}_{\mathrm{KL}}(Q||P)$, where P is the posterior distribution of $\boldsymbol{Z}$. Therefore, maximizing the lower bound amounts to setting $Q(\boldsymbol{Z}) = P(\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{\Theta})$. This is equivalent to forming the function f in the E-step. When obtaining the exact solution $Q(\boldsymbol{Z}) = P(\boldsymbol{Z}|\boldsymbol{Y}, \boldsymbol{\Theta})$ is infeasible, an approximate variational solution for Q can be computed. When a variational Q is used, the resulting EM method is known as a *variational* EM. The M-step is clearly equivalent to maximizing the lower bound, as the term $\mathrm{H}\left[Q\right]$ does not depend on $\boldsymbol{\Theta}$. This information-theoretic perspective allows conceptualizing the EM algorithm as a coordinate ascent approach, where Q and $\boldsymbol{\Theta}$ are alternatively optimized [61], and also allows understanding both standard and variational EM approaches under the same framework.

The EM algorithm is ubiquitously used in graphical modelling, and several classical methods such as K-means or the Baum-Welch method can be seen as instances of this algorithm [9].

3.2.4 Recent Bayesian Nonparametric Methods

A recent area of research in graphical models has been the use of nonparametric Bayesian statistics methods as model selection tools. Specifically, in unsupervised learning models that assume the existence of latent combinatorial structures, *e.g.*, partitions, trees, or binary feature vectors, these nonparametric methods provide prior distributions over those latent structures.

[3] In the EM literature, f is often designated as Q. That terminology is avoided to distinguish f from the variational posterior distribution Q.

The Dirichlet Process

The theoretic basis for several nonparametric priors is the Dirichlet Process (DP) [38]. The DP is a distribution on probability measures over a measurable space Ψ, *i.e.*, each draw from a DP is a probability measure on Ψ. A DP is characterized by a concentration parameter α_0 and a base probability measure G_0. Intuitively, G_0 stipulates the expected mass assigned to any measurable subspace of Ψ, while α_0 governs the variance. Formally, if $G \sim \mathrm{DP}(\alpha_0, G_0)$, then for any finite measurable partition of Ψ, designated by $(A_1, \ldots, A_n)$, the random vector $(G(A_1), \ldots, G(A_n))$, *i.e.*, the mass assigned by the random variable G to each component of the partition of Ψ, follows a Dirichlet distribution

$$(G(A_1), \ldots, G(A_n)) \sim \mathrm{Dir}(\alpha_0 G_0(A_1), \ldots, \alpha_0 G_0(A_n)) \tag{3.19}$$

[138]. As described above, each draw from a DP is a probability measure on Ψ. It has been shown that the drawn probability measures are discrete with probability one, *i.e.*, almost all probability measures drawn from a DP are probability measures over the positive natural numbers [38]. This is the property that enables the DP to be used as a prior distribution on class assignments.

In practice, there exist alternative descriptions of the DP that make it more amenable to use in the context of BNs, namely the stick-breaking [125] and Chinese restaurant process (CRP) [12] formulations. The stick-breaking construction for the DP defines a random variable $G \sim \mathrm{DP}(\alpha_0, G_0)$ as

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}, \tag{3.20}$$

where each $\phi_k | G_0 \sim G_0$ is a draw from G_0, δ_{ϕ_k} is an atomic probability measure that concentrates its mass on ϕ_k, $\pi_k = v_k \prod_{i<k}(1 - v_i)$, and $v_i \sim \mathrm{Beta}(1, \alpha_0)$. Intuitively, this constructive definition of the DP involves sampling a countable number of atoms δ_{ϕ_k} from G_0 and assigning to each atom a probability π_k that vanishes as k increases, in a manner given by the product of terms involving the variables v_k; this product is intuitively akin to a stick-breaking process, hence the name of this alternative description of the DP. The sequence $\boldsymbol{\pi} = (\pi_k)_{k=1}^{\infty}$ is said to follow the Griffiths, Engen, and McCloskey (GEM) distribution, $\boldsymbol{\pi} \sim \mathrm{GEM}(\alpha_0)$ [113].

Another alternative description of the DP is as a Polya urn scheme [12]. The Polya urn scheme describes the joint distribution of a sequence of independent draws from $G \sim \mathrm{DP}(\alpha_0, G_0)$, where G has been integrated out.

The conditional probability of the i-th variable g_i given the DP parameters α_0 and G_0, as well as the previous draws $g_1, \ldots, g_{i-1}$, is given by

$$g_i | g_1, \ldots, g_{i-1}, \alpha_0, G_0 \sim \sum_{k=1}^{i-1} \frac{1}{i - 1 + \alpha_0} \delta_{g_k} + \frac{\alpha_0}{i - 1 + \alpha_0} G_0, \tag{3.21}$$

i.e., g_i has a non-zero probability of being exactly equal to one of the previous draws, while also having a non-zero probability of constituting a novel draw from G_0. This process can be understood intuitively as an urn model. Each ball i of unit size in the urn corresponds to a draw from G, with the colour of the ball corresponding to the atom δ_{g_i}. There is also a special ball with size α_0 that indicates a novel atom. Obtaining a new sample corresponds to drawing a ball from the urn. If the special ball is retrieved, then a new atom and corresponding colour are created, and a ball of unit size with that colour is inserted in the urn, along with the special ball. If, instead, a ball with an existing colour is retrieved, then a new ball with that colour is created and both balls are inserted back in the urn.

The Polya urn scheme can be equivalently described using a gastronomical metaphor known as the Chinese restaurant process [113]. Assuming a Chinese restaurant with an unbounded number of tables and where each table can hold as many clients as needed, when a new client arrives at the restaurant he picks a table with probability proportional to the number of clients already sitting at the table; with probability proportional to α_0, he chooses to sit at a new table.

Two properties stem from the above descriptions of the DP: First, there is a clustering property, in the sense that if a table contains a relatively high number of clients, then there is a high probability that the following client will choose that same table/atom. Second, the distribution of table/atom assignments is exchangeable, *i.e.*, the probability of a collection of table assignments is the same regardless of the order in which the clients are assigned to tables. While the former property constitutes a *desideratum* of a prior distribution over class assignments, the latter property is useful in the development of inference engines for models that incorporate the DP.

The DP has been successfully used in models where objects are partitioned into classes, for instance in mixture models [43, 115, 138]. In fact, the DP can be obtained in the context of a mixture model by use of a limit argument [115]. Both MCMC and variational inference methods have been developed for the DP [16, 60, 105], and both sampling and es-

timation procedures for α_0 exist [14, 36]. Several extensions of the DP are already in place. For instance, while the original DP yields an exponential decay in the number of objects assigned to each cluster, a more general two-parameter process known as the Pitman-Yor process yields a power-law distribution that has been shown to be more adequate in the natural language processing domain [135]. Another extension is the hierarchical DP (HDP), which allows for mixture components to be shared across different data groups [138].

Among the various extensions of the DP is the nested Chinese restaurant process (nCRP) [14]. The nCRP is essentially a recursive version of the CRP that yields a probability distribution over trees. Succinctly, the nCRP starts by running a standard CRP procedure. Then, each cluster obtained via the standard CRP is used as the basis for running another CRP; this recursion continues either indefinitely or until a maximum predefined depth has been reached, as shown in Figure 3.2. In Publication III the nCRP was used as a tree prior, in a model where microarray samples are assigned to leaf nodes in the tree, thus yielding a hierarchical decomposition of a microarray data set.

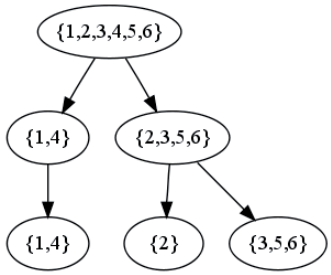


Figure 3.2. Running the nCRP for six objects and two recursion levels. Figure taken from Publication III.

The nCRP is conceptually distinct from the HDP model. The HDP uses a two-level hierarchy of DPs that makes different DPs share the same atoms, allowing different data groups to share the same components within a mixture modelling framework; this process can be described via a gastronomical metaphor known as the Chinese restaurant franchise [138]. On the other hand, in the nCRP the aim is not to share all atoms drawn from different DPs, but rather to partition them into different groups. In general, hierarchical and nesting strategies constitute two different modeling concepts in nonparametric Bayesian modelling [66].

The Indian Buffet Process

An approach akin to the CRP, known as the *Indian buffet process* (IBP), has been proposed for the case when objects, rather than being assigned to mutually exclusive classes, contain each an instantiation of a binary feature vector [45]. Here, the aim is to directly learn the number of features from the data. The IBP has been originally proposed as the limit $K \to \infty$ to the following generative process for a binary matrix $\boldsymbol{Z}$ with n rows and K columns:

$$\pi_k|\alpha \sim \text{Beta}\left(\frac{\alpha}{K}, 1\right), \tag{3.22}$$

$$z_{ik}|\pi_k \sim \text{Bernoulli}(\pi_k), \tag{3.23}$$

where $\alpha > 0$ is a pre-specified parameter. In the limit $K \to \infty$ and after integrating each π_k out of the model, the marginal distribution of $\boldsymbol{Z}$ can be described via the following stochastic process: In an Indian buffet, n clients decide which of a countable number of dishes to try. With probability m_k/i, the i-th client tries a given dish k that has been tried by a previous client, where $m_k \geq 1$ is the number of previous clients that have already tried dish k. Then, client i tries a set of new dishes, with the number of new dishes following a Poisson(α/i) distribution.

The IBP, like the DP, enjoys clustering and exchangeability properties. In addition, any matrix $\boldsymbol{Z}$ generated via the IBP is typically sparse, in the sense that the expected number of ones in $\boldsymbol{Z}$ can be shown to be equal to $n\alpha$.

Two-parameter and power-law variations of the IBP have been proposed [45, 136], and both Gibbs sampling and variational inference methods exist for models that involve the IBP [32]. Finally, both stick-breaking [137] and Beta process formulations [139] have been derived for the IBP. In Publication IV, the finite version of the IBP was used as a prior for bicluster membership.

3.2.5 Model-Based Relevance Measures for Information Retrieval

Often, it is of interest to perform information retrieval (IR) on data which is modelled by a BN. Formally, assume a number of objects, with each object r corresponding to the observed data $\boldsymbol{y}_r$. An object may be for instance a gene associated with a number of expression measurements. Here, the aim is to compute the relevance of each object r to a given query object q corresponding to the observed data $\boldsymbol{y}_q$.

While multivariate data points can be easily related via standard similarity measures such as Spearman correlation, an alternative is to use the BN to perform IR. A formulation originally proposed for the natural language domain is to relate q and r based on how well $\boldsymbol{y}_q$ is modelled by the same variables used to model $\boldsymbol{y}_r$ [21, 128]. Formally, this corresponds to the following definition:

$$rel(q,r) \stackrel{\text{def}}{=} \int_{\boldsymbol{\Theta}} P(\boldsymbol{y}_q|\boldsymbol{\Theta}_q = \boldsymbol{\theta}_r) P(\boldsymbol{\Theta}|\boldsymbol{Y}) d\boldsymbol{\Theta}, \tag{3.24}$$

where $\boldsymbol{\Theta}$ are the model variables and $P(\boldsymbol{y}_q|\boldsymbol{\Theta}_q = \boldsymbol{\theta}_r)$ is defined as the probability of $\boldsymbol{y}_q$ given that the model variables associated with q are equal to the variables associated with r. This approach depends on the ability to map objects to subsets of model variables. For instance, in mixture models such as LDA, each object r is associated with a probability over components θ_r. In this setting, a natural choice for performing model-based IR is to consider the probability of the query data $\boldsymbol{y}_q$ given that $\theta_q = \theta_r$.

Using a model-based relevance metric has the potential advantage that, as a model is used to infer the relevance between data points, the model structures can be used to interpret the retrieval results. It is also a flexible tool in the sense that the same IR principle can be applied to different model families. When a Gibbs sampler is used, (3.24) can be approximated by a mean over samples, while for variational inference engines the posterior probability $P(\boldsymbol{\Theta}|\boldsymbol{Y})$ can be replaced by a variational approximation.

The model-based relevance measure described above was used in Publications II and V, as well as in Publication III, in the context of information retrieval and biclustering, respectively.

4. Biclustering of Gene Expression Data

This chapter presents the contributions on biclustering of gene expression data. It starts with a brief motivation and review of existing work, followed by a description of specific contributions, namely a Bayesian plaid model (Publication I), a hierarchical biclustering method (Publication III), and a general probabilistic biclustering framework (Publication IV).

4.1 Motivation and Earlier Work

Biclustering, also commonly known as co-clustering, is a generalization of clustering in which multidimensional objects are grouped, with each group or *bicluster* being restricted not only to a subset of objects but also to a subset of the objects' features [24, 25, 55, 91, 114]. Biclustering methods aim at capturing local rather than global similarities between objects. For instance, in the context of gene expression, a group of patients may exhibit similar expression patterns for disease-related genes while being dissimilar for the remaining genes, or a family of drugs may yield consistent expression only among genes related to the drugs' common targets.

The overall computational complexity of biclustering can be analyzed via a graph representation of a data matrix. A data matrix can be regarded as a weighted bipartite graph, *i.e.*, a graph in which nodes are partitioned into two sets V_A and V_B, and where all weighted edges (u, v) verify the properties $u \in V_A$ and $v \in V_B$. To see this, set V_A to the set of objects and V_B to the set of conditions in the data matrix. Then, set the weight of each edge (u, v) between object u and condition v to the value of the corresponding entry in the data matrix. Using this graph representation, the problem of finding the largest bicluster of ones in a binary data matrix can be regarded as one of finding the maximum edge biclique in

a bipartite graph, which is an NP-complete problem [91, 111]. Therefore, although the computational complexity depends on the specific biclustering problem at hand, it is expected that most relevant variants of biclustering problems are only approximately solvable. Biclustering can also be related to the classical data mining task of finding frequent itemsets [124].

The first known biclustering algorithm is the direct clustering method proposed by Hartigan [55]. Biclustering has been originally proposed as a meaningful task for gene expression data in several independent studies [25, 44, 142]. A taxonomy for biclustering methods can be derived by considering either the kind of bicluster structures that are obtained by each method, *e.g.*, if the method allows biclusters to overlap, or by considering each method's underlying assumptions [91]. Several methods attempt to find homogeneous blocks in a data matrix (*e.g.*, [25, 96, 142]); other approaches infer a two-way coupling of one-way clustering methods (*e.g.*, [44]); the bipartite graph representation of a data matrix has also been successfully used as a basis for biclustering [133, 134]; linear models are also a common tool for biclustering [82], having the advantage that when coupled with categorical regression frameworks based on link functions, yield methods that are suitable for both continuous and discretized data sets [95]. Other successful approaches include probabilistic multiplicative models [56] and Chernoff bound-based methods for finding dense biclusters in sparse binary data matrices [145]. Finally, probabilistic frameworks such as LDA or the HDP have also been adapted to the biclustering case [39, 43].

4.2 Bayesian Biclustering with the Plaid Model

The plaid model [82] is arguably one of the most general biclustering methods. In the plaid model, biclusters may correspond to any subset of objects and conditions, and may also freely overlap. Biclusters are represented by two binary matrices $\boldsymbol{\rho}$ and $\boldsymbol{\kappa}$ of dimensions $N \times K$ and $M \times K$, respectively, where N is the number of objects, M is the number of conditions, and K is the number of biclusters.[1] Each bicluster k corresponds to an Analysis of Variance (ANOVA) model with parameters μ_k, $\boldsymbol{\alpha}_k$, and

[1]Here, the use of upper case letters for representing matrices is avoided in order to maintain coherence with the notation used in Publication I.

$\boldsymbol{\beta}_k$, where μ_k is a scalar, $\boldsymbol{\alpha}_k$ is a vector of length N, and $\boldsymbol{\beta}_k$ is a vector of length M. Assuming the existence of a bias term μ_0, the expression level for the object-condition pair (i, j) is given by

$$Y_{ij} = \mu_0 + \sum_{k=1}^{K} (\mu_k + \alpha_{ik} + \beta_{jk}) \rho_{ik} \kappa_{jk}, \tag{4.1}$$

i.e., the plaid model consists of an additive combination of biclusters or layers. As in standard ANOVA, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are intended to be interpreted as departures from the mean parameter μ, and so obey the following constraints for every bicluster k:

$$\sum_{i=1}^{N} \rho_{ik} \alpha_{ik} = 0, \tag{4.2}$$

$$\sum_{j=1}^{M} \kappa_{jk} \beta_{jk} = 0. \tag{4.3}$$

In the original plaid model, the parameters are inferred using a greedy, heuristic fitting procedure that aims at minimizing the quadratic error between the model and the observed data matrix, thus assuming a Gaussian noise model. It is therefore of interest to investigate if the model is amenable to more standard optimization or inference approaches.

In Publication I, the main contribution was to endow the plaid model with a Bayesian framework and develop a corresponding collapsed Gibbs sampling method. A connection between the plaid model and two other existing methods was also established [7, 95]. Both $\boldsymbol{\rho}_k$ and $\boldsymbol{\kappa}_k$ were assumed to consist of a set of beta-binomial models,

$$\pi_k \sim \mathrm{Beta}(\delta_\rho^{(k)}, \gamma_\rho^{(k)}), \tag{4.4}$$

$$\lambda_k \sim \mathrm{Beta}(\delta_\kappa^{(k)}, \gamma_\kappa^{(k)}), \tag{4.5}$$

$$\rho_{ik} | \pi_k \sim \mathrm{Bernoulli}(\pi_k), \tag{4.6}$$

$$\kappa_{jk} | \lambda_k \sim \mathrm{Bernoulli}(\lambda_k). \tag{4.7}$$

For simplicity, the parameters μ_k were removed and the constraints (4.2) and (4.3) were eliminated. In effect, keeping the constraints yields positive semidefinite covariance matrices that make the resulting sampler unnecessarily complex [126]. The proposed model does not result in decreased identifiability, although the resulting bicluster parameters have a necessarily different interpretation from the parameters in the original plaid model, as they already incorporate the original mean parameter μ_k. A homoscedastic noise model for the data was also assumed,

$$Y_{ij} | \boldsymbol{\alpha}_i, \boldsymbol{\beta}_j, \boldsymbol{\rho}_i, \boldsymbol{\kappa}_j, \sigma^2 \sim \mathrm{N}\left(\mu_0 + \sum_{k=1}^{K} (\alpha_{ik} + \beta_{jk}) \rho_{ik} \kappa_{jk}, \sigma^2 \right), \tag{4.8}$$

with μ_0, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ having zero-mean Gaussian priors that share the same scalar variance σ^2. The assumption of proportional noise in the data and parameters is a commonly used assumption in Bayesian analysis that facilitates integrating out some of the model variables [40]. Finally, a scaled inverse-chi-square distribution for σ^2 was assumed. The specific selection of conjugate distributions allows integrating out μ_0, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and σ^2 from the model. The proposed collapsed Gibbs sampler aims at obtaining samples from the posterior distribution $P(\boldsymbol{\rho}, \boldsymbol{\kappa}|\boldsymbol{Y})$.

In order to validate the proposed approach, its performance was compared both against a random baseline and a standard hierarchical clustering method in a small proof-of-concept study involving human gene expression data from across a wide range of tissues [130]. In particular, the methods were compared on four independent data subsets corresponding to four Gene Ontology (GO) [5, 72] gene sets related to rhythmic processes, regulation of biosynthetic processes, growth regulation, and cell division. It was first assessed if, for each bicluster, pairs of conditions are significantly more similar when considering only genes included in the same bicluster rather than all the genes in the data, with similarity being measured via Pearson correlation and significance being assessed by random permutations. In all four GO categories the average correlation gain is significant ($p < 0.05$), which shows that the genes assigned to a bicluster typically increase the correlation between conditions assigned to the same bicluster. A classification-based test was also used to show that in three out of four gene sets the method outperforms hierarchical clustering with respect to finding biclusters whose genes are functionally homogeneous.

The main limitation of the proposed approach is that it does not scale up to data matrices with thousands of objects and conditions. Due to its linear model formulation, sampling each variable involves performing matrix inversion. While using the Sherman-Morrison-Woodbury formula [47] allows for faster sampling, this is still not enough. Future work may involve using faster inference engines, *e.g.*, mean field variational inference. Alternatively, Gu and Liu [52] devised a similar collapsed Gibbs sampler with the added constraint that biclusters cannot overlap, which facilitates and accelerates the sampling process.

It was thus shown that the plaid model is amenable to a Bayesian analysis, although the inference engine must still be adapted in order to scale up to thousands of objects and conditions.

4.3 Hierarchical Generative Biclustering

Publication III focussed on investigating the use of nonparametric Bayesian priors, namely the nCRP, for hierarchically organizing a microarray data set. Concretely, the problem of interpreting clustering results was considered. Clustering methods, although being ubiquitously used in biological studies, often provide results that are hard to interpret or translate into biological findings [31, 34]. The aim was to additionally indicate for each cluster which genes tie the clustered conditions together. This yields a biclustering formulation, since each (bi)cluster is now associated not only with a subset of conditions but also with a subset of genes. Additionally, the biclusters were intended to follow a hierarchical organization.

Assuming a tree of biclusters, each microarray maps to a leaf node in the tree, belonging to all biclusters on the unique path from the root to that leaf node. The intuition is to generate *high-level* biclusters, corresponding to nodes closer to the root, to which many samples belong, representing expression patterns that are common across conditions but that typically involve few genes; at the same time, nodes closer to leaf nodes represent specific expression patterns corresponding to fewer microarrays but a larger number of genes. In order to do so, the model assumes that a gene which belongs to a bicluster associated with node u also belongs to all biclusters corresponding to nodes that are descendents of node u.

The main contribution in Publication III was therefore to provide the first hierarchical biclustering model that allows one to simultaneously infer a tree structure where conditions map to leaf nodes, and at the same time explicitly indicate for each node in the tree which genes exhibit homogeneous expression for the conditions under the scope of that node. The model is also the first one to use the nCRP for modelling continuous data. A preliminary study about the applicability of model-based relevance measures to conduct information retrieval using the proposed model was also conducted.

The generative process for the model is divided into three parts: First, a tree structure is created, with nodes representing biclusters, and microarray samples are assigned to leaf nodes in that hierarchy. Second, genes are placed along multiple nodes in the hierarchy, with the placement of gene g in node u meaning that conditions associated with u have homogeneous expression for g. Third, the expression data is assumed to be generated using parameters associated with each node/bicluster.

In order to create a tree structure and assign samples to leaf nodes in the hierarchy, the nCRP was used, as described in the previous chapter. Although the nCRP has been recently extended to handle infinite-depth trees, for simplicity the original finite-depth formulation of the nCRP [15] was used. In the second part of our model, genes were represented as binary features that are originally inactive at the root, that may switch to one along any given directed path, and that do not switch back to zero. The model assumes the existence of *edge length* variables

$$l_{(u,v)} \sim \text{Beta}(\alpha = 1, \beta = 1) \tag{4.9}$$

for every edge (u, v). Then, for every gene j in node v with parent node u, the following activation model is assumed:

$$P(z_{jv} = 1 | z_{ju} = 0, l_{(u,v)}) = l_{(u,v)}, \tag{4.10}$$

$$P(z_{jv} = 1 | z_{ju} = 1, l_{(u,v)}) = 1, \tag{4.11}$$

where z_{jv} is the latent variable that asserts if gene j is active at node v. This activation model guarantees a one-way switching process for the binary latent variables, wherein latent variables may switch from zero to one but cannot switch back. Intuitively, each edge length variable $l_{(u,v)}$ models the fraction of inactive features at node u that are expected to become active in the child node v.

In the third part of the model, it is assumed that for every gene j each node/bicluster u is associated with mean and variance parameters μ_{ju} and σ^2_{ju}, with Gaussian and inverse-Gamma priors, respectively. Assuming that gene j switches to one at node u, the model assumes the following distribution for the expression data of the samples under the subtree that has u as its root, considering only gene j:

$$\boldsymbol{Y}_{jS_u} \sim \text{N}(\mu_{ju}\mathbf{1}, \sigma^2_{ju}\boldsymbol{I}), \tag{4.12}$$

where S_u is the set of samples under node u. This specification highlights the central idea of the model: When a gene switches to one at a given node, the samples under that node are assumed to have a similar expression for that gene. Each node/bicluster is therefore a refinement of its parent node/bicluster, involving less samples but being homogeneous for a potentially larger group of genes. Finally, it is also assumed that when a gene is still zero at a leaf node, then the corresponding data points are independently generated from a baseline standard Gaussian distribution.

The inference engine is based on a collapsed Gibbs sampler, where the edge length, mean, and variance parameters are integrated out. An auxiliary variable scheme to sample the nCRP hyperparameter based on similar schemes originally developed for the Dirichlet process is also used [36, 138]. Finally, the model-based relevance measure described in the previous chapter is used to relate samples.

The proposed method was applied to a large miRNA expression data set profiling human miRNAs across a wide panel of tissues and cell lines [90]. It was shown that the method outperforms other biclustering approaches with respect to GO and tissue class enrichment measures. Furthermore, the inferred tree effectively separates samples from different tissues, and the hierarchical organization allows for the inference of a bicluster of leukemia samples, which is then partitioned into two biclusters that separate leukemia cell lines from leukemic tissue. A case study that highlights how the model's added interpretability can be exploited suggested that miR-224 may have a role in the known association between melanoma and non-Hodgkin lymphoma [84].

Finally, regarding information retrieval, two aspects were tested: First, if the tree structure learned by the model is useful for relating samples, and second, if the generative relevance measure described in the previous chapter yields a strong performance. In order to do so, three methods for relating samples were considered: the inverse of the Euclidean distance, the generative relevance measure, and a simple heuristic measure that counts the number of Gibbs sampler runs in which two samples fall under the same node in the posterior mode. Using a two-class retrieval task in which the aim is to retrieve samples from the same tissue, and evaluating the performance with the standard area under the receiver-operating curve measure [94], a maximal performance was obtained when using the heuristic tree measure, with the inverse Euclidean distance performing better than the generative relevance measure. This suggests that while the structures learned by the model are indeed useful, the generative relevance measure does not appear to effectively leverage those structures for relating samples. Future work may involve using alternative IR measures such as the area under the precision-recall curve, which may be more adequate due to the class imbalance intrinsic to the tissue retrieval task.

4.4 A Mixture-of-Experts Approach to Biclustering

The aim in Publication IV was to develop a general probabilistic biclustering framework that could be easily adapted to varying assumptions and data types. The motivation is based on the fact that existing methods such as the plaid model, despite being flexible in the sense of allowing arbitrary, overlapping bicluster structures, suffer from two main drawbacks: First, the assumption that parameters from different biclusters combine additively is restrictive; second, forcing the practitioner to specify a linear model may yield models that contain artificial and complex assumptions purely for the purpose of model soundness. An alternative leading to more straightforward approaches is to consider that biclusters overlap when they are able to provide roughly equally good models for the corresponding data points. The proposed framework, which is similar to mixture-of-experts models [10, 68], incorporates this notion.

Mixture-of-experts models are applied to data sets consisting of input-output pairs $(\boldsymbol{x}_i, y_i)$, *e.g.*, regression or classification data sets. These models assume that there are K components or *experts*, with each expert being a simple model that is adequate only for a subset of the data. The main idea behind mixture-of-experts models is that the input $\boldsymbol{x}_i$ probabilistically determines which expert is used to model the output y_i. As for the proposed method, it takes as input a data matrix $\boldsymbol{Y}_{N\times M}$, which has the input-output triplet form (i, j, y_{ij}), $i \in \{1, \ldots, N\}$, $j \in \{1, \ldots, M\}$, *i.e.*, every data point includes row and column indices, which are conceptually the input, and the value y_{ij}, which is conceptually the output. The model also assumes the existence of K experts or biclusters, each corresponding to a simple model of the data. Each expert specifies a *region of expertise*, which indicates a data submatrix for which the model is adequate. The choice of expert for each data point is probabilistically determined by membership of the data point to each expert's region of expertise, *i.e.*, it is probabilistically determined by the row and column indices.

The generative process for the model shares some similarities both with the plaid model and overall with mixture models. First, two binary matrices $\boldsymbol{U}_{N\times K}$ and $\boldsymbol{V}_{M\times K}$ define the expertise regions, akin to the binary matrices defined in the plaid model. Each of these matrices is generated via the finite approximation for the IBP, as described in (3.22) and (3.23). Then, it is assumed that each data point chooses a single bicluster, as is standard practice in mixture models. In order to model the notion that the

expertise regions influence the bicluster choice, the following multinomial regression formulation is used:

$$P(z_{ij} = k|\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{\beta}) = \frac{\exp\{u_{ik}v_{jk}\beta_{k1} + (1 - u_{ik}v_{jk})\beta_{k0}\}}{\sum_{k'} \exp\{u_{ik'}v_{jk'}\beta_{k'1} + (1 - u_{ik'}v_{jk'})\beta_{k'0}\}} \quad (4.13)$$

where β_{k0} and β_{k1} are additional model variables. The multinomial regression formulation has the following interpretation: If (i, j) falls within the expertise region of bicluster k, then $u_{ik}v_{jk} = 1$, making the probability of choosing bicluster k proportional to $e^{\beta_{k1}}$. Otherwise, $u_{ik}v_{jk} = 0$ or equivalently $1 - u_{ik}v_{jk} = 1$, making the probability of choosing bicluster k proportional to $e^{\beta_{k0}}$. It is expected that for each bicluster k, β_{k1} is a high value and β_{k0} is a low value, indicating that bicluster k provides a good model for its region of expertise but a bad model for the rest of the data set. Finally, each bicluster is associated with a parameter set $\boldsymbol{\theta}_k$. In the experiments, Gaussian experts are used for continuous data, *i.e.*, $\boldsymbol{\theta}_k = (\mu_k, \sigma_k^2)$, and Bernoulli experts for binary data, *i.e.*, $\boldsymbol{\theta}_k = p_k$. The flexibility of the proposed model relies in the separation between the process of choosing an expert and the model provided by each expert. Although each data point chooses a single bicluster, the regions of expertise may freely overlap, and uncertainty over which bicluster is chosen by a data point reflects the ability of each expert to model that data point.

To perform point estimation and inference, a variational EM algorithm was used. In order to handle the problematic log-sum stemming from the denominator of the multinomial regression formulation, a second-order lower bound was created, using the same approach as Blei and Lafferty [13]. Recently, another type of lower bound to a log-sum term has been proposed [17]. It remains an open question whether it performs better than the currently used approach. All technical details are provided in Publication IV.

The method was applied to a continuous miRNA data set also analyzed in Publication III and to a binary copy number variation data set compiled from several independent studies [102]. In both data sets the method performs well with respect to other biclustering methods, which shows that the approach is robust with respect to different data types and model assumptions. While for the miRNA data set the method partitions the data similarly to the method in Publication III, in the copy number variation data set the inferred biclusters typically confirm existing biological knowledge regarding the association between neoplasm types and chromosomal aberrations. The proposed framework thus appears to be a promising general probabilistic solution to biclustering.

4.5 Discussion

This chapter described the contributions about biclustering of gene expression data.

Regarding Publication I and the plaid model, it remains an open question whether alternative inference engines would allow maintaining the model's flexible assumptions and at the same time accelerate the inference process, allowing the model to be used in larger data sets.

As for the hierarchical model described in Publication III, the most immediate improvement on the model would be the use of unlimited-depth rather than fixed-depth trees. This would also allow some parts of the tree to have a deeper depth than others. As can be seen in Publication III, Figure 4, when the model does not find a hierarchical bicluster structure it still creates two hierarchy levels, with the second level having a single child node, which is clearly an artifact of the model assumptions. Other improvements include allowing microarrays to choose more than one path along the tree, and having a feature activation model that would allow genes to be activated in accordance to pre-specified pathway memberships.

As for the general biclustering framework proposed in Publication IV, it would be interesting to assess model performance when more complex experts are used, instead of the simple Gaussian and Bernoulli experts. The inference engine also appears not to be robust with respect to initialization procedures, which brings in the question of whether global optimization techniques such as simulated annealing [144] would improve this aspect.

Overall, the proposed probabilistic biclustering methods achieve state-of-the-art performance, and are thus promising tools for this computational task. In particular, the proposed mixture-of-experts approach to biclustering can replace several existing biclustering methods as it allows performing biclustering under a variety of data types and model assumptions. The proposed hierarchical biclustering framework is also useful with respect to existing methods, as it is currently the only method that allows inferring a hierarchy of samples and explicitly indicate the genes that tie each node in the hierarchy, an aspect which facilitates the subsequent biological analysis.

5. Model-Based Information Retrieval in Gene Expression

This chapter describes the contributions about information retrieval in gene expression data corresponding to Publications II and V. The chapter starts with a brief motivation and a review of related work. Then, the proposed pipelines involving the use of probabilistic latent variable models are described. The chapter ends with a description of an RT-PCR validation experiment, based on an *in silico* prediction made by the method described in Publication V.

5.1 Motivation and Earlier Work

Information retrieval can be defined as the computational task of obtaining the data which is most relevant to an information need [94]. For instance, Internet search engines [107] are routinely used to relate objects from multiple media types such as movies, images, or hypertext. Given the pace at which biological data is being generated due to the emergence of high-throughput technologies, it makes sense to develop IR tools that systematically find similarities across data from different studies, in order to accelerate research and obtain both novel and robust findings. While this applies to any kind of biological data and suggests the use of integrative IR frameworks that couple multiple data sources, here the focus is on information retrieval for gene expression data obtained with DNA microarrays.

Gene expression IR methods are divided into text-driven and data-driven approaches. Text-driven approaches detect similarities in the textual annotations of the studies [92, 155]. These methods are effective at finding studies corresponding to conditions that are known to be related, which is an important task when the aim is, for example, to collect all avail-

able public data regarding a particular disease and a given treatment. However, data-driven approaches hold a greater potential for uncovering novel biological findings, as the bulk of the information about a study is contained in the measured data itself, rather than in the textual annotations.

Data-driven IR is intrinsically related to gene expression meta-analysis. Since the aim of meta-analysis is to analyze a large set of independent studies, meta-analysis and IR methods ultimately share several technical challenges, including the sparseness of annotation data, variability across samples and experimental protocols, the use of different microarray platforms, and the inherent complexity and stochasticity of the biological conditions under study.

Data-driven IR in gene expression data has been suggested as far as 1999 as a form of alignment [70], and methods exist at least since 2001, with the work of Hunter *et al*., who propose a measure of relevance based on a Bayes factor that compares the hypothesis that two microarrays correspond to the same biological condition vs. two different biological conditions [59]. More recent IR and meta-analysis approaches rely on using differential expression as a solution to the problem of between-study incommensurability. Two known examples of meta-analysis methods based on differential expression are Module Maps [123] and Oncomine [117], with both methods being applied to study the differential expression of gene sets across multiple neoplasms. Another influential approach is the Connectivity Map [80], in which the authors produced a data set of treatment response in cell lines, along with a nonparametric correlation procedure which allows querying the data set with a gene set of interest, retrieving the most relevant compounds.

Current state-of-the-art meta-analysis and IR approaches typically rely on connecting studies using nonparametric correlation measures that take into account the differential expression patterns observed in each study [76]. In general, using differential expression as a basis for encoding and relating studies is an effective way to deal with expression data incommensurability.

5.2 Latent Variable Models for Information Retrieval

Both existing meta-analysis and IR approaches, as well as the proposed methods from Publications II and V, can be seen as instances of a general meta-analysis and information retrieval framework consisting of four components: First, each gene expression profiling study is transformed into a set of comparisons between biological conditions. Second, the differential expression of genes or gene sets is computed, using measures such as fold-change, or statistical methods such as the t-test or GSEA. Third, differential expression patterns are extracted from the data using unsupervised learning methods. Fourth, a relevance measure between studies, conditions, or microarrays is used to find interesting connections in the data.

A flowchart outlining the proposed methods is shown in Figure 5.1. In the following subsections the main parts of the flowchart are described in detail.

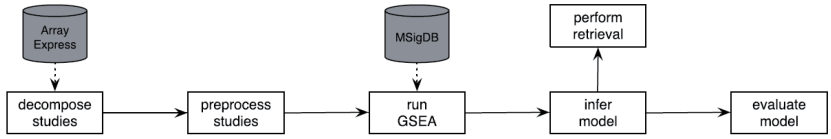


Figure 5.1. Flowchart for the methods presented in Publications II and V. Figure adapted from Publication V.

5.2.1 Study Decomposition

Regarding the decomposition of a study's experimental design into a set of comparisons between pairs of conditions, the proposed approach minimizes the influence of confounding factors and maximizes the interpretability of the derived comparisons.

Using the same terminology as in the ArrayExpress database [108], each study contains a set of *experimental factors*, *e.g.*, "disease-state", "tissue", or "compound". Any given sample in the study contains an instantiation of each experimental factor. For instance, a given sample may have the annotation "disease-state = normal", "tissue = liver", and "compound = none". The instantiation of a given experimental factor is known as an *experimental factor value*.

For every study, a set of comparisons between pairs of conditions is automatically derived. This is done by exhaustively finding all maximally

large sample groups in which all experimental factors share the same value, except for a single experimental factor which has either of two possible values. For instance, a given set of samples may share the annotations “tissue = liver” and “compound = none” for the “tissue” and “compound” experimental factors, but have one of two possible annotations “disease-state = cardiomyopathy” or “disease-state = normal” for the “disease-state” experimental factor. This yields a comparison between cardiomyopathy and normal samples, in the *context* of “tissue = liver” and “compound = none”.

Alternative methods, *e.g.*, Module Maps [123], in which a sample is compared against the average expression of all samples in a study, yield comparisons that depend intrinsically on the conditions that were used to compute the average expression, therefore containing an additional layer of study-specific bias. Also, approaches that compare pairs of conditions without regarding contextual information [35, 57, 58] hinder the subsequent analysis, due to the influence of confounding factors.

The above approach for decomposing a study was proposed in Publication II; in Publication V it was refined by introducing the notion of a *neutral factor*. A neutral factor is a factor that (1) does not yield meaningful comparisons without any further processing, for instance the factor “age”, whose values are typically not categorical, *e.g.*, “age $\geq$ 65” or “age $\leq$ 65”, but rather numerical, and that (2) usually contains a high number of unique factor values, as many as one per sample, *e.g.*, unique patient identifiers. In order to avoid generating a large number of meaningless comparisons, a list of neutral factors was manually built and removed from the experimental design annotation.

5.2.2 Gene Set Enrichment Analysis

After decomposing studies into sets of comparisons between biological conditions, differential expression patterns were extracted from each comparison. Concretely, GSEA was used in each comparison to test for the differential expression of a set of 639 canonical pathways obtained from MSigDB [131].

The use of a gene set test rather than a gene test is due to the fact that procedures that test for the differential expression of gene sets have been observed to be more robust across studies [131]. Using a gene set test also allows re-using biological knowledge of pathways in the context of

meta-analysis and IR.

Intuitively, GSEA tests if the genes in a gene set tend to appear among the most differentially expressed genes. Unlike classical enrichment tests such as the hypergeometric test, GSEA does not require a preprocessing step wherein genes are classified as differentially expressed or non-differentially expressed. Instead, it takes as input the whole list of genes in a given study, sorted according to a differential expression measure such as the log-ratio (sorting the list in descending or ascending order makes GSEA look for differentially over or under-expressed gene sets, respectively).

GSEA computes a running score by traversing the sorted list of genes in the study. The score is increased whenever a gene from the gene set is found and is decreased otherwise. Formally, given a gene set S, the score at position i in the list is computed via the following recursive formula:

$$\text{score}(i) = \begin{cases} \text{score}(i-1) + \frac{|r_i|^p}{N_R} & , i \in S \\ \text{score}(i-1) - \frac{1}{N - N_H} & , i \notin S \end{cases} \tag{5.1}$$

where r_i is the differential expression measure assigned to the gene at position i in the list, $N_R = \sum_{i \in S} |r_i|^p$, N is the total number of genes in the study, N_H is the size of gene set S, and p is an exponent that weighs the contribution from each gene in the gene set. In the original GSEA publication, p was set to zero, which induced a positive score bias for gene sets whose genes are not differentially expressed [131]. In both Publications II and V, p is set to one, as suggested by the GSEA authors [131].

The final *enrichment score* (ES) for a gene set is computed as the maximum of the running score described above. Significance is assessed by performing 1000 random permutations of the phenotype labels and recomputing the corresponding ES's. The ES is also normalized by dividing it by the mean of the random ES's. Finally, the genes in the gene set that were found before the running score reached its maximum are collectively known as the gene set's *leading edge subset*. This subset corresponds to the most differentially expressed genes in the set.

For each comparison, the 50 gene sets with the highest normalized ES were collected, ignoring the direction of differential expression. While this threshold-based selection procedure is partly heuristic, in preliminary studies it was found that it leads to a better IR performance than choosing gene sets based on the standard cut-off value of $q < 0.05$, which yields an excessively sparse encoding where the majority of comparisons

has zero differentially expressed gene sets.

In Publication II, each comparison was encoded as a vector where each entry corresponds to a gene set. For each of the top 50 gene sets, the corresponding entry contained the number of genes in the gene set's leading edge subset. The entries for the remaining gene sets were filled with zeroes. This encoding is equivalent to the *bag-of-words* representation in the natural language domain, where a document is represented by a vector of word frequencies. In Publication V, the actual genes that belong to the leading edge subset of each of the top 50 gene sets were used. This was done in order to also model gene-wise differential expression. Thus, the proposed approaches represent each comparison in terms of its corresponding GSEA output.

5.2.3 Probabilistic Modelling

Given a large number of comparisons represented by their corresponding GSEA output, the next step in the proposed pipeline is to perform a probabilistic modelling of the observed GSEA results. Latent variable mixture models were used in both Publications II and V. Here, the aim is to use the models both for inferring patterns of differential expression that occur across multiple comparisons, and for providing a basis for performing IR.

Both probabilistic models assume the existence of a certain number of so-called mixture components. In Publication II, each component captures a co-occurrence pattern among gene sets, while in Publication V, each component captures a co-occurrence pattern among gene sets coupled with a co-occurrence pattern among genes. The intuitive idea is to detect groups of gene sets or genes that are often differentially expressed together. Both models also create a soft mapping from comparisons to these components, as described in detail below.

In Publication II, LDA [15] was used to model the GSEA output. In LDA, components are known as *topics*, with each topic k corresponding to a multinomial distribution over gene sets ϕ_k. Here, a topic has the biological meaning of representing a group of gene sets that are often simultaneously differentially expressed. LDA assumes a generative process wherein each GSEA comparison i has a multinomial distribution over topics θ_i. Each observation of a gene set in a GSEA comparison is assumed to arise by first selecting a topic k using θ_i and then choosing the gene set using ϕ_k. Both θ_i and ϕ_k are assigned conjugate Dirichlet priors, with in-

ference and estimation being performed using a collapsed Gibbs sampler [51].

In Publication V, a novel latent variable mixture model was proposed. Each component k corresponds to two vectors $\boldsymbol{\phi}_k$ and $\boldsymbol{\psi}_k$, with $\boldsymbol{\phi}_{ks} \in [0, 1]$ representing the activation probability of gene set s and $\boldsymbol{\psi}_{kg} \in [0, 1]$ representing the activation probability of gene g. The generative process assumes that the data for a given comparison i and gene set s arises by choosing a component k, then activating the gene set with probability ϕ_{ks} and finally, if the gene set is active, creating its leading edge subset by sampling the activation status of each gene g in the gene set using ψ_{kg}. The model also assumes a two-level component selection procedure that allows modelling correlations between components [86]. As in LDA, each GSEA comparison i has a distribution $\boldsymbol{\theta}_i$ over so-called *modules*. Each module in turn has a distribution $\boldsymbol{\eta}_m$ over components. A component is chosen by first choosing a module m from $\boldsymbol{\theta}_i$ and then choosing the component from $\boldsymbol{\eta}_m$. Intuitively, a module represents a soft combination of differential expression components that aims at capturing higher-level biological phenomena. Dirichlet and Beta conjugate priors are used for the model variables, and a collapsed Gibbs sampler was used to perform inference and estimation.

Finally, in both probabilistic models, the generative probability formulation described in Chapter 2 was used as a relevance measure between comparisons.

5.3 Results

Both methods were applied to large sets of microarray studies obtained from the ArrayExpress database. In Publication II, the method was applied to slightly less than 800 comparisons derived from 288 human studies, while in Publication V the method was applied to a larger data set consisting of 6925 comparisons derived from 1082 studies involving three species (human, mouse, and rat).

The proposed methods were evaluated according to their IR performance, as well as according to the biological relevance of the inferred components and selected IR case studies.

5.3.1 Retrieval Performance

The analysis of the methods' IR performance was restricted to a subset of so-called *interpretable* comparisons in which a condition is compared against a control. These were named "interpretable" because IR results restricted to these comparisons are inherently easier to systematically assess.

In Publication II the Average Precision performance measure [94] was used, with the proposed method being compared to a random baseline when querying with either of 27 cancer vs. normal comparisons involving different cancer types. The results showed that in 20 out of 27 comparisons the model performance was superior to the 99% confidence interval of the random baseline.

In Publication V, the evaluation approach relied on using a controlled vocabulary known as the Experimental Factor Ontology (EFO) [92] which characterizes some of the existing experimental factor values. A mapping from experimental factor values to ontology terms was successfully obtained for 219 interpretable comparisons, and a ground-truth relevance measure between comparisons was derived based on the shared path between the corresponding terms in the EFO. Since this approach yields a non-binary relevance measure between comparisons, the Average Precision performance measure was replaced by the Normalized Discounted Cumulative Gain (NDCG) measure [63, 94], which effectively handles non-binary relevance measures. The results showed that the proposed method performs competitively with existing approaches.

5.3.2 Qualitative Evaluation

As for the biological interpretation of the components, each component can be analyzed by either looking at its most probable elements or by significance-based approaches such as the one outlined in Publication V. In both cases, it was shown that components correspond to functionally coherent groups related to biological processes such as apoptosis, respiration, metabolism, or inflammation.

Several biological case studies were also provided in both publications II and V, indicating how the model can be used to obtain connections between comparisons that correspond to existing knowledge and hint at potential novel findings.

Finally, in Publication II information visualization tools for visualizing both the model and the information retrieval results were also developed. In Publication II, Figure 1, a proposed graph-based visualization tool allowed for a holistic analysis of the relations between comparisons, components, and gene sets. In Publication II, Figure 4, an IR-based nonlinear dimensionality reduction method [148] was used to visualize both the retrieval results and the importance of each component in each comparison. Both of these tools facilitate the inspection of the results, with the corresponding figures provided in Publication II showing the biological meaningfulness of the inferred model structures and IR results.

5.4 An Application to *SIM2s* Expression in Pleural Malignant Mesothelioma

A more detailed case study in Publication V concerned a computationally predicted connection between malignant pleural mesothelioma (MPM) and the transcription factor single-minded homolog 2, short isoform (*SIM2s*).

MPM is a rare form of cancer that develops in the pleura and which is primarily caused by asbestos exposure. *SIM2* is a basic helix-loop-helix transcription factor with short (*SIM2s*) and long (*SIM2l*) isoforms. *SIM2* is located on chromosome 21 and has been associated with Down syndrome [26]. It has also been observed to be differentially expressed in prostate [53] and breast cancer [77].

In Publication V, when querying the model with a comparison of MPM versus normal pleural tissue in human [49], the third most relevant result was a comparison of an RNA interference (RNAi) knockdown assay of *SIM2s* in a human colon carcinoma cell line [2]. Specifically, the comparison was of “time = 18h” versus the “time = 0h” control. No known connections between MPM and *SIM2s* exist, although it has been shown that Sim2 mutant mice incur in pleural defects [50]. This suggests that *SIM2s* may be differentially expressed in MPM samples compared to a control, although in the MPM study analyzed by the model this was not the case.

In order to follow-up on the computationally predicted connection between MPM and *SIM2s*, an independent set of 10 MPM samples and a healthy pleural control were analyzed. Concretely, an RT-PCR assay was performed to measure the expression of both *SIM2* isoforms, as well as a

small number of genes known to be *SIM2s* targets, namely *MMP2*, *MMP3*, *SNAI1*, *SNAI2*, and *MYOM2*. This set of known *SIM2s* targets was obtained via a literature search which aimed at collecting all known *SIM2s* targets. The expression of *MMP14* was also measured, as *MMP14* has been recently observed to be differentially expressed in MPM [28].

The primary result was that *SIM2s* is significantly under-expressed in the MPM samples ($p < 0.05$). Differential expression of *MMP14* was not confirmed, although its expression was significantly correlated with that of *MMP2* ($r = 0.74$, $p < 0.05$), which is consistent with the knowledge that *MMP2* requires *MMP14* for its activation [28]. *SNAI2* was differentially over-expressed ($p < 0.05$), which is consistent with its putative role as a repressive transcriptional target of *SIM2s* [78].

Observing differential expression of *SIM2s* in an independent set of MPM samples suggests that *SIM2s* may effectively have a role in the disease. In Publication V, it is hypothesized that the role of *SIM2s* in MPM is related to estrogen signalling and the epithelial-mesenchymal transition (EMT) network.

Regarding estrogen signalling, both gender and estrogen receptor-β (ERβ) expression have prognostic power in MPM [112], although it is an open question the extent to which estrogen signalling is important in MPM. The *GADD45A* gene, which was over-expressed in the *SIM2s* RNAi study that the probabilistic model connected to MPM [2], is a transcriptional target of ERβ [109]. *SIM2s* depletion in a mouse model also yields ER-negative tumours [78]. Finally, in the probabilistic model, the three overall most probable gene sets in both the MPM and the *SIM2s* studies are "metabolism of xenobiotics by cytochrome p450", "androgen and estrogen metabolism", and "arachidonic acid metabolism". Cytochrome p450 enzymes mediate estrogen metabolism [143], and the "arachidonic acid metabolism" gene set significantly overlaps with the "metabolism of xenobiotics by cytochrome p450" gene set ($p < 0.05$).

The EMT network is a group of related genes with an important role in tissue development that has also been associated with cancer progression [140]. It has been observed that *SIM2s* depletion in mouse induces a transition similar to the EMT [78]. It has also been recently observed that EMT genes, including *SNAI2*, are over-expressed in MPM [23].

In summary, the computationally predicted connection between MPM and *SIM2s* has been validated via a follow-up RT-PCR study, yielding an hypothesis that *SIM2s* has a role in MPM, potentially via the estrogen

signalling and EMT networks.

5.5 Discussion

This chapter presented the contributions regarding IR in gene expression data. The proposed approach can be extended in several directions. For instance, using a nonparametric Bayesian prior based on the Hierarchical Dirichlet Process model [138] would allow automatically learning the number of components. Additionally, the differential expression encoding could be based on alternative methods, such as sparse linear classifiers [153] instead of GSEA. In the future, it will also become important to adapt the proposed framework to new high-throughput sequencing technologies such as RNA-Seq. Finally, significance-based measures of relevance may be used in the future to automatically suggest the most interesting directions of exploration of the retrieval results.

On the biological side, all described case studies suggest follow-up experiments for uncovering the molecular basis underlying the found connections between biological conditions. In particular, regarding the *SIM2s* case study, future studies may potentially involve studying the expression of *SIM2s* in a human MPM cell line after treatment with estrogen-related compounds.

In general, it is an open question how far *SIM2s* drives the gene expression profile observed in MPM patients. Ultimately, a better understanding of the mechanisms of estrogen signalling and EMT networks, as well as the corresponding role of *SIM2s*, may yield better prognostication and more targeted treatment in MPM.

6. Summary and Conclusions

This thesis summarized contributions on using graphical models for performing biclustering and IR of gene expression data.

Biclustering is a relevant challenge in bioinformatics since it aims at finding local regions of interest in high-throughput data sets, namely subsets of measurements and biological conditions that corresponds to meaningful data patterns. In Publications I, III, and IV, novel methods were proposed for various biclustering tasks. It was shown that the models had a state-of-the-art performance and were able to learn multiple types of bicluster structures, such as a bicluster hierarchy (Publication III) or a set of partially overlapping biclusters (Publications I and IV). Those inferred structures were biologically meaningful and allowed formulating novel hypotheses on associations between diseases and miRNA genes or chromosomal bands.

IR in gene expression data is a timely task due to the ever-increasing number of data sets deposited in public repositories. Reutilizing existing data via a "Google-like" exploration will potentially accelerate the pace of biological research. In Publications II and V, a pipeline for IR in gene expression data was proposed, with the aim being to provide an effective approach for relating the large number of independent microarray gene expression studies deposited in public databases. The proposed IR pipeline combines a differential expression-based representation of studies with probabilistic modelling that finds recurrent patterns of differential expression and provides a sound basis for performing IR. The models were able to learn meaningful patterns of differential expression and provide valid connections between independent studies.

The proposed methods highlight the merits of graphical models. Graphical models are flexible, modular tools for data analysis that allow combining multiple "probabilistic blocks" in order to approach hard compu-

tational tasks. The existence of a large number of graphical modelling approaches, as well as general-purpose inference and estimation methods such as the EM algorithm or the Gibbs sampler, provides a large methodological basis from which novel models and the corresponding inference and estimation procedures can be developed.

As for future work, there is wide scope of applications for the proposed methods. For instance, the proposed IR pipeline may be used for drug repurposing, by connecting diseases, drugs, and pathways, with the aim of explaining the connections via similarities in observed differential expression patterns. There is also the possibility of extending the IR pipeline in order to perform a more thorough pre-processing of microarray data sets. In the proposed approach, fully pre-processed microarray data sets are imported and probe-set-to-gene dictionaries are obtained from sources such as MSigDB. However, as described in Section 2.2, it has been shown that probe-to-transcript mappings are often inaccurate due to the genomic annotations available at the time when the corresponding microarray platforms were developed [29]. A future approach may involve an additional microarray pre-processing step that uses up-to-date probe-to-transcript mappings in order to pre-process a given data set for obtaining more accurate transcript levels. Another direction of research is integration of multiple data types, which was not explored in the methods described in this thesis but is becoming a more relevant task due to the emergence of large-scale projects such as the Cancer Genome Atlas [106].

In general, the bioinformatics field is witnessing the emergence of new high-throughput sequencing technologies and a greater effort towards large-scale, multi-modal data acquisition [80, 106]. The advent of "big data" will bring about fundamental challenges in graphical modelling, placing stronger requirements on model *scalability* and *interpretability*. Regarding scalability, inference and estimation methods will be required to cope with large amounts of data and provide results that are robust with respect to initialization procedures. This will be particularly important for models that infer complex combinatorial structures, which so far lack standard solutions for assessing the robustness of the findings. Regarding interpretability, tools such as information visualization techniques will become increasingly important for providing a holistic analysis of the computational structures inferred by the models.

Finally, the main biological finding described in this thesis is the differential expression of *SIM2s* in MPM described in Publication V. So far,

SIM2s has been shown to be related to components of the EMT and estrogen signalling networks, which are important players in cancer. For instance, the EMT network is known to be a driver of metastasis [140], while estrogen-related signalling is a potential cause for gender differences in survival time in MPM [112]. The exact role of *SIM2s* is not presently understood, although it is known to interact with EMT and estrogen signalling-related genes such as *GADD45A* and *SNAI2*. A better understanding of the role of *SIM2s* may in turn lead to a better understanding of EMT and estrogen signalling, with implications for the treatment of MPM. It is tempting to speculate that ultimately, expression patterns of *SIM2s* and related genes may be used to stratify MPM patients in order to provide more targeted treatments that increase overall survival.

Bibliography

[1] M. Ackermann and K. Strimmer. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, 10(1):47, 2009.

[2] M. J. Aleman, M. P. DeYoung, M. Tress, P. Keating, G. W. Perry, and R. Narayanan. Inhibition of Single Minded 2 gene expression mediates tumor-selective apoptosis and differentiation in human colon cancer cells. *Proc. Natl. Acad. Sci. U.S.A.*, 102(36):12765–12770, 2005.

[3] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.*, 7(1):55–65, 2005.

[4] A. Antoniou and W.-S. Lu. *Practical Optimization*. Springer, New York, NY, 2007.

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

[6] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muertter, and R. Edgar. NCBI GEO: Archive for high-throughput functional genomic data. *Nucleic Acids Res.*, 37(suppl. 1):D885–D890, 2009.

[7] A. Battle, E. Segal, and D. Koller. Probabilistic discovery of overlapping cellular processes and their regulation. *J. Comput. Biol.*, 12(7):909–927, 2005.

[8] J. Bernardo and A. Smith. *Bayesian Theory*. John wiley and Sons, New York, NY, 1994.

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

[10] C. M. Bishop and M. Svensén. Bayesian hierarchical mixture of experts. In C. Meek and U. Kjærulff, editors, *Proceedings of the Nineteenth Conference Conference on Uncertainty in Artificial Intelligence*, pages 57–64. Morgan Kaufmann, San Francisco, CA, 2003.

[11] C. M. Bishop, J. M. Winn, and D. Spiegelhalter. VIBES: A variational inference engine for Bayesian networks. In S. Becker, S. Thrun, and K. Obermeyer, editors, *Advances in Neural Processing Systems*, volume 15, pages 793–800. The MIT Press, Cambridge, MA, 2002.

[12] D. Blackwell and J. B. MacQueen. Ferguson distributions via Polya urn schemes. *Ann. Statist.*, 1(2):353–355, 1973.

[13] D. Blei and J. Lafferty. A correlated topic model of *Science*. *Ann. Appl. Stat.*, 1(1):17–35, 2007.

[14] D. M. Blei, T. L. Griffiths, and M. I. Jordan. The nested chinese restaurant process and Bayesian inference of topic hierarchies. *J. ACM*, 57(2):1–30, 2010.

[15] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 17–24, Cambridge, MA, 2003. The MIT Press.

[16] D. M. Blei and M. I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–144, 2006.

[17] G. Bouchard. Efficient bounds for the softmax and applications to approximate inference in hybrid models. In *NIPS 2007 Workshop on Approximate Inference in Hybrid Models*. 2007.

[18] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

[19] Breiman92. *Probability*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992.

[20] W. Buntine. Chain graphs for learning. In P. Besnard and S. Hanks, editors, *Proceedings of the Eleventh Annual Conference on Uncertainty in Artificial Intelligence*, pages 46–54, San Francisco, CA, 1995. Morgan Kaufmann.

[21] W. Buntine, J. Lofstrom, J. Perkio, S. Perttu, V. Poroshin, T. Silander, H. Tirri, A. Tuominen, and V. Tuulos. A scalable topic-based open source search engine. In N. Zhong, J. Bradshaw, S. K. Pal, D. Talia, J. Liu, and N. Cercone, editors, *Proceedings of the 2004 IEEE/WIC/ACM International Joint Conference on Web Intelligence*, pages 228–234, Los Alamitos, CA, 2004. IEEE Computer Society.

[22] N. Bushati and S. M. Cohen. MicroRNA functions. *Annu. Rev. Cell. Dev. Biol.*, 23:175–205, 2007.

[23] C. Casarsa, N. Bassani, F. Ambrogi, G. Zabucchi, P. Boracchi, E. Biganzoli, and D. Coradini. Epithelial-to-mesenchymal transition, cell polarity and stemness-associated features in malignant pleural mesothelioma. *Cancer Lett.*, 302(2):136–143, 2011.

[24] M. Charrad and M. B. Ahmed. Simultaneous clustering: A survey. In S. O. Kuznetsov, D. P. Mandal, M. K. Kundu, and S. K. Pal, editors, *4th International Conference on Pattern Recognition and Machine Intelligence*, pages 370–375, Springer-Verlag, 2011. Berlin.

[25] Y. Cheng and G. M. Church. Biclustering of expression data. In P. E. Bourne, M. Gribskov, R. B. Altman, N. Jensen, D. A. Hope, T. Lengauer, J. C. Mitchell, E. D. Scheeff, C. Smith, S. Strande, and H. Weissig, editors, *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pages 93–103, Menlo Park, CA, 2000. AAAI Press.

[26] R. Chrast, H. S. Scott, R. Madani, L. Huber, D. P. Wolfer, M. Prinz, A. Aguzzi, H.-P. Lipp, and S. E. Antonarakis. Mice trisomic for a bacterial artificial chromosome with the single-minded 2 gene (sim2) show phenotypes similar to some of those present in the partial trisomy 16 mouse models of Down syndrome. *Hum. Mol. Genet.*, 9(12):1853–1864, 2000.

[27] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. The MIT Press, Cambridge, MA, 2nd edition, 2001.

[28] S. Crispi, R. A. Calogero, M. Santini, P. Mellone, B. Vincenzi, G. Citro, G. Vicidomini, S. Fasano, R. Meccariello, G. Cobellis, S. Menegozzo, R. Pierantoni, F. Facciolo, A. Baldi, and M. Menegozzo. Global gene expression profiling of human pleural mesotheliomas: Identification of matrix metalloproteinase 14 (MMP-14) as potential tumour target. *PLoS One*, 4(9):e7016, 2009.

[29] M. Dai, P. Wang, A. D. Boyd, G. Kostov, B. Athey, E. G. Jones, W. E. Bunney, R. M. Myers, T. P. Speed, H. Akil, S. J. Watson, and F. Meng. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.*, 33(20):e175, 2005.

[30] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B Met.*, 39(1):1–38, 1977.

[31] P. D'Haeseleer. How does gene expression clustering work? *Nat. Biotechnol.*, 23(12):1499–1501, 2005.

[32] F. Doshi-Velez, K. T. Miller, J. Van Gael, and Y. W. Teh. Variational inference for the Indian buffet process. In D. van Dyk and M. Welling, editors, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, pages 137–144, Cambridge, MA, 2009. JMLR.

[33] Sorin Drăghici. *Data Analysis Tools for DNA Microarrays*. Chapman & Hall/CRC, Boca Raton, FL, 2003.

[34] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.*, 95(25):14863–14868, 1998.

[35] J. Engreitz, A. Morgan, J. Dudley, R. Chen, R. Thathoo, R. B. Altman, and A. J. Butte. Content-based microarray search using differential expression profiles. *BMC Bioinformatics*, 11(1):603, 2011.

[36] M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *J. Am. Stat. Assoc.*, 90(430):577–588, 1995.

[37] P. J. Farnham. Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.*, 10(9):605–616, 2009.

[38] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Ann. Stat.*, 1(2):209–230, 1973.

[39] P. Flaherty, G. Giaever, J. Kumm, M. I. Jordan, and A. P. Arkin. A latent variable model for chemogenomic profiling. *Bioinformatics*, 21(15):3286–3293, 2005.

[40] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL, 2nd edition, 2004.

[41] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE T. Pattern Anal.*, 6(6):721–741, 1984.

[42] R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit, editors. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer Science+Business Media, Inc., New York, NY, 2005.

[43] G. K. Gerber, R. D. Dowell, T. S. Jaakkola, and D. K. Gifford. Automated discovery of functional generality of human gene expression programs. *PLoS Comput. Biol.*, 3(8):e148, 2007.

[44] G. Getz, E. Levine, and E. Domany. Coupled two-way clustering analysis of gene microarray data. *P. Natl. Acad. Sci. U. S. A.*, 97(22):12079–12084, 2000.

[45] Z. Ghahramani, T. L. Griffiths, and P. Sollich. Bayesian nonparametric latent feature models. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors, *Bayesian Statistics 8*, pages 201–226, Oxford, UK, 2006. Oxford University Press.

[46] J. J. Goeman and P. Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987, 2007.

[47] G. H. Golub and C. F. van Loan. *Matrix Computations*. The John Hopkins University Press, Baltimore, MD, 3rd edition, 1996.

[48] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

[49] G. J. Gordon, G. N. Rockwell, R. V. Jensen, J. G. Rheinwald, J. N. Glickman, J. P. Aronson, B. J. Pottorf, M. D. Nitz, W. G. Richards, D. J. Sugarbaker, and R. Bueno. Identification of novel candidate oncogenes and tumor suppressors in malignant pleural mesothelioma using large-scale transcriptional profiling. *Am. J. Pathol.*, 166(6):1827–1840, 2005.

[50] E. Goshu, H. Jin, R. Fasnacht, M. Sepenski, J. L. Michaud, and C. M. Fan. *Sim2* mutants have developmental defects not overlapping with those of *Sim1* mutants. *Mol. Cell Biol.*, 22(12):4147–4157, 2002.

[51] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *P. Natl. Acad. Sci. U. S. A.*, 101(suppl. 1):5228–5235, 2004.

[52] J. Gu and J. S. Liu. Bayesian biclustering of gene expression data. *BMC Genomics*, 9(Suppl. 1):S4, 2008.

[53] O. J. Halvorsen, K. Rostad, A. M. Øyan, H. Puntervoll, T. H. BØ, L. Stordrange, S. Olsen, S. A. Haukaas, L. Hood, I. Jonassen, K. H. Kalland, and L. A. Akslen. Increased expression of SIM2-s protein is a novel marker of aggressive prostate cancer. *Clin. Cancer Res.*, 13(3):892–897, 2007.

[54] D. Hanagan and R. A. Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5544):646–674, 2011.

[55] J. A. Hartigan. Direct clustering of a data matrix. *J. Am. Stat. Assoc.*, 67(337):123–129, 1972.

[56] S. Hochreiter, U. Bodenhofer, M. Heusel, A. Mayr, A. Mitterecker, A. Kasim, T. Khamiakova, S. Van Sanden, D. Lin, W. Talloen, L. Bijnens, H. W. H. Gühlmann, Z. Shkedy, and D.-A. Clevert1. FABIA: factor analysis for bicluster acquisition. *Bioinformatics*, 26(12):1520–1527, 2010.

[57] G. Hu and P. Agarwal. Human disease-drug network based on genomic expression profiles. *PLoS One*, 4(8):e6536, 2009.

[58] H. Huang, C.-C. Liu, and X. J. Zhou. Bayesian approach to transforming public gene expression repositories into disease diagnosis databases. *Proc. Natl. Acad. Sci. U.S.A.*, 107(15):6823–6828, 2010.

[59] L. Hunter, R. C. Taylor, S. M. Leach, and R. Simon. GEST: a gene expression search tool based on a novel bayesian similarity metric. *Bioinformatics*, 17(Suppl. 1):S115–S122, 2001.

[60] H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *JASA*, 96(453):161–173, 2001.

[61] T. Jaakkola. Tutorial on variational approximation methods. In M. Opper and D. Saad, editors, *Advanced mean field methods: theory and practice*, pages 129–160. The MIT Press, Cambridge, MA, 2000.

[62] T. Jaakkola and M. I. Jordan. Bayesian parameter estimation via variational methods. *Stat. Comput.*, 10(1):25–37, 2000.

[63] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM T. Inform. Syst.*, 20(4):422–446, 2002.

[64] M. Jeanmougin, A. de Reynies, L. Marisa, C. Paccard, G. Nuel, and M. Guedj. Should we abandon the t-test in the analysis of gene expression microarray data: a comparison of variance modeling strategies. *PLoS One*, 5(9):e12336, 2010.

[65] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-dna interactions. *Science*, 316(5830):1497–1502, 2007.

[66] M. I. Jordan. Hierarchical models, nested models and completely random measures. In M.-H. Chen, D. Dey, P. Mueller, D. Sun, and K. Ye, editors, *Frontiers of Statistical Decision Making and Bayesian Analysis: In Honor of James O. Berger*, pages 207–218. Springer, New York, NY, 2010.

[67] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233, 1999.

[68] M. I. Jordan and R. A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214, 1994.

[69] M. I. Jordan and Y. Weiss. Graphical models: Probabilistic inference. In M. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*. The MIT Press, Cambridge, MA, 2nd edition, 2002.

[70] D. E. Bassett Jr., M. B. Eisen, and M. S. Boguski. Gene expression informatics — it's all in your mine. *Nat. Genet.*, 21(Suppl. 1):51–55, 1999.

[71] G. Karlebach and R. Shamir. Modelling and analysis of gene regulatory networks. *Nat. Rev. Mol. Cell Biol.*, 9(10):770–780, 2008.

[72] P. Khatri and S. Drăghici. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587–3595, 2005.

[73] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, Cambridge, MA, 2009.

[74] J. E. Krebs, E. S. Goldstein, and S. T. Kilpatrick. *Lewin's Genes X*. Jones and Bartlett Publishers, Sudbury, MA, 2009.

[75] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Trans. Info. Theory*, 47(2):498–519, 2001.

[76] I. Kupershmidt, Q. J. Su, A. Grewal, S. Sundaresh, I. Halperin, J. Flynn, M. Shekar, H. Wang, J. Park, W. Cui, G. D. Wall, R. Wisotzkey, S. Alag, S. Akhtari, and M. Ronaghi. Ontology-based meta-analysis of global collections of high-throughput public data. *PLoS One*, 5(9):e13066, 2010.

[77] H. I. Kwak, T. Gustafson, R. P. Metz, B. Laffin, P. Schedin, and W. W. Porter. Inhibition of breast cancer growth and invasion by single-minded 2s. *Carcinogenesis*, 28(2):259–266, 2007.

[78] B. Laffin, E. Wellberg, H.-I. Kwak, R. C. Burghardt, R. P. Metz, T. Gustafson, P. Schedin, and W. W. Porter. Loss of singleminded-2s in the mouse mammary gland induces an epithelial-mesenchymal transition associated with up-regulation of slug and matrix metalloprotease 2. *Mol. Cell. Biol.*, 28(6):1936–1946, 2008.

[79] J. Lamb. The Connectivity Map: a new tool for biomedical research. *Nat. Rev. Cancer*, 7(1):54–60, 2007.

[80] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong, S. J. Haggarty, P. A. Clemons, R. Wei, S. A. Carr, E. S. Lander, and T. R. Golub. The connectivity map: Using gene-expression signatures to connect small molecules, genes, and diseases. *Science*, 313(5795):1929–1935, 2006.

[81] S. L. Lauritzen. *Graphical Models*. Oxford University Press, New York, NY, 1996.

[82] L. Lazzeroni and A. Owen. Plaid models for gene expression data. *Stat. Sinica*, 12(1):61–86, 2002.

[83] J. M. Lee, S. Dedhar, R. Kalluri, and E. W. Thompson. The epithelial-mesenchymal transition: new insights in signaling, development, and disease. *J. Cell Biol.*, 172(7):973–981, 2006.

[84] M. B. Lens and J. A. Newton-Bishop. An association between cutaneous melanoma and non-hodgkin's lymphoma: pooled analysis of published data with a review. *Ann. Oncol.*, 16(3):460–465, 2004.

[85] M. Levine and E. H. Davidson. Gene regulatory networks for development. *P. Natl. Acad. Sci. U. S. A.*, 102(14):4936–4942, 2005.

[86] W. Li and A. McCallum. Packinko allocation: DAG-structured mixture models of topic correlations. In W. W. Cohen and A. Moore, editors, *Proceedings of the Twenty-Third International Conference on Machine Learning*, pages 577–584, New York, NY, 2006. ACM Press.

[87] Jun S. Liu. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *J. Am. Stat. Assoc.*, 89(427):958–966, 1994.

[88] Q. Liu, I. Dinu, A. J. Adewale, J. D. Potter, and Y. Yasui. Comparative evaluation of gene-set analysis methods. *BMC Bioinformatics*, 8(1):431, 2007.

[89] H. Lodish, A. Berk, C. A. Kaiser, M. Krieger, M. P. Scott, A. Bretscher, H. Ploegh, and P. Matsudaira. *Molecular Cell Biology*. W. H. Freeman, New York, NY, 2008.

[90] J. Lu, G. Getz, E. A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B. L. Ebert, R. H. Mak, A. A. Ferrando, J. R. Downing, T. Jacks, H. R. Horvitz, and T. R. Golub. MicroRNA expression profiles classify human cancers. *Nature*, 435(7043):834–838, 2005.

[91] S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: A survey. *IEEE ACM T. Comput Bi.*, 1(1):24–45, 2004.

[92] J. Malone, E. Holloway, T. Adamusiak, M. Kapushesky, J. Zheng, N. Kolesnikov, A. Zhukova, A. Brazma, and H. Parkinson. Modeling sample variables with an experimental factor ontology. *Bioinformatics*, 26(8):1112–1118, 2010.

[93] M. Mann and O. N. Jensen. Proteomic analysis of post-translational modifications. *Nat. Biotechnol.*, 21(3):255–261, 2003.

[94] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK, 2008.

[95] E. Meeds, Z. Ghahramani, R. M. Neal, and S. T. Roweis. Modeling dyadic data with binary latent factors. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 977–984, Cambridge, MA, 2007. The MIT Press.

[96] E. Meeds and S. Roweis. Nonparametric Bayesian biclustering. Technical report, University of Toronto, 2007.

[97] T. Minka. Expectation propagation for approximate bayesian inference. In J. S. Breese and D. Koller, editors, *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pages 361–369, San Francisco, CA, 2001. Morgan Kaufmann.

[98] T. Minka. Divergence measures and message passing. Technical report, Microsoft Research Ltd., 2005.

[99] D. C. Montgomery and G. C. Runger. *Applied Statistics and Probability for Engineers*. John Wiley and Sons, Inc., New York, NY, 3rd edition, 2003.

[100] V. K. Mootha, C. M. Lindgren, K. F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. R idderstrale, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander J. N. Hirschhorn, D. Altshuler, and L. C. Groop. Pgc-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, 34(3):244–245, 2003.

[101] J. S. Morey, J. C. Ryan, and F. M. Van Dolah. Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR. *Biol. Proced. Online*, 8(1):175–193, 2006.

[102] S. Myllykangas, J. Himberg, T. Böhling, B. Nagy, J. Hollmé, and S. Knuutila. DNA copy number amplification profiling of human neoplasms. *Oncogene*, 25(55):7324–7332, 2006.

[103] D. Nam and S.-Y. Kim. Gene-set approach for expression pattern analysis. *Brief. Bioinform.*, 9(3):189–197, 2008.

[104] R. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 355–368. The MIT Press, Cambridge, MA, 1999.

[105] R. M. Neal. Markov chain sampling methods for dirichlet process mixture models. *J. Comput. Graph. Stat.*, 9(2):249–265, 2000.

[106] The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, 2008.

[107] L. Page and S. Brin. The anatomy of a large-scale hypertextual web search engine. In P. W. Enslow Jr. and A. Ellis, editors, *Seventh International World-Wide Web Conference*, pages 107–117, Amesterdam, The Nehterlands, 1998. Elsevier Science Publishers B. V.

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

[109] S. Paruthiyil, A. Cvoro, M. Tagliaferri, I. Cohen, E. Shtivelman, and D. C. Leitman. Estrogen receptor β causes a g2 cell cycle arrest by inhibiting cdk1 activity through the regulation of cyclin b1, gadd45a, and btg2. *Breast Cancer Res. Treat.*, 129(3):777–784, 2011.

[110] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufman Publishers, San Mateo, CA, 1988.

[111] R. Peeters. The maximum edge biclique problem is NP-complete. *Discrete Appl. Math.*, 131(3):651–654, 2003.

[112] G. Pinton, E. Brunelli, B. Murer, R. Puntoni, M. Puntoni, D. A. Fennell, G. Gaudino, L. Mutti, and L. Moro. Estrogen receptor-β affects the prognosis of human malignant mesothelioma. *Cancer Res.*, 69(11):4598–4604, 2009.

[113] J. Pitman. *Combinatorial Stochastic Processes*. Springer-Verlag, New York, NY, 2002.

[114] A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129, 2006.

[115] C. E. Rasmussen. The infinite Gaussian mixture model. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12, pages 554–560, Cambridge, MA, 2004. The MIT Press.

[116] F. Reyal, M. H. van Vliet, N. J. Armstrong, H. M. Horlings, K. E. de Visser, M. Kok, A. E. Teschendorff, S. Mook, L. van't Veer, C. Caldas, R. J. Salmon, M. J. van de Vijver, and L. F. A. Wessels. A comprehensive analysis of prognostic signatures reveals the high predictive capacity of the proliferation, immune response and RNA splicing modules in breast cancer. *Breast Cancer Res.*, 10(6):R93, 2008.

[117] D. R. Rhodes, S. Kalyana-Sundaram, V. Mahavisno, R. Varambally, J. Yu, B. B. Briggs, T. R. Barrette, M. J. Anstet, C. Kincead-Beal, P. Kulkarni, S. Varambally, D. Ghosh, and A. M. Chinnaiyan. Oncomine 3.0: Gene, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia*, 9(2):166–180, 2007.

[118] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, New York, NY, 2005.

[119] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, Inc., Upper Saddle River, NJ, 2nd edition, 2003.

[120] A. Saha, J. Wittmeyer, and B. R. Cairns. Chromatin remodelling: the industrial revoluation of DNA around histones. *Nat. Rev. Mol. Cell Bio.*, 7(6):437–47, 2006.

[121] R. L. Schilsky. Personalized medicine in oncology: the future is now. *Nat. Rev. Drug. Discov.*, 9(5):363–366, 2010.

[122] T. D. Schmittgen and K. J. Livak. Analyzing real-time PCR data by the comparative c_t method. *Nat. Protoc.*, 3(6):1101–1108, 2008.

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

[124] A. Serin and M. Vingron. DeBi: Discovering differentially expressed biclusters using a frequent itemset approach. *Algorithm Mol. Biol.*, 6(1):18, 2011.

[125] J. Sethuraman. A constructive definition of Dirichlet priors. *Stat. Sinica*, 4(2):639–650, 1994.

[126] M. S. Srivastava and D. von Rosen. Regression models with unknown singular covariance matrix. *Linear Algebra and its Applications*, 354(1–3):255–273, 2002.

[127] L. D. Stein. Human genome: End of the beginning. *Nature*, 431(7011):915–916, 2004.

[128] M. Steyvers and T. Griffiths. Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, and W. Kintsch, editors, *Latent Semantic Analysis: A Road to Meaning*, pages 427–448. Lawrence Erlbaum, Hillsdale, NJ, 2007.

[129] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *P. Natl. Acad. Sci. U. S. A.*, 100(16):9440–9445, 2001.

[130] A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *P. Natl. Acad. Sci. U. S. A.*, 101(16):6062–6067, 2004.

[131] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *P. Natl. Acad. Sci. U. S. A.*, 102:15545–15550, 2005.

[132] C. Swanton and C. Caldas. From genomic landscapes to personalized cancer management — is there a roadmap. *Ann. N. Y. Acad. Sci.*, 1210(1):34–44, 2010.

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

[134] 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. *Mol. Syst. Biol.*, 1(1):2005.0002, 2005.

[135] Y. W. Teh. A hierarchical bayesian language model based on pitman-yor processes. In N. Calzolari, C. Cardie, and P. Isabelle, editors, *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 985–992, Stroudsburg, PA, 2006. Association for Computational Linguistics.

[136] Y. W. Teh and D. Görür. Indian buffet processes with power-law behavior. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1838–1846, Cambridge, MA, 2009. The MIT Press.

[137] Y. W. Teh, D. Görür, and Z. Ghahramani. Stick-breaking construction for the Indian buffet process. In M. Meila and X. Shen, editors, *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, pages 556–563, Madison, WI, 2007. Omnipress.

[138] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *J. Am. Stat. Assoc.*, 101(476):1566–1581, 2006.

[139] R. Thibaux and M. I. Jordan. Hierarchical Beta processes and the Indian buffet process. In M. Meila and X. Shen, editors, *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, pages 564–571, Madison, WI, 2007. Omnipress.

[140] J. P. Thiery and J. P. Sleeman. Complex networks orchestrate epithelial-mesenchymal transitions. *Nat. Rev. Mol. Cell Biol.*, 7(2):131–142, 2006.

[141] L. Tian, S. A. Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane, and P. J. Park. Discovering statistically significant pathways in expression profiling studies. *P. Natl. Acad. Sci. U. S. A.*, 102(38):13544–13549, 2005.

[142] R. Tibshirani, T. Hastie, M. Eisen, D. Ross, D. Botstein, and P. Brown. Clustering methods for the analysis of DNA microarray data. Technical report, Department of Health Research and Policy, Department of Genetics and Department of Biochemestry, Stanford University, 1999.

[143] Y. Tsuchiya, M. Nakajima, and T. Yokoi. Cytochrome p450-mediated metabolism of estrogens and its regulation in human. *Cancer Lett.*, 227(2):115–124, 2005.

[144] N. Ueda and R. Nakano. Deterministic annealing EM algorithm. *Neural Networks*, 11(2):271–282, 1998.

[145] M. van Uitert, W. Meuleman, and L. Wessels. Biclustering sparse binary genomic data. *J. Comput. Biol.*, 15(10):1329–1345, 2008.

[146] L. J. van't Veer and R. Bernards. Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature*, 452(7187):564–570, 2008.

[147] L. J. van't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, 2002.

[148] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *J. Mach. Learn. Res.*, 11:451–490, 2010.

[149] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

[150] D. G. Wang, J. B. Fan, C. J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, L. Hsie, T. Topaloglou, E. Hubbell, E. Robinson, M. Mittmann, M. S. Morris, N. Shen, D. Kilburn, J. Rioux, C. Nusbaum, S. Rozen, T. J. Hudson, R. Lipshutz, M. Chee, and E. S. Lander. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, 280(5366):1077–1082, 1998.

[151] Z. Wang, M. Gerstein, and M. Snyder. RNA-SEQ: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63, 2009.

[152] M. L. Wong and J. F. Medrano. Real-time PCR for mRNA quantitation. *Biotechniques*, 39(1):75–85, 2005.

[153] M. C. Wu, L. Zhang, Z. Wang, D. C. Christiani, and X. Lin. Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection. *Bioinformatics*, 25(9):1145–1151, 2009.

[154] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. In G. Lakemeyer and B. Nebel, editors, *Exploring Artificial Intelligence in the New Millenium*, pages 239–270. Morgan Kaufmann, San Francisco, CA, 2002.

[155] Y. Zhu, S. Davis, R. Stephens, P. S. Meltzer, and Y. Chen. GEOmetadb: Powerful alternative search engine for the gene expression omnibus. *Bioinformatics*, 24(23):2798–2800, 2008.

DISSERTATIONS IN INFORMATION AND COMPUTER SCIENCE

TKK-ICS-D19 Lahti, Leo.
Probabilistic Analysis of the Human Transcriptome with Side Information. 2010.

TKK-ICS-D20 Miche, Yoan.
Developing Fast Machine Learning Techniques with Applications to Steganalysis Problems. 2010.

TKK-ICS-D21 Sorjamaa, Antti.
Methodologies for Time Series Prediction and Missing Value Imputation. 2010.

TKK-ICS-D22 Schumacher, André
Distributed Optimization Algorithms for Multihop Wireless Networks. 2010.

Aalto-DD99/2011 Ojala, Markus
Randomization Algorithms for Assessing the Significance of Data Mining Results. 2011.

Aalto-DD111/2011 Dubrovin, Jori
Efficient Symbolic Model Checking of Concurrent Systems. 2011.

Aalto-DD118/2011 Hyvärinen, Antti
Grid Based Propositional Satisfiability Solving. 2011.

Aalto-DD136/2011 Brumley, Billy Bob
Covert Timing Channels, Caching, and Cryptography. 2011.

Aalto-DD11/2012 Vuokko, Niko
Testing the Significance of Patterns with Complex Null Hypotheses. 2012.

Aalto-DD19/2012 Reunanen, Juha
Overfitting in Feature Selection: Pitfalls and Solutions. 2012

Aalto-DD 33/2012

ISBN 978-952-60-4558-0
ISBN 978-952-60-4559-7 (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