TKK Dissertations in Information and Computer Science
Espoo 2010 TKK-ICS-D19

PROBABILISTIC ANALYSIS OF THE HUMAN TRANSCRIPTOME WITH SIDE INFORMATION

Leo Lahti

Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Faculty of Information and Natural Sciences for public examination and debate in Auditorium AS1 at the Aalto University School of Science and Technology (Espoo, Finland) on the 17th of December 2010 at 13 o'clock.

Aalto University School of Science and Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science

Aalto-yliopiston teknillinen korkeakoulu
Informaatio- ja luonnontieteiden tiedekunta
Tietojenkäsittelytieteen laitos

Distribution:

Aalto University School of Science and Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science
P.O.Box 15400
FI-00076 Aalto
FINLAND

Tel. +358-9-470 23272
Fax +358-9-470 23277
Email: series@ics.tkk.fi

ISBN 978-952-60-3367-9 (Print)
ISBN 978-952-60-3368-6 (Online)
ISSN 1797-5050 (Print)
ISSN 1797-5069 (Online)
URL: http://lib.tkk.fi/Diss/2010/isbn9789526033686/

Multiprint Oy
Espoo 2010

ABSTRACT

Lahti, L. (2010): **Probabilistic analysis of the human transcriptome with side information** Doctoral thesis, Aalto University School of Science and Technology, Dissertations in Information and Computer Science, TKK-ICS-D19, Espoo, Finland.

Keywords: data integration, exploratory data analysis, functional genomics, probabilistic modeling, transcriptomics

Recent advances in high-throughput measurement technologies and efficient sharing of biomedical data through community databases have made it possible to investigate the complete collection of genetic material, the genome, which encodes the heritable genetic program of an organism. This has opened up new views to the study of living organisms with a profound impact on biological research.

Functional genomics is a subdiscipline of molecular biology that investigates the functional organization of genetic information. This thesis develops computational strategies to investigate a key functional layer of the genome, the transcriptome. The time- and context-specific transcriptional activity of the genes regulates the function of living cells through protein synthesis. Efficient computational techniques are needed in order to extract useful information from high-dimensional genomic observations that are associated with high levels of complex variation. Statistical learning and probabilistic models provide the theoretical framework for combining statistical evidence across multiple observations and the wealth of background information in genomic data repositories.

This thesis addresses three key challenges in transcriptome analysis. First, new preprocessing techniques that utilize side information in genomic sequence databases and microarray collections are developed to improve the accuracy of high-throughput microarray measurements. Second, a novel exploratory approach is proposed in order to construct a global view of cell-biological network activation patterns and functional relatedness between tissues across normal human body. Information in genomic interaction databases is used to derive constraints that help to focus the modeling in those parts of the data that are supported by known or potential interactions between the genes, and to scale up the analysis. The third contribution is to develop novel approaches to model dependency between co-occurring measurement sources. The methods are used to study cancer mechanisms and transcriptome evolution; integrative analysis of the human transcriptome and other layers of genomic information allows the identification of functional mechanisms and interactions that could not be detected based on the individual measurement sources. Open source implementations of the key methodological contributions have been released to facilitate their further adoption by the research community.

TIIVISTELMÄ

Lahti, L. (2010): **Ihmisen geenien ilmentymisen ja taustatiedon tilastollinen mallitus** Väitöskirja, Aalto-yliopiston teknillinen korkeakoulu, Dissertations in Information and Computer Science, TKK-ICS-D19, Espoo, Suomi.

Avainsanat: aineistojen yhdistely, data-analyysi, toiminnallinen genomiikka, tilastollinen mallitus, geenien ilmentyminen

Mittausmenetelmien kehitys ja tutkimustiedon laajentunut saatavuus ovat mahdollistaneet ihmisen perimän eli genomin kokonaisvaltaisen tarkastelun. Tämä on avannut uusia näkökulmia biologiseen tutkimukseen ja auttanut ymmärtämään elämän syntyä ja rakennetta uusin tavoin. Toiminnallinen genomiikka on molekyylibiologian osa-alue, joka tutkii perimän toiminnallisia ominaisuuksia. Perimän toimintaan liittyvää mittausaineistoa on runsaasti saatavilla, mutta korkeaulotteisiin mittauksiin liittyy monimutkaisia ja tuntemattomia taustatekijöitä, joiden huomiointi mallituksessa on haasteellista. Tehokkaat laskennalliset menetelmät ovat avainasemassa pyrittäessä jalostamaan uusista havainnoista käyttökelpoista tietoa.

Tässä väitöskirjassa on kehitetty yleiskäyttöisiä laskennallisia menetelmiä, joilla voidaan tutkia ihmisen geenien ilmentymistä koko perimän tasolla. Geenien ilmentyminen viittaa lähetti-RNA-molekyylien tuottoon solussa perimän sisältämän informaation nojalla. Tämä on keskeinen perinnöllisen informaation säätelytaso, jonka avulla solu säätelee proteiinien tuottoa ja solun toimintaa ajasta ja tilanteesta riippuen. Tilastollinen oppiminen ja todennäköisyyksin perustuva probabilistinen mallitus tarjoavat teoreettisen kehyksen, jonka avulla rinnakkaisiin mittauksiin ja taustatietoihin sisältyvää informaatiota voidaan käyttää kasvattamaan mallien tilastollista voimaa. Kehitetyt menetelmät ovat yleiskäyttöisiä laskennallisen tieteen tutkimusvälineitä, jotka tekevät vähän, mutta selkeästi ilmaistuja mallitusoletuksia ja sietävät korkeaulotteisiin toiminnallisen genomiikan havaintoaineistoihin sisältyviä epävarmuuksia.

Väitöskirjassa kehitetyt menetelmät tarjoavat ratkaisuja kolmeen keskeiseen mallitusongelmaan toiminnallisessa genomiikassa. Luotettavien esikäsittelymenetelmien kehittäminen on työn ensimmäinen päätulos, jossa tietokantoihin sisältyvää taustatietoa käytetään perimänlaajuisten mittausaineistojen epävarmuuksien vähentämiseksi. Toisena päätuloksena väitöskirjassa kehitetään uusi aliavaruuskasautukseen perustuva menetelmä, jonka avulla voidaan tutkia ja kuvata solubiologisen vuorovaikutusverkon käyttäytymistä kokonaisvaltaisesti ihmiskehon eri osissa. Taustatietoa geenien vuorovaikutuksista käytetään ohjaamaan ja nopeuttamaan mallitusta. Menetelmällä saadaan uutta tietoa geenien säätelystä ja kudosten toiminnallisista yhteyksistä. Kolmanneksi väitöskirjatyössä kehitetään uusia menetelmiä perimänlaajuisten mittausaineistojen yhdistelyyn. Ihmisen geenien ilmentymisen ja muiden aineistojen riippuvuuksien mallitus mahdollistaa sellaisten toiminnallisten yhteyksien ja vuorovaikutusten havaitsemisen, joiden tutkimiseksi yksittäiset havaintoaineistot ovat riittämättömiä. Aineistojen yhdistelyyn kehitettyjä menetelmiä sovelletaan syöpämekanismien ja lajien välisten eroavaisuuksien tutkimiseen. Julkaistuilla avoimen lähdekoodin toteutuksilla on pyritty varmistamaan kehitettyjen menetelmien saatavuus ja laajempi käyttöönotto laskennallisen biologian tutkimuksessa.

Contents

Preface

This work has been carried out at the Neural Networks Research Centre and Adaptive Informatics Research Centre of the Laboratory of Computer and Information Science (Department of Information and Computer Science since 2008), Helsinki University of Technology, i.e., as of 2010 the Aalto University School of Science and Technology. Part of the work was done at the Department of Computer Science, University of Helsinki, when I was visiting there for a year in 2005. I am also pleased to having had the opportunity to be a part of the Helsinki Institute for Information Technology HIIT. The work has been supported by the Graduate School of Computer Science and Engineering, as well as by project funding from the Academy of Finland through the SYSBIO program and from TEKES through the MultiBio research consortium. The Graduate School in Computational Biology, Bioinformatics, and Biometry (ComBi) has supported my participation to scientific conferences and workshops abroad during the thesis work.

I wish to thank my supervisor, professor Samuel Kaski for giving me the opportunity to work in a truly interdisciplinary research field with the freedom and responsibilities of scientific work, and with the necessary amount of guidance. These have been essential parts of the learning process.

I would also like to express my gratitude to the reviewers of this thesis, Professor Juho Rousu and Doctor Simon Rogers for their expert feedback.

Research on computational biology has given me the excellent opportunity to work with and learn from experts in two traditionally distinct disciplines, computational science and genome biology. I am particularly grateful to professor Sakari Knuutila for his enthusiasm, curiosity, and personal example in collaboration and daily research work. Researchers in the Laboratory of Cytomolecular Genetics at the Haartman Institute have provided a friendly and inspiring environment for active collaboration during the last years.

My sincere compliments belong to all of my other co-authors, in particular to Tero Aittokallio, Laura Elo-Uhlgren, Jaakko Hollmén, Juha Knuuttila, Samuel Myllykangas and Janne Nikkilä. It has been a pleasure to work with you, and your contributions extend beyond what we wrote together. I would also like to thank the former and present members of the MI research group for working beside me through these years, as well as for intriguing discussions about science and life in general. I would also like to thank the personnel of the ICS department, in particular professors Erkki Oja and Olli Simula, who have helped to provide an excellent academic research environment, as well as our secretaries Tarja Pihamaa and Leila Koivisto, and Markku Ranta and Miki Sirola, who have given valuable help in so many practical matters during the years.

Science is a community effort. Open sharing of ideas, knowledge, publication

material, data, software, code, experiences and emotions has had a tremendous impact to this thesis. I will express my sincere gratitude to the community by continued participation and contributions.

I would also like to thank my earliest scientific advisors; Reijo, who brought me writings about the chemistry of life and helped me to grow bacteria and prepare space dust in the 1980's, Pekka, who has demonstrated the power of criticism and emphasized that natural science has to be exact, Tapio, for the attitude that maths can be just fun, and Risto, for showing how rational thinking can be applied also in real life. Thanks also go to my science friends, Manu and Ville; we have shared the passion for natural science, and I want to thank you for our continuous and inspiring discussions along the way. I am grateful to my grandfather Osmo, who shared with me the wonder towards life, science, and humanities, and was willing to discuss it all through days and nights when I was a child, questioning himself the self-evident truths again and again, remaining as puzzled as I was. And for Alli and Arja, my grandmothers, for their understanding, and all support and love.

My Friends. With you I have explored other facets of nature, science, and life... Thank you for staying with me through all these years and sharing so many aspects of curiosity, exploration and mutual understanding.

Finally, I am grateful to my parents and sister, Pipsa, Kari, and Tuuli. You have accepted me and loved me, supported me on the paths that I have chosen to follow, and understood that freedom can create the strongest ties.

Cambridge, November 23, 2010
Leo Lahti

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.

1. Laura L. Elo, Leo Lahti, Heli Skottman, Minna Kyläniemi, Riitta Lahesmaa, and Tero Aittokallio. Integrating probe-level expression changes across generations of Affymetrix arrays. *Nucleic Acids Research*, 33(22):e193, 2005.

2. Leo Lahti, Laura L. Elo, Tero Aittokallio, and Samuel Kaski. Probabilistic analysis of probe reliability in differential gene expression studies with short oligonucleotide arrays. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(1):217–225, 2011.

3. Leo Lahti, Juha E.A. Knuuttila, and Samuel Kaski. Global modeling of transcriptional responses in interaction networks. *Bioinformatics*, 26(21):2713–2720, 2010.

4. Leo Lahti, Samuel Myllykangas, Sakari Knuutila, and Samuel Kaski. Dependency detection with similarity constraints. In Tülay Adali, Jocelyn Chanussot, Christian Jutten, and Jan Larsen, editors, *Proceedings of the 2009 IEEE International Workshop on Machine Learning for Signal Processing XIX*, pages 89–94. IEEE, Piscataway, NJ, 2009.

5. Janne Sinkkonen, Janne Nikkilä, Leo Lahti, and Samuel Kaski. Associative clustering. In Boulicaut, Esposito, Giannotti, and Pedreschi (editors), *Machine Learning: ECML2004 (Proceedings of the ECML'04, 15th European Conference on Machine Learning)*, Lecture Notes in Computer Science 3201, 396–406. Springer, Berlin, 2004.

6. Samuel Kaski, Janne Nikkilä, Janne Sinkkonen, Leo Lahti, Juha E.A. Knuuttila, and Cristophe Roos. Associative clustering for exploring dependencies between functional genomics data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics: Special Issue on Machine Learning for Bioinformatics – Part 2*, 2(3):203–216, 2005.

SUMMARY OF PUBLICATIONS AND THE AUTHOR'S CONTRIBUTION

The publications in this thesis have been a joint effort of all authors; key contributions by the author of this thesis are summarized below.

Publication 1 introduces a novel analysis strategy to improve the accuracy and reproducibility of the measurements in genome-wide transcriptional profiling studies. A central part of the approach is the utilization of side information in external genome sequence databases. The author participated in the design of the study, suggested the utilization of external sequence data, implemented this, as well as participated in preparing the manuscript.

Publication 2 provides a probabilistic framework for probe-level gene expression analysis. The model combines statistical power across multiple microarray experiments, and is shown to outperform widely-used preprocessing methods in differential gene expression analysis. The model provides tools to assess probe performance, which can potentially help to improve probe and microarray design. The author had a major role in designing the study. The author derived the formulation, implemented the model, performed the probe-level experiments, as well as coordinated the manuscript preparation. The author prepared an accompanied open source implementation which has been published in BioConductor, a reviewed open source repository for computational biology algorithms.

Publication 3 introduces a novel approach for organism-wide modeling of transcriptional activity in genome-wide interaction networks. The method provides tools to analyze large collections of genome-wide transcriptional profiling data. The author had a major role in designing the study. The author implemented the algorithm, performed the experiments, as well as coordinated the manuscript preparation. The author participated in and supervised the preparation of an accompanied open source implementation in BioConductor.

Publication 4 introduces a regularized dependency modeling framework with particular applications in cancer genomics. The author had a major role in formulating the biomedical modeling task, and in designing the study. The theoretical model was jointly developed by the author and S. Kaski. The author derived and implemented the model, carried out the experiments, and coordinated the manuscript preparation. The author supervised and participated in the preparation of an accompanied open source implementation in BioConductor.

Publication 5 introduces the associative clustering principle, which is a novel data integration framework for dependency detection with direct applications in functional genomics. The author participated in implementation of the method, had the main responsibility in designing and performing the functional genomics experiments, as well as participated in preparing the manuscript.

Publication 6 contains the most extensive treatment of the associative clustering principle. In addition to presenting detailed theoretical considerations, this work introduces new sensitivity analysis of the results, and provides a comprehensive validation in bioinformatics case studies. The author participated in designing the experiments, performed the comparative functional genomics experiments and technical validation, as well as participated in preparing the manuscript.

LIST OF ABBREVIATIONS AND SYMBOLS

In this thesis boldface symbols are used to denote matrices and vectors. Capital symbols ($\mathbf{X}$) signify matrices and lowercase symbols ($\mathbf{x}$) column vectors. *Normal lowercase* symbols indicate scalar variables.

$\mathbb{R}$	Real domain
$\mathbf{X}, \mathbf{Y}$	Data matrices ($D \times N$)
$[\mathbf{X}; \mathbf{Y}]$	Concatenated data
$\mathbf{x}$, $\mathbf{y}$	Data samples, vectors in $\mathbb{R}^D$
x, y	Scalars in $\mathbb{R}$
$\mathcal{X}, \mathcal{Y}$	Random variables
$\mathbf{I}$	Identity matrix
$\mathbf{\Sigma}, \mathbf{\Psi}$	Covariance matrices
$p(\mathbf{x})$	Probability or probability density of $\mathcal{X}$
$p(\mathbf{X})$	Likelihood
$\mathbb{E}[\cdot]$	Expectation
$\|\cdot\|$	Norm of a matrix or vector
Tr	Matrix trace
$I(\mathcal{X}; \mathcal{Y})$	Mutual information between random variables $\mathcal{X}$ and $\mathcal{Y}$
$\mathrm{Beta}(\alpha, \beta)$	Beta distribution with parameters α and β
$\mathrm{Dir}(\boldsymbol{\theta})$	Dirichlet distribution with parameter vector $\boldsymbol{\theta}$
$\mathrm{IG}(\alpha, \beta)$	Inverse Gamma distribution with parameters α and β
$\mathrm{Mult}(N, \boldsymbol{\theta})$	Multinomial distribution with sample size N and parameter vector $\boldsymbol{\theta}$
$N(\boldsymbol{\mu}, \mathbf{\Sigma})$	Normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$
AC	Associative clustering
aCGH	Array Comparative genomics hybridization
CCA	Canonical correlation analysis
cDNA	Complementary DNA
DNA	Deoxyribonucleic acid
DP	Dirichlet process
EM	Expectation – Maximization algorithm
IB	Information bottleneck
KL–divergence	Kullback-Leibler divergence
MAP	Maximum a posteriori
MCMC	Markov chain Monte Carlo
ML	Maximum likelihood
mRNA	Messenger-RNA
tRNA	Transfer-RNA
PCA	Principal component analysis
RNA	Ribonucleic acid

Chapter 1

Introduction

Revolutions in measurement technologies have led to revolutions in science and society. Introduction of the microscope in the 17th century opened a new view to the world of living organisms and enabled the study of life processes at cellular level. Since then, new techniques have been developed to investigate ever smaller objects. The discovery of the molecular structure of the DNA in 1953 (Watson and Crick, 1953) led to the establishment of genes as fundamental units of genetic information that is passed on between generations. The draft sequence of the human genome, covering three billion DNA base pairs, was published in 2001 (International human genome sequencing consortium, 2001; Venter et al., 2001). Modern measurement technologies provide researchers with large volumes of data concerning the structure, function, and interactions of genes and their products. Rapid accumulation of genomic data in shared community databases has accelerated biological research (Cochrane and Galperin, 2010), but the structural and functional organization of genetic information is still poorly understood. While functional roles of individual genes have been characterized, little is known regarding the higher-level regularities and interactions from which the complexity and diversity of life emerges. The quest for systems-level understanding of genome function is a major paradigm in modern biology (Collins et al., 2003).

Computational science has a key role in transforming the genomic data collections into new biological knowledge (Cohen, 2004). New observations allow the formulation of new research questions, but also bring new challenges (Barbour et al., 2005). The sheer size of high-throughput data sets makes them incomprehensible for human mind, and the complexity of biological phenomena and high levels of uncontrolled variation set specific challenges for computational analysis (Tilstone, 2003; Troyanskaya, 2005). Filtering relevant information from statistically uncertain high-dimensional data is a challenging task where new computational methods are needed to organize and summarize the overwhelming volumes of observational data into a comprehensible form to make new discoveries about the structure of life; computation is a new microscope for studying massive data sets.

This thesis develops principled exploratory methods to investigate the *human transcriptome*. It is a central functional layer of the genome and a significant source of phenotypic variation. The transcriptome refers to the complete collection of messenger-RNA transcripts of an organism. The essentially static genome sequence regulates the time- and context-specific patterns of transcriptional ac-

tivity of the genes, and subsequently the function of living cells through protein synthesis. An average cell contains over 300,000 mRNA molecules and the expression levels of individual genes span 4-5 orders of magnitude (Carninci, 2009). A wealth of associated genomic information resources are available in public repositories (Cochrane and Galperin, 2010). By combining heterogeneous information sources and utilizing the wealth of background information in public repositories, it is possible to solve some of the problems that are related to the statistical uncertainties and small sample size of individual data sets, as well as to form a holistic picture of the genome (Huttenhower and Hofmann, 2010).

The observational data can provide the starting point to discover novel research hypotheses of poorly characterized large-scale systems; the analysis proceeds from general observations of the data toward more detailed investigations and hypotheses. This differs from traditional hypothesis testing where the investigation proceeds from hypotheses to measurements that target particular research questions, in order to support or reject a given hypothesis. *Exploratory data analysis* refers to the use of computational tools to summarize and visualize the data in order to identify potentially interesting structure, and to facilitate the generation of new research hypotheses when the search space would be otherwise exhaustively large (Tukey, 1977). When the system is poorly characterized, there is a need for methods that can adapt to the data and extract features in an automated way. This is useful since application-oriented models often require careful preprocessing of the data and a timely model fitting process. They may also require prior knowledge of the investigated system, which is often not available. *Statistical learning* investigates solutions to these problems.

1.1 Contributions and organization of the thesis

This thesis introduces computational strategies for genome- and organism-wide analysis of the human transcriptome. The thesis provides novel tools (i) to increase the reliability of high-throughput microarray measurements by combining statistical evidence from genome sequence databases and across multiple microarray experiments, (ii) to model context-specific transcriptional activation patterns of genome-scale interaction networks across normal human body by using background information of genetic interactions to guide the analysis, and (iii) to integrate measurements of the human transcriptome to other layers of genomic information with novel dependency modeling techniques for co-occurring data sources. The three strategies address widely recognized challenges in functional genomics (Collins et al., 2003; Troyanskaya, 2005).

Obtaining reliable measurements is the crucial starting point for any data analysis task. The first contribution of this thesis is to develop computational strategies that utilize side information in genomic sequence and microarray data collections in order to reduce noise and improve the quality of high-throughput observations. Publication 1 introduces a probe-level strategy for microarray preprocessing, where updated genomic sequence databases are used in order to remove erroneously targeted probes to reduce measurement noise. The work is extended in Publication 2, which introduces a principled probabilistic framework for probe-level analysis. A generative model for probe-level observations combines evidence across multiple experiments, and allows the estimation of probe performance directly from microarray measurements. The model detects a large number of unreliable probes

contaminated by known probe-level error sources, as well as many poorly performing probes where the source of contamination is unknown and could not be controlled based on existing probe-level information. The model provides a principled framework to incorporate prior information of probe performance. The introduced algorithms outperform widely used alternatives in differential gene expression studies.

A novel strategy for organism-wide analysis of transcriptional activity in genome-scale interaction networks in Publication 3 forms the second main contribution of this thesis. The method searches for local regions in a network exhibiting coordinated transcriptional response in a subset of conditions. Constraints derived from genomic interaction databases are used to focus the modeling on those parts of the data that are supported by known or potential interactions between the genes. Nonparametric inference is used to detect a number of physiologically coherent and reproducible transcriptional responses, as well as context-specific regulation of the genes. The findings provide a global view on transcriptional activity in cell-biological networks and functional relatedness between tissues.

The third contribution of the thesis is to integrate measurements of the human transcriptome to other layers of genomic information. Novel dependency modeling techniques for co-occurrence data are used to reveal regularities and interactions, which could not be detected in individual observations. The regularized dependency modeling framework of Publication 4 is used to detect associations between chromosomal mutations and transcriptional activity. Prior biological knowledge is used to constrain the latent variable model and shown to improve cancer gene detection performance. The associative clustering, introduced in Publications 5 and 6, provides tools to investigate evolutionary divergence of transcriptional activity.

Open source implementations of the key methodological contributions of this thesis have been released in order to guarantee wide access to the developed algorithmic tools and to comply with the emerging standards of transparency and reproducibility in computational science, where an increasing proportion of research details are embedded in code and data accompanying traditional publications (Boulesteix, 2010; Carey and Stodden, 2010; Ioannidis et al., 2009) and transparent sharing of these resources can form valuable contributions to public knowledge (Sommer, 2010; Sonnenburg et al., 2007; Stodden, 2010).

The thesis is organized as follows: In Chapter 2, there is an overview of functional genomics, related measurement techniques, and genomic data resources. General methodological background, in particular of exploratory data analysis and the probabilistic modeling paradigm, is provided in Chapter 3. The methodological contributions of the thesis are presented in Chapters 4-6. In Chapter 4, strategies to improve the reliability of high-throughput microarray measurements are presented. In Chapter 5 methods for organism-wide analysis of the transcriptome are considered. In Chapter 6, two general-purpose algorithms for dependency modeling are introduced and applied in investigating functional effects of chromosomal mutations and evolutionary divergence of transcriptional activity. The conclusions of the thesis are summarized in Chapter 7.

Chapter 2

Functional genomics

> *From all we have learnt about the structure of living matter, we must be prepared to find it working in a manner that cannot be reduced to the ordinary laws of physics - - because the construction is different from anything we have yet tested in the physical laboratory.*
>
> E. Schrödinger (1956)

Living organisms are controlled not only by natural laws but also by inheritable *genetic programs* (Mayr, 2004; Schrödinger, 1944). Such *double causation* is a unique feature of life, and in fundamental contrast to purely physical processes of the inanimate world. Life may have emerged on earth more than 3.4 billion years ago (Schopf, 2006; Tice and Lowe, 2004). Genetic information evolves by means of *natural selection* (Darwin, 1859). Living organisms maintain homeostasis, adapt to changing environments, respond to external stimuli, and communicate. Peculiar features of living systems include metabolism, growth and hierarchical organization, as well as the ability to replicate and reproduce. All known life forms share fundamental mechanisms at molecular level, which suggests a common evolutionary origin of the living organisms.

The complete collection of genetic material, *the genome*, encodes the heritable genetic program of an organism. Advances in measurement technology and computational science have opened up new views to the large-scale organization of the genome (Carroll, 2003; Lander, 1996). *Functional genomics* is a subdiscipline of molecular biology investigating the functional organization and properties of genetic information. In this thesis, new computational approaches are developed for investigation of a central functional layer of the genome of our own species, the human *transcriptome*. This chapter gives an overview to the relevant concepts in genome biology in eukaryotic organisms and associated genomic data resources. For further background in molecular genome biology, see Alberts et al. (2002); Brown (2006).

2.1 Universal genetic code

Cells are fundamental building blocks of living organisms. All known life forms maintain a carbon-based cellular form that carries the genetic program (Alberts et al., 2002). Each cell carries a copy of the heritable genetic code, *the genome*.

The human genome is divided in 23 pairs of *chromosomes*, located in the nucleus of the cell, as well as in additional mitochondrial genome. Chromosomes are macroscopic deoxyribonucleic acid (DNA) molecules in which the DNA is wrapped around *histone* molecules and packed into a peculiar *chromatin* structure that will ultimately constitute chromosomes. The *genetic code* in the DNA consists of four *nucleotides*: adenosine (A), thymine (T), guanine (G), and cytosine (C). In ribonucleic acid (RNA), the thymine is replaced by uracil (U). Ordering of the nucleotides carries genetic information. Nucleic acid sequences have a peculiar base pairing property, where only A-T/U and G-C pairs can hybridize with each other. This leads to the well-known double-stranded structure of the DNA, and forms the basis for cellular information processing. The *central dogma of molecular biology* (Crick, 1970) states that DNA encodes the information to construct proteins through the irreversible process of *protein synthesis*. This is a central paradigm in molecular biology, describing the functional organization of life at the cellular level.

2.1.1 Protein synthesis

Genes are basic units of genetic information. The gene is a sequence of DNA that contains the information to manufacture a protein or a set of related proteins. Genetic variation and regulation of gene activity has therefore major phenotypic consequences. The *regulatory region* and *coding sequence* are two key elements of a gene. The regulatory region regulates gene activity, while the coding sequence carries the instructions for protein synthesis (Alberts et al., 2002). Interestingly, the concept of a gene remains controversial despite comprehensive identification of the protein-coding genes in the human genome and detailed knowledge of their structure and function (Pearson, 2006).

Proteins, encoded by the genes, are key functional entities in the cell. They form cellular structures, and participate in cell signaling and functional regulation. *Protein synthesis* refers to the cell-biological process that converts genetic information into final functional protein products (Figure 2.1.1A). Key steps in protein synthesis include transcription, pre-mRNA splicing, and translation. In *transcription*, the double-stranded DNA is opened in a proximity of the gene sequence and the process is initiated on the regulatory region of the gene. The DNA sequence of the gene is then converted into a complementary pre-mRNA by a polymerase enzyme. The pre-mRNA sequence contains both protein coding and non-coding segments. These are called *exons* and *introns*, respectively. In *pre-mRNA splicing*, the introns are removed and the exons are joined together to form mature *messenger-RNA (mRNA)*. A gene can encode multiple splice variants, corresponding to different exon definitions and their combinations; this is called *alternative splicing*. The mature mRNA is exported from nucleus to the cell cytoplasm. In *translation* the mRNA is converted into a corresponding amino acid sequence in ribosomes based on the *universal genetic code* that defines a mapping between nucleic acid triplets, so-called *codons*, and amino acids. The code is common for all known life forms. Each consecutive codon on the mRNA sequence corresponds to an amino acid, and the corresponding sequence of amino acids constitutes a protein. In the final stage of protein synthesis, the amino acid sequence folds into a three-dimensional structure and undergoes *post-translational modifications*. The structural characteristics of a protein molecule will ultimately determine its functional properties (Alberts et al., 2002).

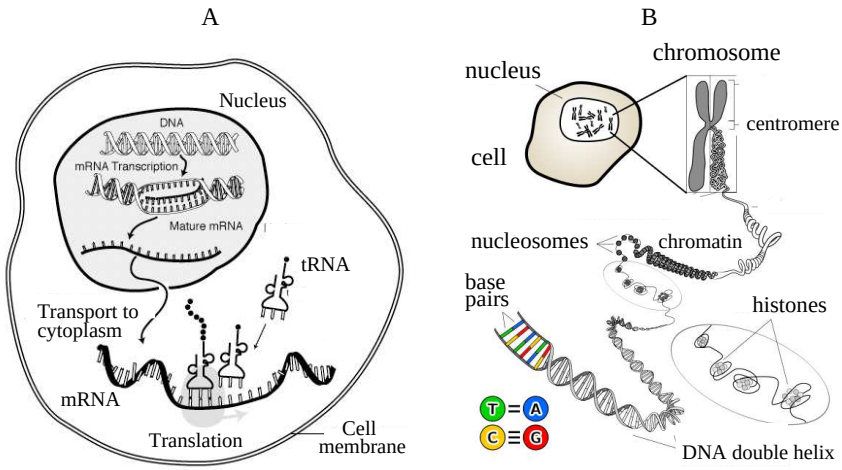


Figure 2.1: **A** Key steps of protein synthesis. The two key processes in protein synthesis are called *transcription* and *translation*, respectively. In transcription, the DNA sequence of the gene is transcribed into pre-mRNA based on the base pairing property of nucleic acid sequences. The pre-mRNA is modified to produce mature messenger-RNA (mRNA), which is then transported to cytoplasm. Transfer-RNA (tRNA) carries the mRNA to ribosomes, where it is translated into an amino acid sequence based on the universal genetic code where each nucleotide triplet of the mRNA sequence, so-called *codon*, corresponds to a particular amino acid. The amino acid sequence is subsequently modified to form the final functional protein product. **B** Organization of the genetic material in an eukaryotic cell. The nucleotide base pairs form the double helix structure of DNA. This is wrapped around histone molecules to form nucleosomes, and the chromatin sequence. The chromatin is tightly packed to form chromosomes that carry the genetic material and are located in the cell nucleus. The image has been modified from http://commons.wikimedia.org/wiki/File:Chromosome_en.svg.

2.1.2 Layers of regulation

Phenotypic changes can rarely be attributed to changes in individual genes; cell function is ultimately determined by coordinated activation of genes and other biomolecular entities in response to changes in cell-biological environment (Hartwell et al., 1999). Gene activity is regulated at all levels of protein synthesis and cellular processes. A major portion of functional genome sequence and protein coding-genes themselves participate in the regulatory system itself (Lauffenburger, 2000).

Epigenetic regulation refers to chemical and structural modifications of chromosomal DNA, the *chromatin*, for instance through methylation, acetylation, and other histone-binding molecules. Such modifications affect the packing of the DNA molecule around *histones* in the cell nucleus. The combinatorial regulation of such modifications regulates access to the gene sequences (Gibney and Nolan, 2010). Epigenetic changes are believed to be heritable and they constitute a major source of variation at individual and population level (Johnson and Tricker, 2010). *Transcriptional regulation* is the next major regulatory layer in protein synthesis. So-called *transcription factor* proteins can regulate the transcription rate by binding to control elements in gene regulatory region in a combinatorial fashion. *Post-transcriptional modifications* will then regulate pre-mRNA splicing. Up to 95% of human multi-exon genes are estimated to have alternative splice variants (Pan et al., 2008). Consequently, a variety of related proteins can be encoded by a single gene. This contributes to the structural and functional diversity of cell function

(Stetefeld and Ruegg, 2005). Several mechanisms will then affect *mRNA degradation* rates. For instance, micro-RNAs that are small, 21-25 basepair nucleotide sequences can inactivate specific mRNA transcripts through complementary base pairing, leading to mRNA degradation, or prevention of translation. Finally, *post-translational modifications*, *protein degradation*, and other mechanisms will affect the three-dimensional structure and life cycle of a protein. The proteins will participate in further cell-biological processes. The processes are in continuous interaction and form complex functional networks, which regulate the life processes of an organism (Alberts et al., 2002).

2.2 Organization of genetic information

The understanding of the structure and functional organization of the genome is rapidly accumulating with the developing genome-scanning technologies and computational methods. This section provides an overview to key structural and functional layers of the human genome.

2.2.1 Genome structure

The genome is a dynamic structure, organized and regulated at multiple levels of resolution from individual nucleotide base pairs to complete chromosomes (Figure 2.1.1B; Brown (2006)). A major portion of heritable variation between individuals has been attributed to differences in the genomic DNA sequence. Traditionally, main genetic variation was believed to arise from small point mutations, so-called *single-nucleotide polymorphisms (SNPs)*, in protein-coding DNA. Recently, it has been increasingly recognized that *structural variation* of the genome makes a remarkable contribution to genetic variation. Structural variation is observed at all levels of organization from single-nucleotide polymorphisms to large chromosomal rearrangements, including deletions, insertions, duplications, copy-number variants, inversions and translocations of genomic regions (Feuk et al., 2006; Sharp et al., 2006). Such modifications can directly and indirectly influence transcriptional activity and contribute to human diversity and health (Collins et al., 2003; Hurles et al., 2008).

The draft DNA sequence of the complete human genome was published in 2001 (International human genome sequencing consortium, 2001; Venter et al., 2001). The human genome contains three billion base pairs and approximately 20,000-25,000 protein-coding genes (International Human Genome Sequencing Consortium, 2004). The protein-coding exons comprise less than 1.5% of the human genome sequence. Approximately 5% of the human genome sequence has been conserved in evolution for more than 200 million years, including the majority of protein-coding genes (The ENCODE Project Consortium, 2007; Mouse Genome Sequencing Consortium, 2002). Half of the genome consists of highly repetitive sequences. The genome sequence contains structural elements such as centromeres and telomeres, repetitive and mobile elements, (Prak and Kazazian Jr., 2000), retroelements (Bannert and Kurth, 2004), and non-coding, non-repetitive DNA (Collins et al., 2003). The functional role of intergenic DNA, which forms 75% of the genome, is to a large extent unknown (Venter et al., 2001). Recent evidence suggests that the three-dimensional organization of the chromosomes, which is to a large extent regulated by the intergenic DNA is under active selection, can have

a remarkable regulatory role (Lieberman-Aiden et al., 2009; Parker et al., 2009). Comparison of the human genome with other organisms, such as the mouse (Mouse Genome Sequencing Consortium, 2002) can highlight important evolutionary differences between species. For a comprehensive review of the structural properties of the human genome, see Brown (2006).

2.2.2 Genome function

In protein synthesis, the gene sequence is transcribed into pre-mRNA, which is then further modified into mature messenger-RNA and transported to cytoplasm. An average cell contains over 300,000 mRNA molecules, and the mRNA concentration, or *expression levels* of individual genes, vary according to Zipf's law, a power-law distribution where most genes are expressed at low concentrations, perhaps only one or few copies of the mRNA per cell on average, and a small number of genes are highly expressed, potentially with thousands of copies per cell (see Carninci, 2009; Furusawa and Kaneko, 2003). Cell-biological processes are reflected at the transcriptional level. Transcriptional activity varies by cell type, environmental conditions and time. Different collections of genes are active in different contexts. *Gene expression*, or mRNA expression, refers to the expression level of an mRNA transcript at particular physiological condition and time point. In addition to protein-coding mRNA molecules that are the main target of analysis in this thesis, the cell contains a variety of other functional and non-functional mRNA transcripts, for instance micro-RNAs, ribosomal RNA and transfer-RNA molecules (Carninci, 2009; Johnson et al., 2005).

The *transcriptome* refers to the complete collection of mRNA sequences of an organism. This is a central functional layer of the genome that regulates protein production in the cells, with a significant role in creating genetic variation (Jordan et al., 2005). According to current estimates, up to 90% of the eukaryotic genome can be transcribed (Consortium, 2005; Gagneur et al., 2009). The protein-coding mRNA transcripts are translated into proteins at ribosomes during protein synthesis.

The proteome refers to the collection of protein products of an organism. The proteome is a main functional layer of the genome. Since the final protein products carry out a main portion of the actual cell functions, techniques for monitoring the concentrations of all proteins and their modified forms in a cell simultaneously would significantly help to improve the understanding of the cellular systems (Collins et al., 2003). However, sensitive, reliable and cost-efficient genome-wide screening techniques for measuring protein expression are currently not available. Therefore genome-wide measurements of the mRNA expression levels are often used as an indirect estimate of protein activity.

In addition to the DNA, RNA and proteins, the cell contains a variety of other small molecules. The extreme functional diversity of living organisms emerges from the complex network of interactions between the biomolecular entities (Barabási and Oltvai, 2004; Hartwell et al., 1999). Understanding of these networks and their functional properties is crucial in understanding cell function (Collins et al., 2003; Schadt, 2009). However, the systemic properties of *the interactome* are poorly characterized and understood due to the complexity of biological phenomena and incomplete information concerning the interactions. The cell-biological processes are inherently modular (Hartwell et al., 1999; Ihmels et al., 2002; Lauffenburger, 2000), and they exhibit complex *pathway cross-talk* between the cell-biological

processes (Li et al., 2008). In modular systems, small changes can have significant regulatory effects (Espinosa-Soto and Wagner, 2010).

2.3 Genomic data resources

Systematic observations from the various functional and regulatory layers of the genome are needed to understand cell-biological systems. Efficient sharing and integration of genomic information resources through digital media has enabled large-scale investigations that no single institution could afford. The public human genome sequencing project (International human genome sequencing consortium, 2001) is a prime example of such project. Results from genome-wide transcriptional profiling studies are routinely deposited to public repositories (Barrett et al., 2009; Parkinson et al., 2009). Sharing of original data is increasingly accepted as the scientific norm, often following explicit data release policies. The establishment of large-scale databases and standards for representing biological information support the efficient use of these resources (Bammler et al., 2005; Brazma et al., 2006). A continuously increasing array of genomic information is available in these databases, concerning aspects of genomic variability across individuals, disease states, and species (Brent, 2008; Church, 2005; Cochrane and Galperin, 2010; G10KCOS consortium, 2009; The Cancer Genome Atlas Research Network, 2008).

2.3.1 Community databases and evolving biological knowledge

Genomic sequence databases

During the human genome project and preceding sequencing projects DNA sequence reads were among the first sources of biological data that were collected in large-scale public repositories, such as GenBank (Benson et al., 2010). GenBank contains comprehensive sequence information of genomic DNA and RNA for a number of organisms, as well as a variety of information concerning the genes, non-coding regions, disease associations, variation and other genomic features. Online analysis tools, such as the Ensembl Genome browser (Flicek et al., 2010), facilitate efficient use of these annotation resources. Next-generation sequencing technologies provide rapidly increasing sequencing capacity to investigate sequence variation between individuals, populations and disease states (Ledford, 2010; McPherson, 2009). In particular, the human and mouse transcriptome sequence collections at the Entrez Nucleotide database of GenBank are utilized in this thesis, in Publications 1 and 2.

Transcriptome databases

Gene expression measurement provides a snapshot of mRNA transcript levels in a cell population at a specific time and condition, reflecting the activation patterns of the various cell-biological processes. While gene expression measurements provide only an indirect view to cellular processes, their wide availability provides a unique resource for investigating gene co-regulation on a genome- and organism-wide scale. Versatile collections of microarray data in public repositories, such as the Gene Expression Omnibus (GEO; Barrett et al. (2009)) and ArrayExpress (Parkinson et al., 2009) are available for human and model organisms, and they

contain valuable information of cell function (Consortium, 2005; DeRisi et al., 1997; Russ and Futschik, 2010; Zhang et al., 2004).

Several techniques are available for quantitative and highly parallel measurements of mRNA or *gene expression*, allowing the measurement of the expression levels of tens of thousands of mRNA transcripts simultaneously (Bradford et al., 2010). Microarray techniques are routinely used to measure the expression levels of tens of thousands of mRNA transcripts in a given sample, and transcriptional profiling is currently a main high-throughput technique used to investigate gene function at genome- and organism-wide scale (Gershon, 2005; Yauk et al., 2004). Increasing amounts of transcriptional profiling data are being produced by sequencing-based methods (Carninci, 2009). A main difference between the microarray- and sequencing-based techniques is that gene expression arrays have been designed to measure predefined mRNA transcripts, whereas sequencing-based methods do not require prior information of the measured sequences, and enable *de novo* discovery of expressed transcripts (Bradford et al., 2010; 't Hoen et al., 2008). Large-scale microarray repositories provide currently the most mature tools for data processing and retrieval, and form the main source of transcriptome data in this thesis.

Microarray technology is based on the base pairing property of nucleic acid sequences where the DNA or RNA sequences in a sample bind to the complementary nucleotide sequences on the array. This is called *hybridization*. The measurement process begins by the collection of cell samples and isolation of the sample mRNA. The isolated mRNA is converted to cDNA, *labeled* with specific marker molecules, and hybridized on complementary probe sequences on the array. The array surface may contain hundreds of thousands of spots, each containing specific probe sequences designed to uniquely match with particular mRNA sequences. The hybridization level reflects the target mRNA concentration in the sample, and it is estimated by measuring the intensity of light emitted by the label molecules with a laser scanner. *Short oligonucleotide arrays* (Lockhart et al., 1996) are among the most widely used microarray technologies, and they are the main source of mRNA expression data in this thesis. Short oligonucleotide arrays utilize multiple, typically 10-20, probes for each transcript target that bind to different regions of the same transcript sequence. Use of several 25-nucleotide probes for each target leads to more robust estimates of transcript activity. Each probe is expected to uniquely hybridize with its intended target, and the detected hybridization level is used as a measure of the activity of the transcript. A short oligonucleotide array measures absolute expression levels of the mRNA sequences; relative differences between conditions can be investigated afterwards by comparing these measurements. A standard whole-genome array measures typically ∼20,000-50,000 unique transcript sequences. A single microarray experiment can therefore produce hundreds of thousands of raw observations.

Comparison and integration of individual microarray experiments is often challenging due to remarkable experimental variation between the experiments. Common standards have been developed to advance the comparison and integration (Brazma et al., 2001, 2006). Carefully controlled integrative datasets, so-called *gene expression atlases*, contain thousands of genome-wide measurements of transcriptional activity across diverse conditions in a directly comparable format. Examples of such data collections include GeneSapiens (Kilpinen et al., 2008), the human gene expression atlas of the European Bioinformatics Institute (Lukk et al.,

2010), as well as the NCI-60 cell line panel (Scherf et al., 2000). Integrative analysis of large and versatile transcriptome collections can provide a holistic view of transcriptional activity of the various cell-biological processes, and opens up possibilities to discover previously uncharacterized cellular mechanisms that contribute to human health and disease.

Other types of microarray data

Microarray techniques can also be used to study other functional aspects of the genome, including epigenetics and micro-RNA regulation, chromosomal aberrations and polymorphisms, alternative splicing, as well as transcription factor binding (Butte, 2002; Hoheisel, 2006). For instance, chromosomal aberrations can be measured with the *array comparative genome hybridization method (aCGH;* Pinkel and Albertson 2005), which is based on hybridization of DNA sequences on the array surface. Copy number changes are a particular type of chromosomal aberrations, which are a major mechanism for cancer development and progression. Copy number alterations can cause changes in gene- and micro-RNA expression, and ultimately cell-biological processes (Beroukhim et al., 2010). A public repository of copy number measurement data is provided for instance by the CanGEM database (Scheinin et al., 2008). In Publication 4, microarray measurements of DNA copy number changes are integrated with transcriptional profiling data to discover potential cancer genes for further biomedical analysis.

Pathway and interaction databases

Curated information concerning cell-biological processes is valuable in both experimental design and validation of computational studies (Blake, 2004). Representation of dynamic biochemical reactions in their full richness is a challenging task beyond a mere listing of biochemical events; a variety of proteins and other compounds interact in a hierarchical manner through various molecular mechanisms (Hartwell et al., 1999; Przytycka et al., 2010). Standardized database formats such as the BioPAX (BioPAX workgroup, 2005) and SBML (Strömbäck and Lambrix, 2005) advance the accumulation of highly structured biological knowledge and automated analysis of such data. A huge body of information concerning cell-biological processes is available in public repositories. The most widely used annotation resources include the Gene Ontology (GO) database (Ashburner et al., 2000) and the KEGG pathway database (Kanehisa et al., 2010). The GO database provides functional annotations for genes and can be used for instance to detect enrichment of certain functional categories among the key findings from computational analysis, as in Publication 6, where enrichment analysis is used for both validation and interpretation purposes. Pathways are more structured representations concerning cellular processes and interactions between molecular entities. Such prior information can be used to guide computational modeling, as in Publication 3, where pathway information derived from the KEGG pathway database is used to guide organism-wide discovery and analysis of transcriptional response patterns.

Evolving biological knowledge

The collective knowledge about genome organization and function is constantly updated and refined by improved measurement techniques and accumulation of

data (Sebat, 2007). This can alter the analysis and interpretation of results from large-scale genomic screens. For instance, evolving gene and transcript definitions are known to significantly affect microarray interpretation. Probe design on microarray technology relies on sequence annotations that may have changed significantly after the original array design. Reinterpretation of microarray data based on updated probe annotations has been shown to improve the accuracy and comparability of microarray results (Dai et al., 2005; Hwang et al., 2004; Mecham et al., 2004b). Bioinformatics studies routinely take into account updates in genome version, *genome build*, in new analyses. The constantly refined biological data highlights the need to account for this uncertainty in computational analyses. In Publications 1 and 2, explicit computational strategies that are robust against evolving transcript definitions are developed for microarray data analysis.

2.3.2 Challenges in high-throughput data analysis

High-throughput genetic screens are inherently noisy. Controlling all potential sources of variation in the measurement process is increasingly difficult when automated measurement techniques can produce millions of data points in a single experiment, concerning extremely complex living systems that are to a large extent poorly understood.

Noise arises from both technical and biological sources (Butte, 2002), and systematic variation between laboratories, measurement batches and measurement platforms has to be taken into account when combining the results across individual studies (Heber and Sick, 2006; MAQC Consortium, 2006). Moreover, genomic knowledge is constantly evolving, which can potentially change the interpretation of previous experiments (see e.g. Dai et al., 2005). The various sources of noise and uncertainty in microarray studies are discussed in more detail in Chapter 4.

High dimensionality of the data and small sample size form another challenge for the analysis of high-throughput functional genomics data. Tens of thousands of transcripts can be measured simultaneously in a single microarray experiment, which greatly exceeds the number of available samples in most biomedical studies. Small sample sizes leave considerable uncertainty in the analyses; few observations contain very limited information concerning the complex and high-dimensional phenomena and potential interactions between different parts of the system. Overfitting of the models and the problem of multiple testing forms considerable challenges in such situations. While automated analysis methods can generate thousands of hypotheses concerning the system, prioritizing the findings and characterizing uncertainty in the predictions become central issues in the analysis. The *curse of dimensionality*, coupled with the high levels of noise in functional genomics studies, is therefore posing particular challenges for computational modeling (Saeys et al., 2007).

The challenges in controlling the various sources of uncertainty have led to remarkable problems in reproducing microarray results (Ioannidis et al., 2009), but maturing technology and the development of common standards and analytical procedures are constantly improving the reliability of high-throughput screens (Allison et al., 2006; Reimers, 2010; MAQC Consortium, 2006). The models developed in this thesis combine statistical evidence across related experiments to improve the reliability of the analysis and to increase modeling power. Generative probabilistic models provide a rigorous framework for handling noise and uncertainty in the data and models.

2.4 Genomics and health

Genomic variation between individuals has remarkable and to a large extent unknown contribution to health and disease susceptibility. Large-scale characterization of the variability between individuals and populations is expected to elucidate genomic mechanisms associated with disease, as well as to lead to the discovery of novel medical treatments. High-throughput genomics can provide new tools to understand disease mechanisms (Braga-Neto and Marques, 2006; Lage et al., 2008), to 'hack the genome' (Evanko, 2006) to treat diseases (Volinia et al., 2010), and to guide *personalized therapies* that take into account the individual variability in sensitivity and responses to treatments (Church, 2005; Downward, 2006; Foekens et al., 2008; Ocana and Pandiella, 2010; van 't Veer and Bernards, 2008). Disease signatures are potentially robust across tissues and experiments (Dudley et al., 2009; Hu et al., 2006). Genomic screens have revealed new disease subtypes (Bhattacharjee et al., 2001), and led to the discovery of various *diagnostic* (Lee et al., 2008; Su et al., 2009; Tibshirani et al., 2002) and *prognostic* (Beer et al., 2002) biomarkers. Diseases cause coordinated changes in gene activity through biomolecular networks (Cabusora et al., 2005). Integration of chemical, genomic and pharmacological functional genomics data can also help to predict new drug targets and responses (Lamb et al., 2006; Yamanishi et al., 2010). Genomic mutations can also affect genome function and cause diseases (Taylor et al., 2008). Cancer is an example of a prevalent genomic disease. Boveri (1914) discovered that cancer cells have chromosomal imbalances, and since then the understanding of genomic changes associated with cancer has continuously improved (Stratton et al., 2009; Wunderlich, 2007). For instance, many human micro-RNA genes are located at cancer-associated genomic regions and are functionally altered in cancers (see Calin and Croce, 2006). Genomic changes also affect transcriptional activity of the genes (Myllykangas et al., 2008). Publication 4 introduces a novel computational approach for screening cancer-associated DNA mutations with functional implications by genome-wide integration of chromosomal aberrations and transcriptional activity.

This chapter has provided an overview to central modeling challenges and research topics in functional genomics. In the following chapters, particular methodological approaches are introduced to solve research tasks in large-scale analysis of the human transcriptome. In particular, methods are introduced to increase the reliability of high-throughput measurements, to model large-scale collections of transcriptome data and to integrate transcriptional profiling data to other layers of genomic information. The next chapter provides general methodological background for these studies.

Chapter 3

Statistical learning and exploratory data analysis

Essentially, all models are wrong, but some are useful.

G.E.P. Box and N.R. Draper (1987)

Models are condensed, simplified representations of observed phenomena. Models can be used to describe observations and to predict future events. Two key aspects in modeling are the construction and learning of formal representations of the observed data. Complex real-world observations contain large amounts of uncontrolled variation, which is often called *noise*; all aspects of the data cannot be described within a single model. Therefore, a *modeling compromise* is needed to decide what aspects of data to describe and what to ignore. The second step in modeling is to fill in, to *learn*, details of the formal representation based on the actual empirical observations. Various learning algorithms are typically available that differ in efficiency and accuracy. For instance, improvements in computation time can often be achieved by potential decrease in accuracy. An *inference compromise* is needed to decide how to balance between these and other potentially conflicting objectives of the learning algorithm; the relative importance of each factor depends on the particular application and available resources, and affects the choice of the learning procedure. The modeling and inference compromises are at the heart of data analysis. Ultimately, the value of a model is determined by its ability to advance the solving of practical problems.

This chapter gives an overview of the key concepts in statistical modeling central to the topics of this thesis. The objectives of exploratory data analysis and statistical learning are considered in Section 3.1. The methodological framework is introduced in Section 3.2, which contains an overview of central concepts in probabilistic modeling and the Bayesian analysis paradigm. Key issues in implementing and validating the models are discussed in Section 3.3.

3.1 Modeling tasks

Understanding requires generalization beyond particular observations. While empirical observations contain information of the underlying process that generated

the data, a major challenge in computational modeling is that empirical data is always finite and contains only limited information of the system. Traditional statistical models are based on careful hypothesis formulation and systematic collection of data to support or reject a given hypothesis. However, successful hypothesis formulation may require substantial prior knowledge. When minimal knowledge of the system is available, there is a need for *exploratory methods* that can recognize complex patterns and extract features from empirical data in an automated way (Baldi and Brunak, 1999). This is a central challenge in computational biology, where the investigated systems are extremely complex and contain large amounts of poorly characterized and uncontrolled sources of variation. Moreover, the data of genomic systems is often very limited and incomplete. General-purpose algorithms that can learn relevant features from the data with minimal assumptions are therefore needed, and they provide valuable tools in functional genomics studies. Classical examples of such exploratory methods include clustering, classification and visualization techniques. The extracted features can provide hypotheses for more detailed experimental testing and reveal new, unexpected findings. In this work, general-purpose exploratory tools are developed for central modeling tasks in functional genomics.

3.1.1 Central concepts in data analysis

Let us start by defining some of the basic concepts and terminology. *Data set* in this thesis refers to a finite collection of observations, or *samples*. In experimental studies, as in biology, a sample typically refers to the particular object of study, for instance a patient or a tissue sample. In computational studies, sample refers to a numerical observation, or a subset of observations, represented by a numerical *feature vector*. Each element of the feature vector describes a particular *feature* of the observation. Given D features and N samples, the data set is presented as a matrix $\mathbf{X} \in \mathbb{R}^{D \times N}$, where each column vector $\mathbf{x} \in \mathbb{R}^D$ represents a sample and each row corresponds to a particular feature. The features can represent for instance different experimental conditions, time points, or particular summaries about the observations. This is the general structure of the observations investigated in this work.

The observations are modeled in terms of probability densities; the samples are modeled as independent instances of a random variable. A central modeling task is to characterize the underlying probability density of the observations, $p(\mathbf{x})$. This defines a topology in the sample space and provides the basis for generalization beyond empirical observations. As explained in more detail in Section 3.2, the models are formulated in terms of observations $\mathbf{X}$, model parameters $\boldsymbol{\theta}$, and *latent variables* $\mathbf{Z}$ that are not directly observed, but characterize the underlying process that generated the data.

Ultimately, all models describe relationships between objects. *Similarity* is therefore a key concept in data analysis; the basis for characterizing the relations, for summarizing the observations, and for predicting future events. Measures of similarity can be defined for different classes of objects such as feature vectors, data sets, or random variables. Similarity in general is a vague concept. *Euclidean distance*, induced by the Euclidean metrics, is a common (dis-)similarity measure for multivariate observations. *Correlation* is a standard choice for univariate random variables. *Mutual information* is an information-theoretic measure of statistical dependency between two random variables, characterizing the decrease in

the uncertainty concerning the realization of one variable, given the other one. The uncertainty of a random variable $\mathcal{X}$ is measured in terms of *entropy*[1] (Shannon, 1948). The mutual information between two random variables is then given by $I(\mathcal{X}, \mathcal{Y}) = H(\mathcal{X}) - H(\mathcal{X}|\mathcal{Y})$ (see e.g. Gelman et al., 2003). The Kullback-Leibler divergence, or *KL–divergence*, is a closely related non-symmetric dissimilarity measure for probability distributions p, q, defined as $d_{KL}(p, q) = \int_{\mathbf{x}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x}$ (see e.g. Bishop, 2006). Mutual information between two random variables can be alternatively formulated as the KL–divergence between their joint density $p(\mathbf{x}, \mathbf{y})$ and the product of their independent marginal densities, $q(\mathbf{x}, \mathbf{y}) = p_x(\mathbf{x})p_y(\mathbf{y})$, which gives the connection $I(\mathcal{X}, \mathcal{Y}) = d_{KL}(p(\mathbf{x}, \mathbf{y}), p_x(\mathbf{x})p_y(\mathbf{y}))$. Mutual information and KL-divergence are central information-theoretic measures of dependency employed in the models of this thesis.

It is important to notice that measures of similarity are inherently coupled to the statistical representation of data and to the goals of the analysis; different representations can reveal different relationships between observations. For instance, the Euclidean distance is sensitive to scaling of the features; representation in natural or logarithmic scale, or with different units can potentially lead to very different analysis results. Not all measures are equally sensitive; mutual information can naturally detect non-linear relationships, and it is invariant to the scale of the variables. On the other hand, estimating mutual information is computationally demanding.

Feature selection refers to computational techniques for selecting, scaling and transforming the data into a suitable form for further analysis. Feature selection has a central role in data analysis, and it is implicitly present in all analysis tasks in selecting the investigated features for the analysis.

There are no universally optimal stand-alone feature selection techniques, since the problem is inherently entangled with the analysis task and multiple equally optimal feature sets may be available for instance in classification or prediction tasks Guyon and Elisseeff (2003); Saeys et al. (2007). Successful feature selection can reduce the dimensionality of the data with minimal loss of relevant information, and focus the analysis on particular features. This can reduce model complexity, which is expected to yield more efficient, generalizable and interpretable models. Feature selection is particularly important in genome-wide profiling studies, where the dimensionality of the data is large compared to the number of available samples, and only a small number of features are relevant for the studied phenomenon. This is also known as the *large p, small n* problem (West, 2003). Advanced feature selection techniques can take into account dependencies between the features, consider weighted combinations of them, and can be designed to interact with the more general modeling task, as for instance in the nearest shrunken centroids classifier of Tibshirani et al. (2002). The constrained subspace clustering model of Publication 3 can be viewed as a feature selection procedure, where high-dimensional genomic observations are decomposed into distinct feature subsets, each of which reveals different relationships of the samples. In Publication 4, identification of maximally informative features between two data sets forms a central part of a regularized dependency modeling framework. In Publications 3-4 the procedure and representations are motivated by biological reasoning and analysis goals.

[1] Entropy is defined as $H(\mathcal{X}) = -\int_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x}$ for a continuous variable.

3.1.2 Exploratory data analysis

Exploratory data analysis refers to the use of computational techniques to summarize and visualize data in order to facilitate the generation of new hypotheses for further study when the search space would be otherwise exhaustively large (Tukey, 1977). The analysis strategy takes the observations as the starting point for discovering interesting regularities and novel research hypotheses for poorly characterized large-scale systems without prior knowledge. The analysis can then proceed from general observations of the data toward *confirmatory data analysis*, more detailed investigations and hypotheses that can be tested in independent data sets with standard statistical procedures. Exploratory data analysis differs from traditional hypothesis testing where the hypothesis is given. Light-weight exploratory tools are particularly useful with large data sets when prior knowledge on the system is minimal. Standard exploratory approaches in computational biology include for instance clustering, classification and visualization techniques (Evanko, 2010; Polanski and Kimmel, 2007).

Cluster analysis refers to a versatile family of methods that partition data into internally homogeneous groups of similar data points, and often at the same time minimize the similarity between distinct clusters. Clustering techniques enable *class discovery* from the data. This differs from classification where the target is to assign new observations into known classes. The partitions provided by clustering can be nested, partially overlapping or mutually exclusive, and many clustering methods generalize the partitioning to cover previously unseen data points (Jain and Dubes, 1988). Clustering can provide compressed representations of the data based on a shared parametric representation of the observations within each cluster, as for instance in K-means or Gaussian mixture modeling (see e.g. Bishop, 2006). Certain clustering approaches, such as the hierarchical clustering (see e.g. Hastie et al., 2009), apply recursive schemes that partition the data into internally homogeneous groups without providing a parametric representation of the clusters. Cluster structure can be also discovered by linear algebraic operations on the distance matrices, as for instance in spectral clustering. The different approaches often have close theoretical connections. Clustering in general is an ill-defined concept that refers to a set of related but mutually incompatible objectives (Ben-David and Ackerman, 2008; Kleinberg, 2002). Cluster analysis has been tremendously popular in computational biology, and a comprehensive review of the different applications are beyond the scope of this thesis. It has been observed, for instance, that genes with related functions have often similar expression profiles and are clustered together, suggesting that clustering can be used to formulate hypotheses concerning the function of previously uncharacterized genes (DeRisi et al., 1997; Eisen et al., 1998), or to discover novel cancer subtypes with biomedical implications (Sørlie et al., 2001).

Visualization techniques are another widely used exploratory approach in computational biology. Visualizations can provide compact and intuitive summaries of complex, high-dimensional observations on a lower-dimensional display, for instance by linear projection methods such as principal component analysis, or by explicitly optimizing a lower-dimensional representation as in the self-organizing map (Kohonen, 1982). Visualization can provide the first step in investigating large data sets (Evanko, 2010).

3.1.3 Statistical learning

Statistical learning refers to computational models that can learn to recognize structure and patterns from empirical data in an automated way. Unsupervised and supervised models form two main categories of learning algorithms.

Unsupervised learning approaches seek compact descriptions of the data without prior knowledge. In probabilistic modeling, unsupervised learning can be formulated as the task of finding a probability distribution that describes the observed data and generalizes to new observations. This is also called *density estimation*. The parameter values of the model can be used to provide compact representations of the data. Examples of unsupervised analysis tasks include methods for clustering, visualization and dimensionality reduction. In cluster analysis, groups of similar observations are sought from the data. Dimensionality reduction techniques provide compact lower-dimensional representations of the original data, which is often useful for subsequent modeling steps. Not all observations of the data are equally valuable, and assessing the relevance of the observed regularities is problematic in fully unsupervised analysis.

In *supervised learning* the task is to learn a function that maps the inputs $\mathbf{x}$ to the desired outputs $\mathbf{y}$ based on a set of training examples in a generalizable fashion, as in regression for continuous outputs, and classification for discrete output variables. The supervised learning tasks are inherently asymmetric; the inference proceeds from inputs to outputs, and prior information of the modeling task is used to supervise the analysis; the training examples also include a desired output of the model.

The models developed in this thesis can be viewed as unsupervised exploratory techniques. However, the distinction between supervised and unsupervised models is not strict, and the models in this thesis borrow ideas from both categories. The models in Publications 2-3 are unsupervised algorithms that utilize prior information derived from background databases to guide the modeling by constraining the solutions. However, since no desired outputs are available for these models, the modeling tasks differ from supervised analysis. The dependency modeling algorithms of Publications 4-6 have close theoretical connections to the supervised learning task. In contrast to supervised learning, the learning task in these algorithms is symmetric; modeling of the co-occurring data sets is unsupervised, but coupled. Each data set affects the modeling of the other data set in a symmetric fashion, and, in analogy to supervised learning, prediction can then proceed to either direction. Compared to supervised analysis tasks, the emphasis in the dependency detection algorithms introduced in this thesis is in the discovery and characterization of symmetric dependencies, rather than in the construction of asymmetric predictive models.

3.2 Probabilistic modeling paradigm

The main contributions of this thesis follow the generative probabilistic modeling paradigm. Generative probabilistic models describe the observed data in terms of probability distributions. This allows the calculation of expectations, variances and other standard summaries of the model parameters, and at the same time allows to describe the independence assumptions and relations between variables, and uncertainty in the modeling process in an explicit manner. Measurements

are regarded as noisy observations of the general, underlying processes; generative models are used to characterize the processes that generated the observations.

The first task in modeling is the selection of a *model family* - a set of potential formal representations of the data. As discussed in Section 3.2.2, the representations can also to some extent be learned from the data. The second task is to define the *objective function*, or cost function, which is used to measure the descriptive power of the models. The third task is to identify the optimal model within the model family that best describes the observed data with respect to the objective function. This is called *learning* or *model fitting*. The details of the modeling process are largely determined by the exact modeling task and particular nature of the observations. The objectives of the modeling task are encoded in the selected model family, the objective function and to some extent to the model fitting procedure. The model family determines the space of possible descriptions for the data and has therefore a major influence on the final solution. The objective function can be used to prefer simple models or other aspects in the modeling process. The model fitting procedure affects the efficiency and accuracy of the learning process. For further information of these and related concepts, see Bishop (2006). A general overview of the probabilistic modeling framework is given in the remainder of this section.

3.2.1 Generative modeling

Generative probabilistic models view the observations as random samples from an underlying probability distribution. The model defines a probability distribution $p(\mathbf{x})$ over the feature space. The model can be parameterized by model parameters $\boldsymbol{\theta}$ that specify a particular model within the model family. For convenience, we assume that the model family is given, and leave it out from the notation. In this thesis, the appropriate model families are selected based on biological hypotheses and analysis goals. Generative models allow efficient representation of dependencies between variables, independence assumptions and uncertainty in the inference (Koller and Friedman, 2009). Let us next consider central analysis tasks in generative modeling.

Finite mixture models

Classical probability distributions provide well-justified and convenient tools for probabilistic modeling, but in many practical situations the observed regularities in the data cannot be described with a single standard distribution. However, a sufficiently rich mixture of standard distributions can provide arbitrarily accurate approximations of the observed data. In *mixture models*, a set of distinct, latent processes, or *components*, is used to describe the observations. The task is to identify and characterize the components and their associations to the individual observations. The standard formulation assumes independent and identically distributed observations where each observation has been generated by exactly one component. In a standard mixture model the overall probability density of the data is modeled as a weighted sum of component distributions:

$$p(\mathbf{x}) = \sum_{r=1}^{R} \pi_r p_r(\mathbf{x}|\boldsymbol{\theta}_r), \tag{3.1}$$

where the components are indexed by r, and $\int p(\mathbf{x})d\mathbf{x} = 1$. Each mixture component can have a different distributional form. The mixing proportion, or weight, and model parameters of each component are denoted by π_r and $\boldsymbol{\theta}_r$, respectively, with $\sum_r \pi_r = 1$. Many applications utilize convenient standard distributions, such as Gaussians, or other distributions from the exponential family. Then the mixture model can be learned for instance with the EM algorithm described in Section 3.3.1.

In practice, the mixing proportions of the components are often unknown. The mixing proportions can be estimated from the data by considering them as standard model parameters to be fitted with a ML estimate. However, the procedure is potentially prone to overfitting and local optima, i.e., it may learn to describe the training data well, but fails to generalize to new observations. An alternative, probabilistic way to determine the weights is to treat the mixing proportions as latent variables with a prior distribution $p(\boldsymbol{\pi})$. A standard choice is a symmetric Dirichlet prior[2] $\boldsymbol{\pi} \sim Dir(\frac{\boldsymbol{\alpha}}{R})$. This gives an equal prior weight for each component and guarantees the standard exchangeability assumption of the mixture component labels. A label determines cluster identity. Intuitively, exchangeability corresponds to the assumption that the analysis is invariant to the ordering of the data samples and mixture components. Compared to standard mixture models, probabilistic mixture models have increased computational complexity.

Further prior knowledge can be incorporated in the model by defining prior distributions for the other parameters of the mixture model. This can also be used to regularize the learning process to avoid overfitting. A typical prior distribution for the components of a Gaussian mixture model, parameterized by $\boldsymbol{\theta}_r = \{\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r\}$, is the normal-inverse-Gamma prior (see e.g. Gelman et al., 2003).

Interpreting the mixture components as clusters provides an alternative, probabilistic formulation of the clustering task. This has made probabilistic mixture models a popular choice in the analysis of functional genomics data sets that typically have high dimensionality but small sample size. Probabilistic analysis takes the uncertainties into account in a rigorous manner, which is particularly useful when the sample size is small. The number of mixture components is often unknown in practical modeling tasks, however, and has to be inferred based on the data. A straightforward solution can be obtained by employing a sufficiently large number of components in learning the mixture model, and selecting the components having non-zero weights as a post-processing step. An alternative, model-based treatment for learning the number of mixture components from the data is provided by infinite mixture models considered in Section 3.2.2.

Latent variables and marginalization

The observed variables are often affected by *latent variables* that describe relevant structure in the model, but are not directly observed. The latent variable values can be, to some extent, inferred based on the observed variables. Combination of latent and observed variables allows the description of complex probability spaces in terms of simple component distributions and their relations. Use of simple component distributions can provide an intuitive and computationally tractable characterization of complex generative processes underlying the observations.

[2] *Dirichlet distribution* is the probability density $Dir(\boldsymbol{\pi}|\mathbf{n}) \sim \prod_r \pi_r^{n_r - 1}$ where the multivariate random variable $\boldsymbol{\pi}$ and the positive parameter vector $\mathbf{n}$ have their elements indexed by r, $0 < \pi_r < 1$, and $\sum_r \pi_r = 1$.

A generative latent variable model specifies the distributional form and relationships of the latent and observed variables. As a simple example, consider the probabilistic interpretation of probabilistic component analysis (PCA), where the observations $\mathbf{x}$ are modeled with a linear model $\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\varepsilon}$ where a normally distributed latent variable $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I})$ is transformed with the parameter matrix $\mathbf{W}$ and isotropic Gaussian noise ($\boldsymbol{\varepsilon}$) is assumed on the observations. More complex models can be constructed by analogous reasoning. A *complete-data likelihood* $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ defines a joint density for the observed and latent variables. Only a subset of variables in the model is typically of interest for the actual analysis task. For instance, the latent variables may be central for describing the generative process of the data, but their actual values may be irrelevant. Such variables are called *nuisance variables*. Their integration, or *marginalization*, provides probabilistic averaging over the potential realizations. Marginalization over the latent variables in the complete-data likelihood gives the likelihood

$$p(\mathbf{X}|\boldsymbol{\theta}) = \int_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) d\mathbf{Z}. \tag{3.2}$$

Marginalization over the latent variables collapses the modeling task to finding optimal values for model parameters $\boldsymbol{\theta}$, in a way that takes into account the uncertainty in latent variable values. This can reduce the number of variables in the learning phase, yield more straightforward and robust inferences, as well as speed up computation. However, marginalization may lead to analytically intractable integrals. As certain latent variables may be directly relevant, marginalization depends on the overall goals of the analysis and may cover only a subset of the latent variables. In this thesis latent variables are utilized for instance in Publication 3, which treats the sample-cluster assignments as discrete latent variables, as well as in Publication 4, where a regularized latent variable model is introduced to model dependencies between co-occurring observations.

3.2.2 Nonparametric models

Finite mixture models and latent variable models require the specification of model structure prior to the analysis. This can be problematic since for instance the number and distributional shape of the generative processes is unknown in many practical tasks. However, the model structure can also to some extent be learned from the data. Non-parametric models provide principled approaches to learn the model structure from the data. In contrast to parametric models, the number and use of the parameters in nonparametric models is flexible (see e.g. Hjort et al., 2010; Müller and Quintana, 2004). The infinite mixture of Gaussians, used as a part of the modeling process in Publication 3, is an example of a non-parametric model where both the number of components, as well as mixture proportions of the component distributions are inferred from the data. Learning of Bayesian network structure is another example of nonparametric inference, where relations between the model variables are learned from the data (see e.g. Friedman, 2003). While more complex models can describe the training data more accurately, an increasing model complexity needs to be penalized to avoid overfitting and to ensure generalizability of the model.

Nonparametric models provide flexible and theoretically principled approaches for data-driven exploratory analysis. However, the flexibility often comes with an increased computational cost, and the models are potentially more prone to

overfitting than less flexible parametric models. Moreover, complex models can be difficult to interpret.

Many nonparametric probabilistic models are defined by using the theory of stochastic processes to impose priors over potential model structures. Stochastic processes can be used to define priors over function spaces. For instance, the *Dirichlet process (DP)* defines a probability density over the function space of Dirichlet distributions[3]. The *Chinese Restaurant Process (CRP)* provides an intuitive description of the Dirichlet process in the cluster analysis context. The CRP defines a prior distribution over the number of clusters and their size distribution. The CRP is a random process in which n customers arrive in a restaurant, which has an infinite number of tables. The process goes as follows: The first customer chooses the first table. Each subsequent customer m will select a table based on the state F_{m-1} of the restaurant tables after $m-1$ customers have arrived. The new customer m will select a previously occupied table i with a probability which is proportional to the number of customers seated at table i, i.e. $p(i|F_{m-i}) \propto n_i$. Alternatively, the new customer will select an empty table with a probability which is proportional to a constant α. The model prefers tables with a larger number of customers, and is analogous to clustering, where the customers and tables correspond to samples and clusters, respectively. This provides an intuitive prior distribution for clustering tasks. The prior prefers compact models with relatively few clusters, but the number of clusters is potentially infinite, and ultimately determined based on the data.

Infinite mixture models

Infinite mixture models are a general class of nonparametric methods where the number of mixture components are determined in a data-driven manner; the number of components is potentially infinite (see e.g. Müller and Quintana, 2004; Rasmussen, 2000). An infinite mixture is obtained by letting $R \to \infty$ in the finite mixture model of Equation 3.1 and replacing the Dirichlet distribution prior of the mixing proportions $\boldsymbol{\pi}$ by a Dirichlet process. The formal probability distribution of the Dirichlet process can be intuitively derived with the so-called *stick-breaking presentation*. Consider a unit length stick and a stick-breaking process, where the breakpoint β is stochastically determined, following the beta distribution $\beta \sim Beta(1, \alpha)$, where α tunes the expected breaking point. The process can be viewed as consecutively breaking off portions of a unit length stick to obtain an infinite sequence of stick lengths $\pi_1 = \beta_1$; $\pi_i = \beta_i \prod_{l=1}^{i-1}(1-\beta_l)$, with $\sum_{i=1}^{\infty} \pi_i = 1$ (Ishwaran and James, 2001). This defines the probability distribution $\text{Stick}(\alpha)$ over potential partitionings of the unit stick. A truncated stick-breaking representation considers only the first T elements. Setting the prior $\boldsymbol{\pi} \sim \text{Stick}(\alpha)$, defined by the stick-breaking representation in Equation 3.1 assigns a prior on the number of mixture components and their mixing proportions that are ultimately learned from the observed data. The prior helps to find a compromise between increasing model complexity and likelihood of the observations.

Traditional approaches used to determine the number mixture components are based on objective functions that penalize increasing model complexity, for instance in certain variants of the K-means or in spectral clustering (see e.g. Hastie

[3]If G is a distribution drawn from a Dirichlet process with the probability measure P over the sample space, $G \sim \text{DP}(P)$, then each finite partition $\{A_k\}_k$ of the sample space is distributed as $(G(A_1), ..., G(A_k)) \sim Dir(P(A_1), ..., P(A_k))$.

et al., 2009). Other model selection criteria include cross-validation and comparison of the models in terms of their likelihood or various information-theoretic criteria that seek a compromise between model complexity and fit (see e.g. Gelman et al., 2003). However, the sample size may be insufficient for such approaches, and the models may lack a rigorous framework to account for uncertainties in the observations and model parameters. Modeling uncertainty in the parameters while learning the model structure can lead to more robust inference in nonparametric probabilistic models but also adds inherent computational complexity in the learning process.

3.2.3 Bayesian analysis

The term 'Bayesian' refers to interpretation of model parameters as variables. The uncertainty over the parameter values, arising from limited empirical evidence, is described in terms of probability distributions. This is in contrast to the traditional view where parameters have fixed values with no distribution and the uncertainty is ignored. The Bayesian approach leads to a learning task where the objective is to estimate the *posterior distribution* $p(\boldsymbol{\theta}|\mathbf{X})$ of the model parameters $\boldsymbol{\theta}$, given the observations $\mathbf{X}$. The posterior is given by the *Bayes' rule* (Bayes, 1763):

$$p(\boldsymbol{\theta}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{X})}. \tag{3.3}$$

The two key elements of the posterior are *the likelihood* and *the prior*. The likelihood $p(\mathbf{X}|\boldsymbol{\theta})$ describes the probability of the observations, given the parameter values $\boldsymbol{\theta}$. The parameters can also characterize alternative model structures. The prior $p(\boldsymbol{\theta})$ encodes prior beliefs about the model and rewards solutions that match with the prior assumptions or yield simpler models. Such regularizing properties can be particularly useful when training data is scarce and there is considerable uncertainty in the parameter estimates. With strong, informative priors, new observations have little effect on the posterior. In the limit of large sample size the posterior converges to the ordinary likelihood $p(\mathbf{X}|\boldsymbol{\theta})$. The Bayesian inference provides a robust framework for taking the uncertainties into account when the data is scarce, as it often is in practical modeling tasks. Moreover, the Bayes' rule provides a formal framework for sequential update of beliefs based on accumulating evidence. The prior predictive density $p(\mathbf{X}) = \int p(\mathbf{X}, \boldsymbol{\theta}) d\boldsymbol{\theta}$ is a normalizing constant, which is independent of the parameters $\boldsymbol{\theta}$ and can often be ignored during model fitting.

The involved distributions can have complex non-standard forms and limited empirical data can only provide partial evidence regarding the different aspects of the data-generating process. Often only a subset of the parameters and other variables and their interdependencies can be directly observed. The Bayesian approach provides a framework for making inferences on the unobserved quantities through hierarchical models, where the probability distribution of each variable is characterized by higher-level parameters, so-called *hyperparameters*. A similar reasoning can be used to model the uncertainty in the hyperparameters, until the uncertainties become modeled at an appropriate detail. Prior information can help to compensate the lack of data on certain aspects of a model, and explicit models for the noise can characterize uncertainty in the empirical observations. Distributions can also share parameters, which provides a basis for pooling evidence from multiple sources, as for instance in Publication 4. In many applications only a

subset of the parameters in the model are of interest and the modeling process can be considerably simplified by marginalizing over the less interesting nuisance variables to obtain an expectation over their potential values.

The Bayesian paradigm provides a principled framework for modeling the uncertainty at all levels of statistical inference, including the parameters, the observed and latent variables and the model structure; all information of the model is incorporated in the posterior distribution, which summarizes empirical evidence and prior knowledge, and provides a complete description of the expected outcomes of the data-generating process. When the data does not contain sufficient information to decide between the alternative model structures and parameter values, the Bayesian framework provides tools to take expectations over all potential models, weighted by their relative evidence.

A central challenge in the Bayesian analysis is that the models often include analytically intractable posterior distributions, and learning of the models can be computationally demanding. Widely-used approaches for estimating posterior distributions include *Markov Chain Monte Carlo (MCMC)* methods and variational learning. Stochastic MCMC methods provide a widely-used family of algorithms to estimate intractable distributions by drawing random samples from these distributions (see e.g. Gelman et al., 2003); a sufficiently large pool of random samples will converge to the underlying distribution, and sample statistics can then be used to characterize the distribution. However, sampling-based methods are computationally intensive and slow. In variational learning, considered in Section 3.3.1, the intractable distributions are approximated by more convenient tractable distributions, which yields faster learning procedure, but potentially less accurate results. While analysis of the full posterior distribution will provide a complete description of the uncertainties regarding the parameters, simplified summary statistics, such as the mean, variance and quantiles of the posterior can provide a sufficient characterization of the posterior in many practical applications. They can be obtained for instance by summarizing the output of sampling-based or variational methods. Moreover, when the uncertainty in the results can be ignored, point estimates can provide simple, interpretable summaries that are often useful in further biomedical analysis, as for instance in Publication 2. Point estimates are single optimal values with no distribution. However, point estimates are not necessarily sufficient for instance in biomedical diagnostics and other prediction tasks, where different outcomes are associated with different costs and it may be crucial to assess the probabilities of the alternative outcomes. For further discussion on learning the Bayesian models, see Section 3.3.1.

In this thesis the Bayesian approach provides a formal framework to perform robust inference based on incomplete functional genomics data sets and to incorporate prior information of the models in the analysis. The Bayesian paradigm can alternatively be interpreted as a philosophical position, where probability is viewed as a subjective concept (Cox, 1946), or considered a direct consequence of making rational decisions under uncertainty (Bernardo and Smith, 2000). For further concepts in model selection, comparison and averaging in the Bayesian analysis, see Gelman et al. (2003). For applications in computational biology, see Wilkinson (2007).

3.3 Learning and inference

The final stage in probabilistic modeling is to learn the optimal statistical presentation for the data, given the model family and the objective function. This section highlights central challenges and methodological issues in statistical learning.

3.3.1 Model fitting

Learning in probabilistic models often focuses on optimizing the model parameters $\boldsymbol{\theta}$. In addition, posterior distribution of the latent variables, $p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta})$, can be calculated. Estimating the latent variable values is called statistical *inference*. In the Bayesian analysis, the model parameters can also be treated as latent variables with a prior probability density, in which case the distinction between model parameters and latent variables will disappear. A comprehensive characterization of the variables and their uncertainty would be achieved by estimating the full posterior distribution. However, this can be computationally very demanding, in particular when the posterior is not analytically tractable. The posterior is often approximated with stochastic or analytical procedures, such as stochastic MCMC sampling methods or variational approximations, and appropriate summary statistics. In many practical settings, it is sufficient to summarize the full posterior distribution with a point estimate. Point estimates do not characterize the uncertainties in the analysis result, but are often more convenient to interpret than full posterior distributions.

Various optimization algorithms are available to learn statistical models, given the learning procedure. The potential challenges in the optimization include *computational complexity* and the presence of *local optima* on complex probability density topologies, as well as *unidentifiability* of the models. Finding a global optimum of a complex model can be computationally exhaustive, and it can become intractable with increasing sample size. In unidentifiable models, the data does not contain sufficient information to choose between alternative models with equal statistical evidence. Ultimately, the uncertainty in inference arises from limited sample size and the lack of computational resources.

In the remainder of this section, let us consider more closely the particular learning procedures central to this thesis: point estimates and variational approximation, and the standard optimization algorithms used to learn such representations.

Point estimates

Assuming independent and identically distributed observations, the *likelihood* of the data, given model parameters, is $p(\mathbf{X}|\boldsymbol{\theta}) = \prod_i p(\mathbf{x}_i|\boldsymbol{\theta})$. This provides a probabilistic measure of model fit and the objective function to maximize. Maximization of the likelihood $p(\mathbf{X}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ yields a *maximum likelihood (ML)* estimate of the model parameters, and specifies an optimal model that best describes the data. This is a standard point estimate used in probabilistic modeling. Practical implementations typically operate on *log-likelihood*, the logarithm of the likelihood function. As a monotone function, this yields the same optima, but has additional desirable properties: it factorizes the product into a sum and is less prone to numerical overflows during optimization.

The *maximum a posteriori (MAP)* estimate additionally takes prior information of the model parameters into account. While the ML estimate maximizes the likelihood $p(\mathbf{X}|\boldsymbol{\theta})$ of the observations, the MAP estimate maximizes the posterior $p(\boldsymbol{\theta}|\mathbf{X}) \sim p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ of the model parameters. The objective function to maximize is the log-likelihood

$$logp(\boldsymbol{\theta}|\mathbf{X}) \sim logp(\mathbf{X}|\boldsymbol{\theta}) + logp(\boldsymbol{\theta}). \tag{3.4}$$

The prior is explicit in MAP estimation and the model contains the ML estimate as a special case; assuming large sample size, or non-informative, uniform prior $p(\boldsymbol{\theta}) \sim 1$, the likelihood of the data $p(\mathbf{X}|\boldsymbol{\theta})$ will dominate and the MAP estimation becomes equivalent to optimizing $p(\mathbf{X}|\boldsymbol{\theta})$, yielding the traditional ML estimate. The ML and MAP estimates are asymptotically consistent approximations of the posterior distribution, since the posterior will converge a point distribution with a large sample size. The computation and interpretation of point estimates is straightforward compared to the use of posterior distributions in the full Bayesian treatment. The differences between ML and MAP estimates highlight the role of prior information in the modeling when training data is limited.

Variational inference

In certain modeling tasks the uncertainty in the model parameters needs to be taken into account. Then point estimates are not sufficient. The uncertainty is characterized by the posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$. However, the posterior distributions are often intractable and need to be estimated by approximative methods. *Variational approximations* provide a fast and principled optimization scheme (see e.g. Bishop, 2006) that yields only approximative solutions, but can accelerate posterior inference by orders of magnitude compared to stochastic, sampling-based MCMC methods that can in principle provide exact solutions, assuming that infinite computational resources are available. The potential decrease in accuracy in variational approximations is often acceptable, given the gains in efficiency. Variational approximation characterizes the uncertainty in $\boldsymbol{\theta}$ with a tractable distribution $q(\boldsymbol{\theta})$ that approximates the full, potentially intractable posterior $p(\boldsymbol{\theta}|\mathbf{X})$,

Variational inference is formulated as an optimization problem where an intractable posterior distribution $p(\mathbf{Z}, \boldsymbol{\theta}|\mathbf{X})$ is approximated by a more easily tractable distribution $q(\mathbf{Z}, \boldsymbol{\theta})$ by minimizing the KL–divergence between the two distributions. This is also shown to maximize a lower bound of the marginal likelihood $p(\mathbf{X})$, and subsequently the likelihood of the data, yielding an approximation of the overall model. The log-likelihood of the data can be decomposed into a sum of the lower bound $\mathcal{L}(q)$ of the observed data and the KL–divergence $d_{KL}(q, p)$ between the approximative and the exact posterior distributions:

$$logp(\mathbf{X}) = \mathcal{L}(q) + d_{KL}(q, p), \tag{3.5}$$

where

$$\begin{array}{rcl} \mathcal{L}(q) & = & \int_{\mathbf{z}} q(\mathbf{Z}, \boldsymbol{\theta}) log \frac{p(\mathbf{Z}, \boldsymbol{\theta}, \mathbf{X})}{q(\mathbf{Z}, \boldsymbol{\theta})}; \\ d_{KL}(q, p) & = & -\int_{\mathbf{z}} q(\mathbf{Z}, \boldsymbol{\theta}) log \frac{p(\mathbf{Z}, \boldsymbol{\theta}|\mathbf{X})}{q(\mathbf{Z}, \boldsymbol{\theta})}. \end{array} \tag{3.6}$$

The KL-divergence is non-negative, and equals to zero if and only if the approximation and the exact distribution are identical. Therefore $\mathcal{L}(q)$ gives a

lower bound for the log-likelihood $logp(\mathbf{X})$ in Equation 3.5. Minimization of d_{KL} with respect to q will provide an analytically tractable approximation $q(\mathbf{Z}, \boldsymbol{\theta})$ of $p(\mathbf{Z}, \boldsymbol{\theta}|\mathbf{X})$. Minimization of d_{KL} will also maximize the lower bound $\mathcal{L}(q)$ since the log-likelihood $logp(\mathbf{X})$ is independent of q. The approximation typically assumes independent parameters and latent variables, yielding a *factorized* approximation $q(\mathbf{Z}, \boldsymbol{\theta}) = q_{\mathbf{z}}(\mathbf{Z})q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ based on tractable standard distributions. It is also possible to factorize $q_{\mathbf{z}}$ and $q_{\boldsymbol{\theta}}$ into further components. Variational approximations are used for efficient learning of infinite multivariate Gaussian mixture models in Publication 3.

Expectation–Maximization (EM)

The *EM algorithm* is a general procedure for learning probabilistic latent variable models (Dempster et al., 1977), and a special case of variational inference. The algorithm provides an efficient algorithm for finding point estimates for model parameters in latent variable models. The objective of the EM algorithm is to maximize the marginal likelihood

$$p(\mathbf{X}|\boldsymbol{\theta}) = \int_{\mathbf{z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) d\mathbf{Z} \tag{3.7}$$

of the observations $\mathbf{X}$ with respect to the model parameters $\boldsymbol{\theta}$. Marginalization over the probability density of the latent variables provides an inference procedure that is robust to uncertainty in the latent variable values. The algorithm iterates between estimating the posterior of the latent variables, and optimizing the model parameters (see e.g. Bishop, 2006). Given initial values $\boldsymbol{\theta}_0$ of the model parameters, the *expectation step* evaluates the posterior density of the latent variables, $p(\mathbf{z}|\mathbf{x}, \boldsymbol{\theta}_t)$, keeping $\boldsymbol{\theta}_t$ fixed. If the posterior is not analytically tractable, variational approximation $q(\mathbf{z})$ can be used to obtain a lower bound for the likelihood in Equation 3.7. The *maximization step* optimizes the model parameters $\boldsymbol{\theta}$ with respect to the following objective function:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}_t) = \int_{\mathbf{z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_t) logp(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) d\mathbf{Z}. \tag{3.8}$$

This is the expectation of the *complete-data log-likelihood* $logp(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ over the latent variable density $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}_t)$, obtained from the previous expectation step. The new parameter estimate is then

$$\boldsymbol{\theta}_{t+1} = argmax_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}_t).$$

The expectation and maximization steps determine an iterative learning procedure where the latent variable density and model parameters are iteratively updated until convergence. The maximization step will also increase the target likelihood of Equation 3.7, but potentially with a remarkably smaller computational cost (Dempster et al., 1977). In contrast to the marginal likelihood in Equation 3.7, the complete-data likelihood in Equation 3.8 is logarithmized before integration in the maximization step. When the joint distribution $p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})$ belongs to the exponential family, the logarithm will cancel the exponential in algebraic manipulations. This can considerably simplify the maximization step. When the likelihoods in the optimization are of suitable form, the iteration steps can be solved analytically, which can considerably reduce required evaluations of

the objective function. Convergence is guaranteed, if the optimization can increase the likelihood at each iteration. However, the identification of a global optimum is not guaranteed in the EM algorithm.

Incorporating prior information of the parameter values through Bayesian priors can be used to avoid overfitting and focus the modeling on particular features in the data, as in the regularized dependency modeling framework of Publication 4, where the EM algorithm is used to learn Gaussian latent variable models.

Standard optimization methods

Optimization methods provide standard tools to implement selected learning procedures. Optimization algorithms are used to identify parameter values that minimize or maximize the objective function, either globally, or in local surroundings of the optimized value. Selection of optimization method depends on smoothness and continuity properties of the objective function, required accuracy, and available resources.

Gradient-based approaches optimize the objective function by assuming smooth, continuous topology over the probability density where setting the derivatives to zero will yield local optima. If a closed form solution is not available, it is often possible to estimate gradient directions in a given point. Optimization can then proceed by updating the parameters towards the desired direction along the gradient, gradually improving the objective function value in subsequent gradient ascent steps. So-called *quasi-Newton methods* use function values and gradients to characterize the optimized manifold, and to optimize the parameters along the approximated gradients. An appropriate step length is identified automatically based on the curvature of the objection function surface. The Broyden-Fletcher-Goldfarb-Shanno (BFGS) (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970) method is a quasi-Newton approach used for standard optimization tasks in this thesis.

3.3.2 Generalizability and overlearning

Probabilistic models are formulated in terms of probability distributions over the sample space and parameter values. This forms the basis for generalization to new, unobserved events. A generalizable model can describe essential characteristics of the underlying process that generated the observations; a generalizable model is also able to characterize future observations. *Overlearning*, or *overfitting* refers to models that describe the training data well, but do not generalize to new observations. Such models describe not only the general processes underlying the observations, but also noise in the particular observations. Avoiding overfitting is a central aspect in modeling. Overlearning is particularly likely when training data is scarce. While overfitting could in principle be avoided by collecting more data, this is often not feasible since the cost of data collection can be prohibitively large.

Generalizability can be measured by investigating how accurately the model describes new observations. A standard approach is to split the data into a *training set*, used to learn the model, and a *test set*, used to measure model performance on unseen observations that were not used for training. In *cross-validation* the test is repeated with several different learning and test sets to assess the variability in the testing procedure. Cross-validation is used for instance in Publication 5 of

this thesis. *Bootstrap analysis* (see, for instance, Efron and Tibshirani, 1994) is another widely used approach to measure model performance. The observed data is viewed as a finite realization of an underlying probability density. New samples from the underlying density are obtained by re-sampling the observed data points with replacement to simulate variability in the original data; observations from the more dense regions of the probability space become re-sampled more often than rare events. Each bootstrap sample resembles the probability density of the original data. Modeling multiple data sets obtained with the bootstrap helps to estimate the sensitivity of the model to variations in the data. Bootstrap is used to assess model performance in Publication 6.

3.3.3 Regularization and model selection

In general, increasing model complexity will yield more flexible models, which have higher descriptive power but are, on the other hand, more likely to overfit. Therefore relatively simple models can often outperform more complex models in terms of generalizability. A compromise between simplicity and descriptive power can be obtained by imposing additional constraints or soft penalties in the modeling to prefer compact solutions, but at the same time retain the descriptive power of the original, flexible model family. This is called *regularization*. Regularization is particularly important when the sample size is small, as demonstrated for instance in Publication 4, where explicit and theoretically principled regularization is achieved by setting appropriate priors on the model structure and parameter values. The priors will then affect the MAP estimate of the model parameters. One commonly used approach is to prefer *sparse* solutions that allow only a small number of the potential parameters to be employed at the same time to model the data (see e.g. Archambeau and Bach, 2008). A family of probabilistic approaches to balance between model fit and model complexity is provided by information-theoretic criteria (see e.g. Gelman et al., 2003). The *Bayesian Information Criterion (BIC)* is a widely used information criterion that introduces a penalty term on the number of model parameters to prefer simpler models. The log-likelihood $\mathcal{L}$ of the data, given the model, is balanced by a measure of model complexity, $q log(N)$, in the final objective function $-2\mathcal{L} + q log(N)$. Here q denotes the number of model parameters and N is the constant sample size of the investigated data set. The BIC has been criticized since it does not address changes in prior distributions, and its derivation is based on asymptotic considerations that hold only approximately with finite sample size (see e.g. Bishop, 2006). On the other hand, BIC provides a principled regularization procedure that is easy to implement. In this thesis, the BIC has been used to regularize the algorithms in Publication 3.

3.3.4 Validation

After learning a probabilistic model, it is necessary to confirm the quality of the model and verify potential findings in further, independent experiments. *Validation* refers to a versatile set of approaches used to investigate model performance, as well as in model criticism, comparison and selection. Internal and external approaches provide two complementary categories for model validation. *Internal validation* refers to procedures to assess model performance based on training data alone. For instance, it is possible to estimate the sensitivity of the model to initialization, parameterization, and variations in the data, or convergence of

the learning process. Internal analysis can help to estimate the weaknesses and generalizability of the model, and to compare alternative models. Bootstrap and cross-validation are widely used approaches for internal validation and the analysis of model performance (see e.g. Bishop, 2006). Bootstrap can provide information about the sensitivity of the results to sampling effects in the data. Cross-validation provides information about the model generalization performance and robustness by comparing predictions of the model to real outcomes. *External validation* approaches investigate model predictions and fit on new, independent data sets and experiments. Exploratory analysis of high-throughput data sets often includes massive multiple testing, and provides potentially thousands of automatically generated hypotheses. Only a small set of the initial findings can be investigated more closely by human intervention and costly laboratory experiments. This highlights the need to prioritize the results and assess the uncertainty in the models.

Chapter 4

Reducing uncertainty in high-throughput microarray studies

As far as the laws of mathematics refer to reality, they are not certain, as far as they are certain, they do not refer to reality.

A. Einstein (1956)

Gene expression microarrays are currently the most widely used technology for genome-wide transcriptional profiling, and they constitute the main source of data in this thesis. An overview of microarray technology is provided in Section 2.3.1. Microarray measurements are associated with high levels of noise from technical and biological sources. Appropriate preprocessing techniques can help to reduce noise and obtain reliable measurements, which is the crucial starting point for any data analysis task. This chapter presents the first main contribution of the thesis, preprocessing techniques that utilize side information in genomic sequence databases and microarray data collections in order to improve the accuracy of high-throughput gene expression data. The chapter is organized as follows: Section 4.1 gives an overview of the various sources of noise in high-throughput microarray studies. Section 4.2 introduces a strategy for noise reduction based on side information in external genomic sequence databases. Section 4.3 extends this model by describing a model-based approach that additionally combines statistical evidence across multiple microarray experiments in order to provide quantitative information of probe performance and utilizes this information to improve the reliability of high-throughput observations. The results are summarized in Section 4.4.

4.1 Sources of uncertainty

Measurement data obtained with novel high-throughput technologies comes with high levels of uncontrolled biological and technical variation. This is often called *noise* as it obscures the measurements, and adds potential bias and variance on the *signal* of interest. Biological noise is associated with natural biological variation between cell populations, cellular processes and individuals. Single-nucleotide

polymorphisms, alternative splicing and non-specific hybridization add biological variation in the data (Dai et al., 2005; Zhang et al., 2005). More technical sources of noise in the measurement process include RNA extraction and amplification, experiment-specific variation, as well as platform- and laboratory-specific effects (Choi et al., 2003; MAQC Consortium, 2006; Tu et al., 2002).

A significant source of noise on gene expression arrays comes from individual probes that are designed to measure the activity of a given transcript in a biological sample. Figure 4.1A shows probe-level observations of differential gene expression for a collection of probes designed to target the same mRNA transcript. One of the probes is highly contaminated and likely to add unrelated variation to the analysis. A number of factors affect probe performance. For instance, it has been reported in Publication 1 and elsewhere (Hwang et al., 2004; Mecham et al., 2004b) that a large portion of microarray probes may target unintended mRNA sequences. Moreover, although the probes have been designed to uniquely hybridize with their intended mRNA target, remarkable cross-hybridization with the probes by single-nucleotide polymorphisms (Dai et al., 2005; Sliwerska et al., 2007) and other mRNAs with closely similar sequences (Zhang et al., 2005) have been reported; high-affinity probes with high GC-content may have higher likelihood of cross-hybridization with nonspecific targets (Mei et al., 2003). Alternative splicing (MAQC Consortium, 2006) and mRNA degradation (Auer et al., 2003) may cause differences between probes targeting different positions of the gene sequence. Such effects will contribute to probe-level contamination in a probe- and condition-specific manner. However, sources of probe-level noise are still poorly understood (Irizarry et al., 2005; Li et al., 2005) despite their importance for expression analysis and probe design.

High levels of noise set specific challenges for analysis. Better understanding of the technical aspects of the measurement process will lead to improved analytical procedures and ultimately to more accurate biological results (Reimers, 2010). Publication 2 provides computational tools to investigate probe performance and the relative contributions of the various sources of probe-level contamination on short oligonucleotide arrays.

4.2 Preprocessing microarray data with side information

Preprocessing of the *raw data* obtained from the original measurements can help to reduce noise and improve comparability between microarray experiments. Preprocessing can be defined in terms of statistical transformations on the raw data, and this is a central part of data analysis in high-throughput studies. This section outlines the standard preprocessing steps for short oligonucleotide arrays, the main source of transcriptional profiling data in this thesis. However, the general concepts also apply to other microarray platforms (Reimers, 2010).

Standard preprocessing steps

A number of preprocessing techniques for short oligonucleotide arrays have been introduced (Irizarry et al., 2006; Reimers, 2010). The standard preprocessing steps in microarray analysis include quality control, background correction, normalization and summarization.

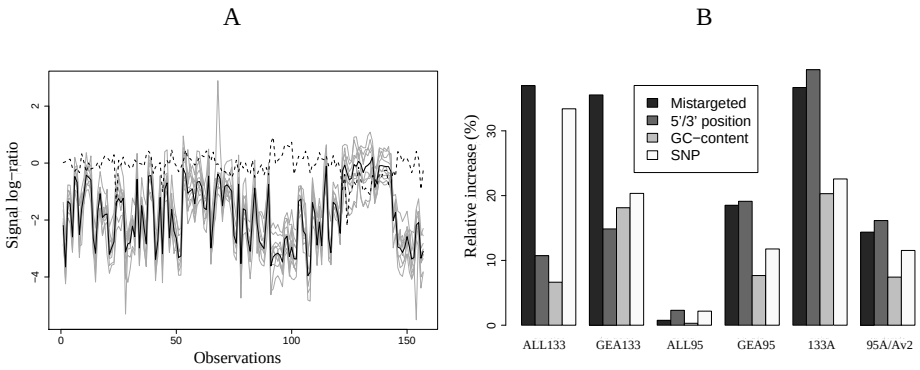


Figure 4.1: **A** Example of a probe set that contains a probe with high contamination levels (dashed line) detected by the probabilistic RPA model. The probe-level observations of differential gene expression for the different probes that measure the same target transcript are indicated by gray lines. The black line shows the estimated signal of the target transcript across a number of conditions. **B** Increase in the average variance of the probes associated with the investigated noise sources: mistargeted probes having errors in the genomic alignment, most 5'/3' probes of each probe set, GC-rich, and SNP-associated probes. The variances were estimated by RPA and describe the noise level of the probes. The results are shown for the individual ALL and GEA data sets, and for their combined results on both platforms (133A and 95A/Av2). ©IEEE. Reprinted with permission from Publication 2.

Microarray quality control is used to identify arrays with remarkable experimental defects, and to remove them from subsequent analysis. The typical tests consider RNA degradation levels and a number of other summary statistics to guarantee that the array data is of reasonable quality. The arrays that pass the microarray quality control are preprocessed further. Each array typically has spatial biases that vary smoothly across the array, arising from technical factors in the experiment. *Background correction* is used to detect and remove such spatial effects from the array data, and to provide a uniform background signal, enhancing the comparability of the probe-level observations between different parts of the array. Moreover, background correction can estimate the general noise level on the array; this helps to detect probes whose signal differs significantly from the background noise. Robust multi-array averaging (RMA) is one of the most widely used approaches for preprocessing short oligonucleotide array data (Irizarry et al., 2003a). The background correction in RMA is based on a global model for probe intensities. The observed intensity, Y, is modeled as a sum of an exponential signal component, S and Gaussian noise B. Background corrected data is then obtained as the expectation $\mathbb{E}_B(S|Y)$. While background correction makes the observations comparable within array, *normalization* is used to improve the comparability between arrays. Quantile normalization is a widely used method that forces all arrays to follow the same empirical intensity distribution (see e.g. Bolstad et al., 2003). Quantile normalization makes the measurements across different arrays comparable, assuming that the overall distribution of mRNA concentration is approximately the same in all cell populations. This has proven to be a feasible assumption in transcriptional profiling studies. As always, there are exceptions. For instance, human brain tissues have systematic differences in gene expression compared to other organs. On short oligonucleotide arrays, a number of probes

target the same transcript. In the final *summarization step*, the individual probe-level observations of each target transcript are summarized into a single summary estimate of transcript activity. Standard algorithmic implementations are available for each preprocessing step.

Probe-level preprocessing methods

Differences in probe characteristics cause systematic differences in probe performance. The use of several probes for each target leads to more robust estimates on transcript activity but it is clear that probe quality may significantly affect the results of a microarray study (Irizarry et al., 2003b). Widely used preprocessing algorithms utilize probe-specific parameters to model probe-specific effects in the probe summarization step. Some of the first and most well-known probe-level preprocessing algorithms include dChip/MBEI (Li and Wong, 2001), RMA (Irizarry et al., 2003a), and gMOS (Milo et al., 2003). Taking probe-level effects into account can considerably improve the quality of a microarray study (Reimers, 2010). Publications 1 and 2 incorporate side information of the probes to preprocessing, and introduce improved probe-level analysis methods for differential gene expression studies.

In order to introduce probe-level preprocessing methods in more detail, let us consider the probe summarization step of the RMA algorithm (Irizarry et al., 2003a). RMA has a Gaussian model for probe effects with probe-specific mean parameters and a shared variance parameter for the probes. The mean parameters characterize probe-specific binding affinities that cause systematic differences in the signal levels captured by each probe. Estimating the probe-specific effects helps to remove this effect in the final probeset-level summary of the probe-level observations. To briefly outline the algorithm, let us consider a collection of probes (a *probeset*) that measure the expression level of the same target transcript g in condition i. The probe-level observations are modeled as a sum of the true, underlying expression signal g_i, which is common to all probes, probe-specific binding affinity μ_j, and Gaussian noise ϵ. A probe-level observation for probe j in condition i is then modeled in RMA as

$$s_{ij} = g_i + \mu_j + \epsilon. \tag{4.1}$$

Measurements from multiple conditions are needed to estimate the probe-specific effects μ_j. RMA and other models that measure absolute gene expression have an important drawback: the probe affinity effects $\{\mu_j\}$ are unidentifiable. In order to obtain an identifiable model, the RMA algorithm includes an additional constraint that the probe affinity effects are zero on average: $\Sigma_j \mu_j = 0$. This yields a well-defined algorithm that has been shown to produce accurate measurements of gene expression in practical settings. Further extensions of the RMA algorithm include gcRMA, which has a more detailed chemical model for the probe effects (Wu and Irizarry, 2004), refRMA (Katz et al., 2006), which utilizes probe-specific effects derived from background data collections, and fRMA (McCall et al., 2010), which also models batch-specific effects in microarray studies. The estimation of unidentifiable probe affinities is a main challenge for most probe-level preprocessing models.

RMA and other probe-level models for short oligonucleotide arrays have been designed to estimate absolute expression levels of the genes. However, gene expression studies are often ultimately targeted at investigating *differential expression*

levels, that is, differences in gene expression between experimental conditions. Measurements of differential expression is obtained for instance by comparing the expression levels, obtained through the RMA algorithm or other methods, between different conditions. However, the summarization of the probe-level values is then performed prior to the actual comparison. Due to the unidentifiability of the probe affinity parameters in the RMA and other probe-level models, this is potentially suboptimal. Publication 1 demonstrates that reversing the order, i.e., calculating differential gene expression already at the probe level before probeset-level summarization, leads to improved estimates of differential gene expression. The explanation is that the procedure circumvents the need to estimate the unidentifiable probe affinity parameters. This is formally described in Publication 2, which provides a probabilistic extension of the Probe-level Expression Change Averaging (PECA) procedure of Publication 1. In PECA, a standard weighted average statistics summarizes the probe level observations of differential gene expression. PECA does not model probe-specific effects, but it is shown to outperform widely used probe-level preprocessing methods, such as the RMA, in estimating differential expression. Publication 2, considered in more detail in Section 4.3, provides an extended probabilistic framework that also models probe-specific effects.

Utilizing side information in transcriptome databases

Probe-level preprocessing models and microarray analysis can be further improved by utilizing external information of the probes (Eisenstein, 2006; Hwang et al., 2004; Katz et al., 2006). Although any given microarray is designed on most up-to-date sequence information available, rapidly evolving genomic sequence data can reveal inaccuracies in probe annotations when the body of knowledge grows. In recent studies, including Publication 1, a remarkable number of probes on various oligonucleotide arrays have been detected not to uniquely match their intended target (Hwang et al., 2004; Mecham et al., 2004a). A remarkable portion of probes on several popular microarray platforms in human and mouse did not match with their intended mRNA target, or were found to target unintended mRNA transcripts in the Entrez Nucleotide (Wheeler et al., 2005) sequence database in Publication 1 (Table 4.2). The observations are in general concordant with other studies, although the exact figures vary according to the utilized database and comparison details (Gautier et al., 2004; Mecham et al., 2004b). In this thesis, strategies are developed to improve microarray analysis with background information from genomic sequence databases, and with model-based analysis of microarray collections.

Probe verification is increasingly used in standard preprocessing, and to confirm the results of a microarray study. Matching the probe sequences of a given array to updated genomic sequence databases and constructing an alternative interpretation of the array data based on the most up-to-date genomic annotations has been shown to increase the accuracy and cross-platform consistency of microarray analyses in Publication 1 and elsewhere (Dai et al., 2005; Gautier et al., 2004).

Publication 1 combines probe verification with a novel probe-level preprocessing method, PECA, to suggest a novel framework for comparing and combining results across different microarray platforms. While huge repositories of microarray data are available, the data for any particular experimental condition is typically scarce, and coming from a number of different microarray platforms. Therefore

reliable approaches for integrating microarray data are valuable. Integration of results across platforms has proven problematic due to various sources of technical variation between array technologies. Matching of probe sequences between microarray platforms has been shown to increase the consistency of microarray measurements (Hwang et al., 2004; Mecham et al., 2004b). However, probe matching between array platforms guarantees only technical comparability (Irizarry et al., 2005). Probe verification against external sequence databases is needed to confirm that the probes are also biologically accurate. This can also improve the comparability across array platforms, as confirmed by the validation studies in Publication 1 (Figure 4.2A).

The PECA method of Publication 1 utilizes genomic sequence databases to reduce probe-level noise by removing erroneous probes based on updated genomic knowledge. The strategy relies on external information in the databases and can therefore only remove known sources of probe-level contamination. Publication 2 introduces a probabilistic framework to measure probe reliability directly based on microarray data collections. The analysis can reveal both well-characterized and unknown sources of probe-level contamination, and leads to improved estimates of gene expression. This model, coined Robust Probabilistic Averaging (RPA), also provides a theoretically justified framework for incorporating prior knowledge of the probes into the analysis.

Array type	Number of probes	Verified probes (%)
HG-U133 Plus2.0	604,258	58.2
HG-U133A	247,965	82.5
HG-U95Av2	199,084	82.6
MOE430 2.0	496,468	68.2
MG-U74Av2	197,993	73.1

Table 4.1: The proportion of sequence-verified probes on three popular human microarray platforms and two mouse platforms, as observed in Publication 1. Probes that matched to mRNA sequences corresponding to unique genes (defined by a GeneID identifier) in the Entrez database are considered verified. A remarkable portion of the probes on the investigated arrays did not match the Entrez transcript sequences, or had ambiguous targets.

4.3 Model-based noise reduction

Standard approaches for investigating probe performance typically rely on external information, such as genomic sequence data (see Mecham et al. 2004b; Zhang et al. 2005 and Publication 1) or physical models (Naef and Magnasco, 2003; Wu et al., 2005). However, such models cannot reveal probes with uncharacterized sources of contamination, such as cross-hybridization with alternatively spliced transcripts or closely related mRNA sequences. Vast collections of microarray data are available in public repositories. These large-scale data sets contain valuable information of both biological and technical aspects of gene expression studies. Publication 2 introduces a data-driven strategy to extract and utilize probe-level information in microarray data collections.

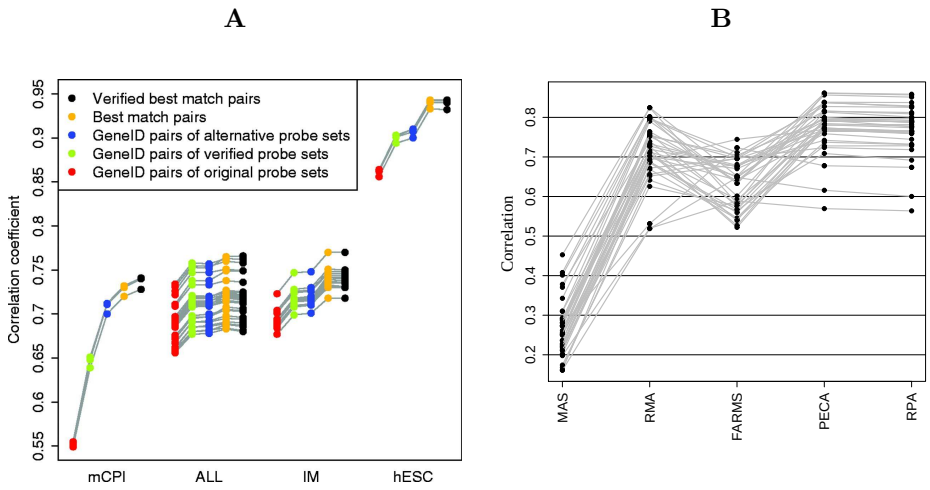


Figure 4.2: **A** Effect of sequence verification on comparability between microarray platforms. Correlations between RMA-preprocessed technical replicates on two array platforms where the same samples have been hybridized on the two array types. The Pearson correlations were calculated for each pair of arrays measuring the same biological sample. The gray lines show correlations obtained with the different probe matching criteria. In the hESC array comparison, the best match probe sets contained exactly the same probes on both array generations, which resulted in very high correlations. The advantages of probe verification and alternative mappings were largest when arrays with different probe collections were compared in the mCPI, ALL and IM array comparisons. **B** Reproducibility of signal estimates in real data sets between the technical replicates, i.e., the 'best match' probe sets between the HG-U95Av2 and HG-U133A platforms. The consistency was measured by the Pearson correlation between the pairs of arrays, to which the same sample was hybridized. ©Published by Oxford University Press. Reprinted with permission from Publication 1.

The model, *Robust Probabilistic Averaging (RPA)*, is a probabilistic preprocessing procedure that is based on explicit modeling assumptions to analyze probe reliability and quantify the uncertainty in measurement data based on gene expression data collections, independently of external information of the probes. The model can be viewed as a probabilistic extension of the probe-level preprocessing approach for differential gene expression studies presented in Publication 1. The explicit Bayesian formulation quantifies the uncertainty in the model parameters, and allows the incorporation of prior information concerning probe reliability into the analysis. RPA provides estimates of probe reliability, and a probeset-level estimate of differential gene expression directly from expression data and independently of the noise source. The RPA model is independent of physical models or external and constantly updated information such as genomic sequence data, but provides a framework for incorporating such prior information of the probes in gene expression analysis.

Other probabilistic methods for microarray preprocessing include BGX (Hein et al., 2005), gMOS (Milo et al., 2003) and its extensions (Liu et al., 2005). The key difference to the RPA procedure of Publication 2 is that these methods are designed to provide probeset-level summaries of absolute gene expression levels, and suffer from the same unidentifiability problem of probe affinity parameters as the RMA algorithm (Irizarry et al., 2003a). In contrast, RPA models probe-level estimates of differential gene expression. This removes the unidentifiability

issue, which is advantageous when the objective is to compare gene expression levels between experimental conditions. Another important difference is that the other preprocessing methods do not provide explicit estimates of probe-specific parameters, or tools to investigate probe performance. Publication 2 assigns an explicit probabilistic measure of reliability to each probe. This gives tools to analyze probe performance and to guide probe design.

Robust Probabilistic Averaging

Let us now consider in more detail the probabilistic preprocessing framework, RPA, introduced in Publication 2. Probe performance is ultimately determined by its ability to accurately measure the expression level of the target transcript, which is unknown in practical situations. Although the performance of individual probes varies, the collection of probes designed to measure the same transcript will provide ground truth for assessing probe performance (Figure 4.1A). RPA captures the shared signal of the probes within a probeset, and assumes that the shared signal characterizes the expression of the common target transcript of the probes. The reliability of individual probes is estimated with respect to the strongest shared signal of the probes. RPA assumes normally distributed probe effects, and quantifies probe reliability based on probe variance around the probeset-level signal across a large number of arrays. This extends the formulation of the RMA model in Equation 4.1 by introducing an additional probe-specific Gaussian noise component:

$$s_{ij} = g_i + \mu_j + \varepsilon_{ij}. \tag{4.2}$$

In contrast to RMA, the variance is probe-specific in this model, and distributed as $\varepsilon_{ij} \sim N(0, \tau_j^2)$. The variance parameters $\{\tau_j^2\}$ are of interest in probe reliability analysis; they reflect the noise level of the probe, in contrast to probe-level preprocessing methods that focus on estimating the unidentifiable mean parameter of the Gaussian noise model, corresponding to probe affinity (see e.g. Irizarry et al., 2003a; Li and Wong, 2001). In Publication 2, probe-level calculation of differential expression avoids the need to model unidentifiable probe affinities, the key probe-specific parameter in other probe-level preprocessing methods. More formally, the unidentifiable probe affinity parameters $\mu_\cdot$ cancel out in RPA when the signal log-ratio between a user-specified 'reference' array and the remaining arrays is computed for each probe: the differential expression signal between arrays $t = \{1, \ldots, T\}$ and the reference array c for probe j is obtained by $m_{tj} = s_{tj} - s_{cj} = g_t - g_c + \varepsilon_{tj} - \varepsilon_{cj} = d_t + \varepsilon_{tj} - \varepsilon_{cj}$. In vector notation, the differential expression profile of probe j across the T arrays is then written as $\mathbf{m}_j = \mathbf{d} + \boldsymbol{\varepsilon}_j$, i.e., a noisy observation of the true underlying differential expression signal $\mathbf{d}$ and probe-specific noise $\boldsymbol{\varepsilon}_j$.

The unidentifiable probe affinity parameters cancel out in the RPA model of Publication 2. This can partly explain the previous empirical observations that calculating differential expression already at probe-level improves the analysis of differential gene expression (Zhang et al., 2002; Elo et al., 2005). However, the previous models are non-probabilistic preprocessing methods that do not aim at quantifying the uncertainty in the probes. Use of a single parameter for probe effects in RPA also gives more straightforward interpretations of probe reliability.

Posterior estimates of the model parameters are derived to estimate probe reliability and differential gene expression. The differential expression vector $\mathbf{d} =$

$\{d_t\}$ and the probe-specific variances $\boldsymbol{\tau}^2 = \{\tau_j^2\}$ are estimated simultaneously. The posterior density of the model parameters is obtained from the likelihood of the data and the prior according to Bayes' rule (Equation 3.3) as

$$p(\mathbf{d}, \boldsymbol{\tau}^2|\mathbf{m}) \sim p(\mathbf{m}|\mathbf{d}, \boldsymbol{\tau}^2)p(\mathbf{d}, \boldsymbol{\tau}^2). \tag{4.3}$$

To obtain this posterior, let us consider the likelihood $p(\mathbf{m}|\mathbf{d}, \boldsymbol{\tau}^2)$ of the data and the prior $p(\mathbf{d}, \boldsymbol{\tau}^2)$ of the model parameters. The noise on the selected control array ε_{cj} is a latent variable, and marginalized out in the model to obtain the likelihood:

$$\begin{aligned} p(\mathbf{m}|\mathbf{d}, \boldsymbol{\tau}^2) &= \prod_{tj} \int N(m_{tj}|d_t - \varepsilon_{cj}, \tau_j^2) N(\varepsilon_{cj}|0, \tau_j^2) d\varepsilon_{cj} \\ &\sim \prod_j (2\pi\tau_j^2)^{-\frac{T}{2}} exp(-\frac{\sum_t (m_{tj} - d_t)^2 - \frac{[\sum_t (m_{tj} - d_t)]^2}{T+1}}{2\tau_j^2}). \end{aligned} \tag{4.4}$$

Let us assume independent priors, $p(\mathbf{d}, \boldsymbol{\tau}^2) = p(\mathbf{d})p(\boldsymbol{\tau}^2)$, flat non-informative prior $p(\mathbf{d}) \sim 1$ and conjugate priors for the variance parameters in $\boldsymbol{\tau}^2$ (inverse Gamma function, see Gelman et al. 2003). With these standard assumptions, the prior takes the form

$$p(\mathbf{d}, \boldsymbol{\tau}^2) \sim \prod_j IG(\tau_j^2; \alpha_j, \beta_j), \tag{4.5}$$

where α_j and β_j are the shape and scale parameters of the inverse Gamma distribution. Prior information of the probes can be incorporated in the analysis through these parameters. Probe-level differential expression is then described by two sets of parameters; the differential gene expression vector $\mathbf{d} = [d_1 \dots d_T]$, and the probe-specific variances $\boldsymbol{\tau}^2 = [\tau_1^2 \dots \tau_J^2]$. High variance τ_j^2 indicates that the probe-level observation $\mathbf{m}_j$ is strongly deviated from the estimated true signal $\mathbf{d}$. Denoting $\hat{\alpha}_j = \alpha_j + \frac{T}{2}$ and $\hat{\beta}_j = \beta_j + \frac{1}{2}\sum_t (m_{tj} - d_t)^2 - \frac{1}{2}\frac{(\sum_t (m_{tj} - d_t))^2}{T+1}$, the posterior of the model parameters in Equation 4.3 takes the form

$$p(\mathbf{d}, \boldsymbol{\tau}^2|\mathbf{m}) \sim \prod_j (\tau_j^2)^{-(\hat{\alpha}_j + 1)} exp(-\frac{\hat{\beta}_j}{\tau_j^2}). \tag{4.6}$$

The formulation allows estimating the uncertainty in the expression estimates and probe-level parameters. In practice, a MAP point estimate of the parameters, obtained by maximizing the posterior, is often sufficient. In the limit of a large sample size ($T \to \infty$), the model will converge to estimating ordinary mean and variance parameters. With limited sample sizes that are typical in microarray studies the prior parameters provide regularization that makes the probabilistic formulation more robust to overfitting and local optima, compared to direct estimation of the mean and variance parameters. Moreover, the probabilistic analysis takes the uncertainty in the data and model parameters into account in an explicit manner.

The model also provides a principled framework for incorporating prior knowledge probe reliability in microarray preprocessing through the probe-specific hyperparameters α, β. Estimation and use of probe-specific effects from external microarray data collections has been previously suggested in the context of the

refRMA method by Katz et al. (2006), where such side information was shown to improve gene expression estimates. The RPA method of Publication 2 provides an alternative probabilistic treatment.

Model validation

The probabilistic RPA model introduced in Publication 2 was validated by comparing the preprocessing performance to other preprocessing methods, and additionally by comparing the estimates of probe-level noise to known sources of probe-level contamination. The comparison methods include the FARMS (Hochreiter, 2006), MAS5 (Hubbell et al., 2002), PECA (Publication 1), and RMA (Irizarry et al., 2003a) preprocessing algorithms. FARMS has a more detailed model for probe effects than the other methods, and it contains implicitly a similar probe-specific variance parameter than our RPA model. FARMS is based on a factor analysis model, and is defined as $s_{ij} = z_i\lambda_j + \mu_j + \varepsilon_{ij}$, where z_i captures the underlying gene expression. In contrast to RMA and RPA that have a single probe-specific parameter, FARMS has three probe-specific parameters $\{\lambda_j, \mu_j, \varepsilon_{ij}\}$. MAS5 is a standard preprocessing algorithm provided by the array manufacturer. The algorithm performs local background correction, utilizes so-called mismatch probes to control for non-specific hybridization, and scales the data from each array to the same average intensity level to improve comparability across arrays. MAS5 summarizes probe-level observations of absolute gene expression levels using robust summary statistics, Tukey biweight estimate, but unlike FARMS, RMA and RPA, MAS5 does not model probe-specific effects.

The preprocessing performance of these methods was investigated in spike-in experiments where certain target transcripts measured by the array have been spiked in at known concentrations, as well as on real data sets. The results from the spike-in experiments were compared in terms of receiver operating characteristics (ROC). The standard RMA, PECA (Publication 1) and RPA (Publication 2) had comparable performance in spike-in data, and they outperformed the MAS5 (Hubbell et al., 2002) and FARMS (Hochreiter, 2006) preprocessing algorithms in estimating differential gene expression. On real data sets, PECA and RPA outperformed the other methods, providing higher reproducibility between technical replicates measured on different microarray platforms (Figure 4.2B).

In contrast to standard preprocessing algorithms, RPA provides explicit quantitative estimates of probe performance. The model has been validated on widely used human whole-genome arrays by comparing the estimates of probe reliability with known probe-level error sources: errors in probe-genome alignment, interrogation position of a probe on the target sequence, GC-content, and the presence of SNPs in the probe target sequences; a good model for assessing probe reliability should detect probes contaminated by the known error sources. The results from our analysis can be used to characterize the relative contribution of different sources of probe-level noise (Figure 4.1B). In general, the probes with known sources of contamination were more noisy than the other probes, with 7-39% increase in the average variance, as detected by RPA. Any single source of error seems to explain only a fraction of the most highly contaminated probes. A large portion (35-60%) of the detected least reliable probes were not associated with the investigated known noise sources. This suggests that previous methods that remove probe-level noise based on external information, such as genomic alignments will fail to detect a significant portion of poorly performing probes. The RPA

model of Publication 2 provides rigorous algorithmic tools to investigate the various probe-level error sources. Better understanding of the factors affecting probe performance can advance probe design and contribute to reducing probe-related noise in future generations of gene expression arrays.

4.4 Conclusion

The contributions presented in this Chapter provide improved preprocessing strategies for differential gene expression studies. The introduced techniques utilize probe-level analysis, as well as side information in sequence and microarray databases. Probe-level studies have led to the establishment of probe verification and alternative microarray interpretations as a standard step in microarray preprocessing and analysis. The alternative interpretations for microarray data based on updated genomic sequence data (Gautier et al., 2004; Dai et al., 2005) are now implemented as routine tools in popular preprocessing algorithms such as the RMA, or the RPA method of Publication 2. The probe-level analysis strategy has been recently extended to exon array context, where expression levels of alternative splice variants of the same genes are compared under particular experimental conditions. The probe-level approach has shown superior preprocessing performance also with exon arrays (Laajala et al., 2009). A convenient access to the algorithmic tools developed in Publications 1 and 2 for microarray preprocessing and probe-level analysis is provided by the accompanied open source implementation in BioConductor.[1]

[1] http://www.bioconductor.org/packages/release/bioc/html/RPA.html

Chapter 5

Global analysis of the human transcriptome

> *When we try to pick out anything by itself, we find that it is bound fast by a thousand invisible cords that cannot be broken, to everything in the universe.*
>
> J. Muir (1869)

Measurements of transcriptional activity provide only a partial view to physiological processes, but their wide availability provides a unique resource for investigating gene activity at a genome- and organism-wide scale. Versatile and carefully controlled *gene expression atlases* have become available for normal human tissues, cancer as well as for other diseases (see, for instance, Kilpinen et al., 2008; Lukk et al., 2010; Roth et al., 2006; Su et al., 2004). These data sources contain valuable information about shared and unique mechanisms between disparate conditions, which is not available in smaller and more specific experiments (Lage et al., 2008; Scherf et al., 2000). While standard methods for gene expression analysis have focused on comparisons between particular conditions, versatile transcriptome atlases allow for global organism-wide characterization of transcriptional activation patterns (Levine et al., 2006). Novel methodological approaches are needed in order to realize the full potential of these information sources, as many traditional methods for expression analysis are not applicable to versatile large-scale collections. This chapter provides an overview to current approaches for global transcriptome analysis in Section 5.1 and introduces the second main contribution of the thesis, a novel exploratory approach that can be used to investigate context-specific responses in genome-scale interaction networks across organism-wide collections of measurement data in Section 5.2. The conclusions are summarized in Section 5.3.

5.1 Standard approaches

Global observations of transcriptional activity reflect known and previously uncharacterized cell-biological processes. Exploratory analysis of the transcriptome can provide research hypotheses and material for more detailed investigations.

Widely-used standard approaches for global transcriptome analysis include various clustering, dimensionality reduction and visualization techniques (see e.g. Huttenhower and Hofmann, 2010; Polanski and Kimmel, 2007; Quackenbush, 2001). The large data collections open up new possibilities to investigate functional relatedness between physiological conditions, disease states, as well as cellular processes, and to discover previously uncharacterized connections and functional mechanisms (Bergmann et al., 2004; Kilpinen et al., 2008; Lukk et al., 2010).

Gene expression studies have traditionally focused on the analysis of relatively small and targeted data sets, such as particular diseases or cell types. A typical objective is to detect genes, or gene groups, that are differentially expressed between particular conditions, for instance to predict disease outcomes, or to identify potentially unknown disease subtypes. The increasing availability of large and versatile transcriptome collections that may cover thousands of experimental conditions allows global, data-driven analysis, and the formulation of novel research questions where the traditional analysis methods are often insufficient (Huttenhower and Hofmann, 2010).

A variety of approaches have been proposed and investigated in the recent years in the global transcriptome analysis context. An actively studied modeling problem in transcriptome analysis is the *discovery of transcriptional modules*, i.e., identification of coherent gene groups that show coordinated transcriptional responses under particular conditions (Segal et al., 2003a, 2004; Stuart et al., 2003). Models have also been proposed to predict gene regulators (Segal et al., 2003b), and to infer cellular processes and networks based on transcriptional activation patterns (Friedman, 2004; Segal et al., 2003c). An increasing number of models are being developed to integrate transcriptome measurements to other sources of genomic information, such as regulation and interactions between the genes to detect and characterize cellular processes and disease mechanisms (Barash and Friedman, 2002; Chari et al., 2010; Vaske et al., 2010). Findings from transcriptome analysis have potential biomedical implications, as in Lamb et al. (2006), where chemically perturbed cancer cell lines were screened to enhance the detection of drug targets based on shared functional mechanisms between disparate conditions, or in Sørlie et al. (2001), where cluster analysis of cancer patients based on genome-wide transcriptional profiling experiments led to the discovery of a novel breast cancer subtype. In the remainder of this section, the modeling approaches that are particularly closely related to the contributions of this thesis are considered in more detail.

Investigating known processes

A popular strategy for genome-wide gene expression analysis is to consider known biological processes and their activation patterns across diverse collections measurement data from various experimental conditions. Biomedical databases contain a variety of information concerning genes and their interactions. For instance, the Gene Ontology database (Ashburner et al., 2000) provides functional and molecular classifications for the genes in human and a number of other organisms. Other categories are based on micro-RNA regulation, chromosomal locations, chemical perturbations and other features (Subramanian et al., 2005). Joint analysis of functionally related genes can increase the statistical power of the analysis. So-called *gene set-based approaches* are typically designed to test differential expression between two particular conditions (Goeman and Buhlmann, 2007; Nam

and Kim, 2008), but they can also be used to build global maps of transcriptional activity of the known processes (Levine et al., 2006). However, gene set-based approaches typically ignore more detailed information of the interactions between individual genes. Pathway and interaction databases contain more detailed information concerning molecular interactions and cell-biological processes (Kanehisa et al., 2008; Vastrik et al., 2007). *Network-based methods* utilize relational information of the genes to guide expression analysis. For instance, Draghici et al. (2007) demonstrated that taking into account aspects of pathway topology, such as gene and interaction types, can improve the estimation of pathway activity between two predefined conditions. Another recent approach which utilizes pathway topology in inferring pathway activity is PARADIGM (Vaske et al., 2010), which also integrates other sources of genomic information in pathway analysis. However, these methods have been designed for the analysis of particular experimental conditions, rather than comprehensive expression atlases. MATISSE (Ulitsky and Shamir, 2007) is a network-based approach that searches for functionally related genes that are connected in the network, and have correlated expression profiles across many conditions. The potential shortcoming of this approach is that it assumes global correlation across all conditions between the interacting genes, while many genes can have multiple, context-sensitive functional roles. Different conditions induce different responses in the same genes, and the definition of 'gene set' is vague (Montaner et al., 2009; Nacu et al., 2007). Therefore methods have been suggested to identify 'key condition-responsive genes' of predefined gene sets (Lee et al., 2008), or to decompose predefined pathways into smaller and more specific functional modules (Chang et al., 2009). These approaches rely on predefined functional classifications for the genes. The data-driven analysis in Publication 3 provides a complementary approach where the gene sets are learned directly from the data, guided by prior knowledge of genetic interactions. This avoids the need to refine suboptimal annotations, and enables the discovery of new processes. The findings demonstrate that simply measuring whether a gene set, or a network, is differentially expressed between particular conditions is often not sufficient for measuring the activity of cell-biological processes. Since gene function and interactions are regulated in a context-specific manner, it is important to additionally characterize how, and in which conditions the expression changes. Global analysis of transcriptional activation patterns interaction networks, introduced in Publication 3, can address such questions.

Biclustering and subspace clustering

Approaches that are based on previously characterized genes and processes are biased towards well-characterized phenomena. This limits their value in *de novo* discovery of functional patterns. Unsupervised methods provide tools for such analysis, but often with an increased computational cost and a higher proportion of false positive findings.

Cluster analysis is widely used for unsupervised analysis of gene expression data, providing tools for class discovery, gene function prediction and for visualization purposes. Examples of widely used clustering approaches include hierarchical clustering and K-means (see e.g. Polanski and Kimmel, 2007). Clustering of patient samples with similar expression profiles has led to the discovery of novel cancer subtypes with biomedical implications (Sørlie et al., 2001); clustering of genes with coordinated activation patterns can be used, for instance, to predict novel

functional associations for poorly characterized genes (Allocco et al., 2004). The self-organizing map (Kohonen, 1982, 2001) is a related approach that provides efficient tools to *visualize* high-dimensional data on lower-dimensional displays, with particular applications in transcriptional profiling studies (Tamayo et al., 1999; Törönen et al., 1999). The standard clustering methods are based on comparison of global expression patterns, and therefore are relatively coarse tools for analyzing large transcriptome collections. Different genes respond in different ways, as well as in different conditions. Therefore it is problematic to find clusters in high-dimensional data spaces, such as in whole-genome expression profiling studies; different gene groups can reveal different relationships between the samples. Detection of smaller, coherent subspaces with a particular structure can be useful in biomedical applications, where the objective is to identify sets of interesting genes for further analysis. Both genes and the associated conditions may be unknown, and the learning task is to detect them from the data. This can help, for instance, in identifying responses to drug treatments in particular genes (Ihmels et al., 2002; Tanay et al., 2002), or in identifying functionally coherent transcriptional modules in gene expression databases (Segal et al., 2004; Tanay et al., 2005).

Subspace clustering methods (Parsons et al., 2004) provide a family of algorithms that can be used to identify subsets of dependent features revealing coherent clustering for the samples; this defines a subspace in the original feature space. Subspace clustering models are a special case of a more general family of *biclustering* algorithms (Madeira and Oliveira, 2004). Closely related models are also called co-clustering (Cho et al., 2004), two-way clustering Gad et al. (2000), and plaid models (Lazzeroni and Owen, 2002). Biclustering methods provide general tools to detect co-regulated gene groups and associated conditions from the data, to provide compact summaries and to aid interpretation of transcriptome data collections. Biclustering models enable the discovery of *gene expression signatures* (Hu et al., 2006) that have emerged as a central concept in global expression analysis context. A signature describes a co-expression state of the genes, associated with particular conditions. Established signatures have been found to be reliable indicators of the physiological state of a cell, and commercial signatures have become available for routine clinical practice (Nuyten and van de Vijver, 2008). However, the established signatures are typically designed to provide optimal classification performance between two particular conditions. The problem with the classification-based signatures is that their associations to the underlying physiological processes are not well understood (Lucas et al., 2009). In Publication 3 the understanding is enhanced by deriving transcriptional signatures that are explicitly connected to well-characterized processes through the network.

Role of side information

Standard clustering models ignore prior information of the data, which could be used to supervise the analysis, to connect the findings to known processes, as well as to improve scalability. For instance, standard model-based feature selection, or subspace clustering techniques would consider all potential connections between the genes or features (Law et al., 2004; Roth and Lange, 2004). Without additional constraints on the solution space they can typically handle at most tens or hundreds of features, which is often insufficient in high-throughput genomics applications. Use of side information in clustering can help to guide unsupervised analysis, for instance based on known or potential interactions between the genes.

This has been shown to improve the detection of functionally coherent gene groups (Hanisch et al., 2002; Shiga et al., 2007; Ulitsky and Shamir, 2007; Zhu et al., 2005). However, while these methods provide tools to cluster the genes, they do not model differences between conditions. Extensions of biclustering models that can utilize relational information of the genes include cMonkey (Reiss et al., 2006) and a modified version of SAMBA biclustering (Tanay et al., 2004). However, cMonkey and SAMBA are application-oriented tools that rely on additional, organism-specific information, and their implementation is currently not available for most organisms, including that of the human. Further application-oriented models for utilizing side information in the discovery of transcriptional modules have recently been proposed for instance by Savage et al. (2010) and Suthram et al. (2010). Publication 3 introduces a complementary method where the exhaustively large search space is limited with side information concerning known relations between the genes, derived from genomic interaction databases. This is a general algorithmic approach whose applicability is not limited to particular organisms.

Other approaches

Prior information on the cellular networks, regulatory mechanisms, and gene function is often available, and can help to construct more detailed models of gene function and network analysis, as well as to summarize functional aspects of genomic data collections (Huttenhower et al., 2009; Segal et al., 2003b; Troyanskaya, 2005). Versatile transcriptome collections also enable *network reconstruction*, i.e., *de novo* discovery (Lezon et al., 2006; Myers et al., 2005) and augmentation (Novak and Jain, 2006) of genetic interaction networks. Other methodological approaches for global transcriptome analysis are provided by probabilistic latent variable models (Rogers et al., 2005; Segal et al., 2003a), hierarchical Dirichlet process algorithms (Gerber et al., 2007), as well as matrix and tensor computations (Alter and Golub, 2005). These methods provide further model-based tools to identify and characterize transcriptional programs by decomposing gene expression data sets into smaller, functionally coherent components.

5.2 Global modeling of transcriptional activity in interaction networks

Molecular interaction networks cover thousands of genes, proteins and small molecules. Coordinated regulation of gene function through molecular interactions determines cell function, and is reflected in transcriptional activity of the genes. Since individual processes and their transcriptional responses are in general unknown (Lee et al., 2008; Montaner et al., 2009), data-driven detection of condition-specific responses can provide an efficient proxy for identifying distinct transcriptional states of the network with potentially distinct functional roles. While a number of methods have been proposed to compare network activation patterns between particular conditions (Draghici et al., 2007; Ideker et al., 2002; Cabusora et al., 2005; Noirel et al., 2008), or to use network information to detect functionally related gene groups (Segal et al., 2003d; Shiga et al., 2007; Ulitsky and Shamir, 2007), general-purpose algorithms for a global analysis of context-specific network activation patterns in a genome- and organism-wide scale have been missing.

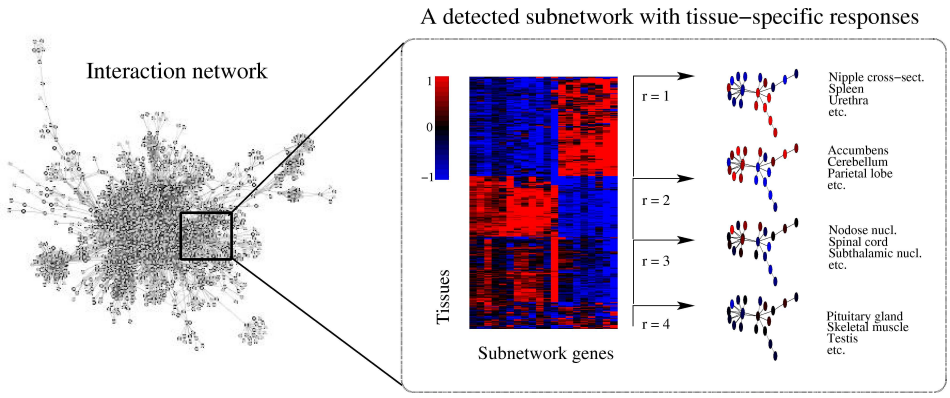


Figure 5.1: Organism-wide analysis of transcriptional responses in a human pathway interaction network reveals physiologically coherent activation patterns and condition-specific regulation. One of the subnetworks and its condition-specific responses, as detected by the NetResponse algorithm is shown in the Figure. The expression of each gene is visualized with respect to its mean level of expression across all samples. ©The Author 2010. Published by Oxford University Press. Reprinted with permission from Publication 3.

Publication 3 introduces and validates two general-purpose algorithms that provide tools for global modeling of transcriptional responses in interaction networks. The motivation is similar to biclustering approaches that detect functionally coherent gene groups that show coordinated response in a subset of conditions (Madeira and Oliveira, 2004). The network ties the findings more tightly to cell-biological processes, focusing the analysis and improving interpretability. In contrast to previous network-based biclustering models for global transcriptome analysis, such as cMonkey (Reiss et al., 2006) or SAMBA (Tanay et al., 2004), the algorithms introduced in Publication 3 are general-purpose tools, and do not depend on organism-specific annotations.

A two-step approach

The first approach in Publication 3 is a straightforward extension of network-based gene clustering methods. In this two-step approach, the functionally coherent subnetworks, and their condition-specific responses are detected in separate steps. In the first step, a network-based clustering method is used to detect functionally coherent subnetworks. In Publication 3, MATISSE, a state-of-the-art algorithm described in Ulitsky and Shamir (2007), is used to detect the subnetworks. MATISSE finds connected subgraphs in the network that have high internal correlations between the genes. In the second step, condition-specific responses of each identified subnetwork are searched for by a nonparametric Gaussian mixture model, which allows a data-driven detection of the responses. However, the two-step approach, coined MATISSE+, can be suboptimal for detecting subnetworks with particular condition-specific responses. The main contribution of Publication 3 is to introduce a second general-purpose algorithm, coined NetResponse, where the detection of condition-specific responses is used as the explicit key criterion for subnetwork search.

The NetResponse algorithm

The network-based search procedure introduced in Publication 3 searches for local *subnetworks*, i.e., functionally coherent network modules where the interacting genes show coordinated responses in a subset of conditions (Figure 5.1). Side information of the gene interactions is used to guide modeling, but the algorithm is independent of predefined classifications for genes or measurement conditions. Transcriptional responses of the network are described in terms of subnetwork activation. Regulation of the subnetwork genes can involve simultaneous activation and repression of the genes: sufficient amounts of mRNA for key proteins has to be available while interfering genes may need to be silenced. The model assumes that a given subnetwork n can have multiple transcriptional states, associated with different physiological contexts. A transcriptional state is reflected in a unique expression signature $\mathbf{s}^{(n)}$, a vector that describes the expression levels of the subnetwork genes, associated with the particular transcriptional state. Expression of some genes is regulated at precise levels, whereas other genes fluctuate more freely. Given the state, expression of the subnetwork genes is modeled as a noisy observation of the transcriptional state. With a Gaussian noise model with covariance $\Sigma^{(n)}$, the observation is described by $\mathbf{x}^{(n)} \sim N(\mathbf{s}^{(n)}, \Sigma^{(n)})$. A given subnetwork can have $R^{(n)}$ latent transcriptional states indexed by r. In practice, the states, including their number $R^{(n)}$, are unknown, and they have to be estimated from the data. In a specific measurement condition, the subnetwork n can be in any one of the latent physiological states indexed by r. Associations between the observations and the underlying transcriptional states are unknown and they are treated as latent variables. Gene expression in subnetwork n is then modeled with a Gaussian mixture model:

$$\mathbf{x}^{(n)} \sim \sum_{r=1}^{R^{(n)}} w_r^{(n)} p(\mathbf{x}^{(n)} | \boldsymbol{\theta}_r), \tag{5.1}$$

where each component distribution p is assumed to be Gaussian with parameters $\boldsymbol{\theta}_r = \{\mathbf{s}_r^{(n)}, \boldsymbol{\Sigma}_r^{(n)}\}$. In practice, we assume a diagonal covariance matrix $\boldsymbol{\Sigma}_r^{(n)}$, leaving the dependencies between the genes unmodeled within each transcriptional state. Use of diagonal covariances is justified by considerable gains in computational efficiency when the detection of distinct responses is of primary interest. It is possible, however, that such simplified model will fail to detect certain subnetworks where the transcriptional levels of the genes have strong linear dependencies within the individual transcriptional states; signaling cascades could be expected to manifest such activation patterns, for instance. More detailed models of transcriptional activity could help to distinguish the individual states in particular when the transcriptional states are partially overlapping, but with increased computational cost. A particular transcriptional response is then characterized with the triple $\{\mathbf{s}_r^{(n)}, \boldsymbol{\Sigma}_r^{(n)}, w_r^{(n)}\}$. This defines the shape, fluctuations and frequency of the associated transcriptional state of subnetwork n. A posterior probability of each latent state can be calculated for each measurement sample from the Bayes' rule (Equation 3.3). The posterior probabilities can be interpreted as soft component memberships for the samples. A hard, deterministic assignment is obtained by selecting for each sample the component with the highest posterior probability.

The remaining task is to identify the subnetworks having such distinct transcriptional states. Detection of the distinct states is now used as a search criterion

for the subnetworks. In order to achieve fast computation, an agglomerative procedure is used where interacting genes are gradually merged into larger subnetworks. Initially, each gene is assigned in its own singleton subnetwork. Agglomeration proceeds by at each step merging the two neighboring subnetworks where joint modeling of the genes leads to the highest improvement in the objective function value. Joint modeling of dependent genes reveals coordinated responses and improves the likelihood of the data in comparison with independent models, giving the first criterion for merging the subnetworks. However, increasing subnetwork size tends to increase model complexity and the possibility of overfitting, since the number of samples remains constant while the dimensionality (subnetwork size) increases. To compensate for this effect, the Bayesian information criterion (see Gelman et al., 2003) is used to penalize increasing model complexity and to determine optimal subnetwork size. The final cost function for a subnetwork G is $C(G) = -2\mathcal{L} + qlog(N)$, where $\mathcal{L}$ is the (marginal) log-likelihood of the data, given the mixture model in Equation 5.1, q is the number of parameters and N denotes sample size. The algorithm then compares independent and joint models for each subnetwork pair that has a direct link in the network, and merges at each step the subnetwork pair G_i, G_j that minimizes the cost

$$\Delta\mathcal{C} = -2(\mathcal{L}_{i,j} - (\mathcal{L}_i + \mathcal{L}_j)) + (q_{i,j} - (q_i + q_j))log(N). \quad (5.2)$$

The iteration continues until no improvement is obtained by merging the subnetworks. The combination of modeling techniques yields a scalable algorithm for genome- and organism-wide investigations: First, the analysis focuses on those parts of the data that are supported by known interactions, which increases modeling power and considerably limits the search space. Second, the agglomerative scheme finds a fast approximative solution where at each step the subnetwork pair that leads to the highest improvement in cost function is merged. Third, an efficient variational approximation is used to learn the mixture models (Kurihara et al., 2007b). Note that the algorithm does not necessarily identify a globally optimal solution. However, detection of physiologically coherent and reproducible responses is often sufficient for practical applications.

Global view on network activation patterns

The NetResponse algorithm introduced in Publication 3 was applied to investigate transcriptional activation patterns of a pathway interaction network of 1800 genes based on the KEGG database of metabolic pathways (Kanehisa et al., 2008) provided by the SPIA package (Tarca et al., 2009) across 353 gene expression samples from 65 tissues. The two algorithms proposed in Publication 3, MATISSE+ and NetResponse were shown to outperform an unsupervised biclustering approach in terms of reproducibility of the finding. The introduced NetReponse algorithm, where the detection of transcriptional response patterns is used as a search criterion for subnetwork identification, was the best-performing method. The algorithm identified 106 subnetworks with 3-20 genes, with distinct transcriptional responses across the conditions. One of the subnetworks is illustrated in Figure 5.1; the other findings are provided in the supplementary material of Publication 3. The detected transcriptional responses were physiologically coherent, suggesting a potential functional role. The reproducibility of the responses was confirmed in an independent validation data set, where 80% of the predicted responses were detected ($p < 0.05$). The findings highlight context-specific regulation of the genes.

Some responses are shared by many conditions, while others are more specific to particular contexts such as the immune system, muscles, or the brain; related physiological conditions often exhibit similar network activation patterns. Tissue relatedness can be measured in terms of shared transcriptional responses of the subnetworks, giving an alternative formulation of the tissue connectome map suggested by Greco et al. (2008) in order to highlight functional connectivity between tissues based on the number of shared differentially expressed genes. In Publication 3, shared network responses are used instead of shared gene count. The use of co-regulated gene groups is expected to be more robust to noise than the use of individual genes. The analysis provides a global view on network activation across the normal human body, and can be used to formulate novel hypotheses of gene function in previously unexplored contexts.

5.3 Conclusion

Gene function and interactions are often subject to condition-specific regulation (Liang et al., 2006; Rachlin et al., 2006), but these have been typically studied only in particular experimental conditions. Organism-wide analysis can potentially reveal new functional connections and help to formulate novel hypotheses of gene function in previously unexplored contexts, and to detect highly specialized functions that are specific to few conditions. Changes in cell-biological conditions induce changes in the expression levels of co-regulated genes, in order to produce specific physiological responses, typically affecting only a small part of the network. Since individual processes and their transcriptional responses are in general unknown (Lee et al., 2008; Montaner et al., 2009), data-driven detection of condition-specific responses can provide an efficient proxy for identifying distinct transcriptional states of the network, with potentially distinct functional roles.

Publication 3 provides efficient model-based tools for global, organism-wide discovery and characterization of context-specific transcriptional activity in genome-scale interaction networks, independently of predefined classifications for genes and conditions. The network is used to bring in prior information of gene function, which would be missing in unsupervised models, and allows data-driven detection of coordinately regulated gene sets and their context-specific responses. The algorithm is readily applicable in any organism where gene expression and pairwise interaction data, including pathways, protein interactions and regulatory networks, are available. It has therefore a considerably larger scope than previous network-based models for global transcriptome analysis, which rely on organism-specific annotations, but lack implementations for most organisms (Reiss et al., 2006; Tanay et al., 2004).

While biomedical implications of the findings require further investigation, the results highlight shared and reproducible responses between physiological conditions, and provide a global view of transcriptional activation patterns across the normal human body. Other potential applications for the method include large-scale screening of drug responses and disease subtype discovery. Implementation of the algorithm is freely available through BioConductor.[1]

[1] http://bioconductor.org/packages/devel/bioc/html/netresponse.html

Chapter 6

Human transcriptome and other layers of genomic information

The way to deal with the problem of big data is to beat it senseless with other big data.

J. Quackenbush (2006)

This chapter presents the third main contribution of the thesis, computational strategies to integrate measurements of human transcriptome to other layers of genomic information. Genomic, transcriptomic, proteomic, epigenomic and other sources of measurement data characterize different aspects of genome organization (Hawkins et al., 2010; Montaner and Dopazo, 2010; Sara et al., 2010); any single source provides only a limited view to the cellular system. Understanding functional organization of the genome and ultimately the cell function requires integration of data from the various levels of genome organization and modeling of their dynamical interplay. Such an holistic approach, which is also called *systems biology*, is a key to understanding living organisms, which are "rich in emergent properties because forever new groups of properties emerge at every level of integration" (Mayr, 2004). Combining evidence across multiple sources can help to discover functional mechanisms and interactions, which are not seen in the individual data sets, and to increase statistical power in noisy and incomplete high-throughput experiments (Huttenhower and Hofmann, 2010; Reed et al., 2006).

Integration of heterogeneous genomic data comes with a variety of technical and methodological challenges (Hwang et al., 2005; Troyanskaya, 2005), and the particular modeling approaches vary according to the analysis task and particular properties of the investigated measurement sources. Integrative studies have been limited by poor availability of co-occurring genomic observations, but suitable data sets are now becoming increasingly available in both in-house and public biomedical data repositories (The Cancer Genome Atlas Research Network, 2008). New observations highlight the need for novel, integrative approaches in functional genomics (Coe et al., 2008). Recent studies have proposed for instance methods to integrate epigenetic modifications (Sadikovic et al., 2008), micro-RNA (Qin, 2008),

transcription factor binding (Savage et al., 2010), as well as protein expression (Johnson et al., 2008). Given the complex stochastic nature of biological systems, computational efficiency, robustness against uncertainty and interpretability of the results are key issues. Prior information of biological systems is often incomplete, and subject to high levels of uncontrolled variation and complex interdependencies between different parts of the cellular system (Troyanskaya, 2005). These issues emphasize the need for principled approaches requiring minimal prior knowledge about the data, as well as minimal model fitting procedures. Section 6.1 gives an overview of the standard models for high-throughput data integration methods, which have close connections to the modeling approaches developed in this work.

6.1 Standard approaches for genomic data integration

The integrative approaches can be roughly classified in three categories: methods that (i) combine statistical evidence across related studies in order to obtain more accurate inferences of target variables, (ii) utilize side information in order to guide the analysis of a single, primary data source, and (iii) detect and characterize dependencies between the measurement sources in order to discover new functional connections between the different layers of genomic information. The contributions in Chapters 4 and 5 are associated with the first two categories; the contributions presented in this chapter, the regularized dependency detection framework of Publication 4, and associative clustering of Publications 5 and 6, belong to the third category.

6.1.1 Combining statistical evidence

The first general category of methods for genomic data integration consists of approaches where evidence across similar studies is combined to increase statistical power, for instance by comparing and integrating data from independent microarray experiments targeted at studying the same disease. In Publications 2 and 3, joint analysis of a large number of commensurable microarray experiments, where the observed data is directly comparable between the arrays, helps to increase statistical power and to reveal weak, shared signals in the data that can not be detected in more restricted experimental setups and smaller datasets.

However, the related observations are often not directly comparable, and further methodological tools are needed for integration. *Meta-analysis* provides tools for such analysis (Ramasamy et al., 2008). Meta-analysis forms part of the microarray analysis procedure introduced in Publication 1, where methods to integrate related microarray measurements across different array platforms are developed. Meta-analysis emphasizes shared effects between the studies over statistical significance in individual experiments. In its standard form, meta-analysis assumes that each individual study measures the same target variable with varying levels of noise. The analysis starts from identifying a measure of *effect size* based on differences, means, or other summary statistics of the observations such as the Hedges' g, used in Publication 1. Weighted averaging of the effect sizes provides the final, combined result. Weighting accounts for differences in reliability of the individual studies, for instance by emphasizing studies with large sample size, or low measurement variance. Averaging is expected to yield more accurate

estimates of the target variable than individual studies. This can be particularly useful when several studies with small sample sizes are available for instance from different laboratories, which is a common setting in microarray analysis context, where the data sets produced by individual laboratories are routinely deposited to shared community databases. Ultimately, the quality of meta-analysis results rests on the quality of the individual studies. Modeling choices, such as the choice of the effect size measure and included studies will affect the analysis outcome.

Kernel methods (see e.g. Schölkopf and Smola, 2002) provide another widely used approach for integrating statistical evidence across multiple, potentially heterogeneous measurement sources. Kernel methods operate on similarity matrices, and provide a natural framework for combining statistical evidence to detect similarity and patterns that are supported by multiple observations. The modeling framework also allows for efficient modeling of nonlinear feature spaces.

Multi-task learning refers to a class of approaches where multiple, related modeling tasks are solved simultaneously by combining statistical power across the related tasks. A typical task is to improve the accuracy of individual classifiers by taking advantage of the potential dependencies between them (see e.g. Caruana, 1997).

6.1.2 Role of side information

The second category of approaches for genomic data integration consists of methods that are asymmetric by nature; integration is used to support the analysis of one, primary data source. Side information can be used, for instance, to limit the search space and to focus the analysis to avoid overfitting, speed up computation, as well as to obtain potentially more sensitive and accurate findings (see e.g. Eisenstein, 2006). One strategy is to impose hard constraints on the model, or model family, based on side information to target specific research questions. In gene expression context, functional classifications or known interactions between the genes can be used to constrain the analysis (Goeman and Buhlmann, 2007; Ulitsky and Shamir, 2009). In factor analysis and mixed effect models, clinical annotations of the samples help to focus the modeling on particular conditions (see e.g. Carvalho et al., 2008). Hard constraints rely heavily on the accuracy of side information. Soft, or probabilistic approaches can take the uncertainty in side information into account, but they are computationally more demanding. Examples of such methods in the context of transcriptome analysis include for instance the supervised biclustering models, such as cMonkey and modified SAMBA, as well as other methods that guide the analysis with additional information of genes and regulatory mechanisms, such as transcription factor binding (Reiss et al., 2006; Savage et al., 2010; Tanay et al., 2004). Publication 3 uses gene interaction network as a hard constraint for modeling transcriptional co-regulation of the genes, but the condition-specific responses of the detected gene groups are identified in an unsupervised manner.

A complementary approach for utilizing side information of the experiments is provided by *multi-way learning*. A classical example is the analysis of variance (ANOVA), where a single data set is modeled by decomposing it into a set of basic, underlying effects, which characterize the data optimally. The effects are associated with multiple, potentially overlapping attributes of the measurement samples, such as disease state, gender and age, which are known prior to the analysis. Taking such prior knowledge of systematic variation between the samples into account

helps to increase modeling power and can reveal the attribute-specific effects. An interesting subtask is to model the interactions between the attributes, so-called *interaction effects*. These are manifested only with particular combinations of attributes, and indicate dependency between the attributes. For instance, simultaneous cigarette smoking and asbestos exposure will considerably increase the risk of lung cancer, compared to any of the two risk factors alone (see e.g. Nymark et al., 2007). *Factor analysis* is a closely related approach where the attributes, also called *factors*, are not given but instead estimated from the data. *Mixed effect models* combine the supervised and unsupervised approaches by incorporating both *fixed* and *random effects* in the model, corresponding to the known and latent attributes, respectively (see e.g. Carvalho et al., 2008). The standard factorization approaches for individual data sets are related to the dependency-seeking approaches in Publications 4-6, where co-occurring data sources are decomposed in an unsupervised manner into components that are maximally informative of the components in the other data set.

6.1.3 Modeling of mutual dependency

Symmetric models for dependency detection form the third main category of methods for genomic data integration, as well as the main topic of this chapter. Dependency modeling is used to distinguish the *shared* signal from *dataset-specific* variation. The shared effects are informative of the commonalities and interactions between the observations, and are often the main focus of interest in integrative analysis. This motivates the development of methods that can allocate computational resources efficiently to modeling of the shared features and interactions.

Multi-view learning is a general category of approaches for symmetric dependency modeling tasks. In multi-view learning, multiple measurement sources are available, and each source is considered as a different view on the same objects. The task is to enhance modeling performance by combining the complementary views. A classical example of such a model is canonical correlation analysis (Hotelling, 1936). Related approaches that have recently been applied in functional genomics include for instance probabilistic variants of meta-analysis (Choi et al., 2007; Conlon et al., 2007), generalized singular value decomposition (see e.g. Alter et al., 2003; Berger et al., 2006) and simultaneous non-negative matrix factorization (Badea, 2008).

The dependency modeling approaches in this thesis make an explicit distinction between statistical representation of data and the modeling task. Let us denote the representations of two co-occurring multivariate observations, $\mathbf{x}$ and $\mathbf{y}$, with $f_x(\mathbf{x})$ and $f_y(\mathbf{y})$, respectively. The selected representations depend on the application task. The representation can be for instance used to perform feature selection as in *canonical correlation analysis (CCA)* Hotelling (1936), capture nonlinear features in the data as in kernelized versions of CCA (see e.g. Yamanishi et al., 2003), or partition the data as in information bottleneck (Friedman et al., 2001) and associative clustering (Publications 5-6). *Statistical independence* of the representations implies that their joint probability density can be decomposed as $p(f_x(\mathbf{x}), f_y(\mathbf{y})) = p(f_x(\mathbf{x}))p(f_y(\mathbf{y}))$. Deviations from this assumption indicate statistical dependency. The representations can have a flexible parametric form which can be optimized by the dependency modeling algorithms to identify dependency structure in the data.

Recent examples of such dependency-maximizing methods include probabilistic

canonical correlation analysis (Bach and Jordan, 2005), which has close theoretical connection to the regularized models introduced in Publication 4, and the associative clustering principle introduced in Publications 5-6. Canonical correlations and contingency table analysis form the methodological background for the contributions in Publications 4-6. In the remainder of this section these two standard approaches for dependency detection are considered more closely.

Classical and probabilistic canonical correlation analysis

Canonical correlation analysis (CCA) is a classical method for detecting linear dependencies between two multivariate random variables (Hotelling, 1936). While ordinary correlation characterizes the association strength between two vectors with paired scalar observations, CCA assumes paired vectorial values, and generalizes correlation to multidimensional sources by searching for maximally correlating low-dimensional representation of the two sources, defined by linear projections $\mathbf{X}\mathbf{v}_x, \mathbf{Y}\mathbf{v}_y$. Multiple projection components can be obtained iteratively, by finding the most correlating projection first, and then consecutively the next ones after removing the dependencies explained by the previous CCA components; the lower-dimensional representations are defined by projections to linear hyperplanes. The model can be formulated as a generalized eigenvalue problem that has an analytical solution with two useful properties: the result is invariant to linear transformations of the data, and the solution for any fixed number of components maximizes mutual information between the projections for Gaussian data (Kullback, 1959; Bach and Jordan, 2002). Extensions of the classical CCA include generalizations to multiple data sources (Kettenring, 1971; Bach and Jordan, 2002), regularized solutions with non-negative and sparse projections (Sigg et al., 2007; Archambeau and Bach, 2008; Witten et al., 2009), and non-linear extensions, for instance with kernel methods (Bach and Jordan, 2002; Yamanishi et al., 2003). Direct optimization of correlations in the classical CCA provides an efficient way to detect dependencies between data sources, but it lacks an explicit model to deal with the uncertainty in the data and model parameters.

Recently, the classical CCA was shown to correspond to the ML solution of a particular generative model where the two data sets are assumed to stem from a shared Gaussian latent variable $\mathbf{z}$ and normally distributed data-set-specific noise (Bach and Jordan, 2005). Using linear assumptions, the model is formally defined as

$$\begin{cases} \mathbf{x} & \sim \mathbf{W}_x\mathbf{z} + \boldsymbol{\varepsilon}_x \\ \mathbf{y} & \sim \mathbf{W}_y\mathbf{z} + \boldsymbol{\varepsilon}_y. \end{cases} \tag{6.1}$$

The manifestation of the shared signal in each data set can be different. This is parameterized by $\mathbf{W}_x$ and $\mathbf{W}_y$. Assuming a standard Gaussian model for the shared latent variable, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and data set-specific effects where $\boldsymbol{\varepsilon}_x \sim \mathcal{N}(\mathbf{0}, \Psi_x)$ (and respectively for $\mathbf{y}$), the correlation-maximizing projections of the traditional CCA introduced in Section 6.1 can be retrieved from the ML solution of the model (Archambeau et al., 2006; Bach and Jordan, 2005). The model decomposes the observed co-occurring data sets into *shared* and *data set-specific* components based on explicit modeling assumptions (Figure 6.1). The dataset-specific effects can also be described in terms of latent variables as $\boldsymbol{\varepsilon}_x = \mathbf{B}_x\mathbf{z}_x$ and $\boldsymbol{\varepsilon}_y = \mathbf{B}_y\mathbf{z}_y$, allowing the construction of more detailed models for the dataset-specific effects (Klami and

Kaski, 2008). The shared signal $\mathbf{z}$ is treated as a latent variable and marginalized out in the model, providing the marginal likelihood for the observations:

$$p(\mathbf{X}, \mathbf{Y}|\mathbf{W}, \Psi) = \int p(\mathbf{X}, \mathbf{Y}|\mathbf{Z}, \mathbf{W}, \Psi)p(\mathbf{Z})d\mathbf{Z}, \tag{6.2}$$

where Ψ denotes the block-diagonal matrix of Ψ_x, Ψ_y, and $\mathbf{W} = [\mathbf{W}_x; \mathbf{W}_y]$. The probabilistic formulation of CCA has opened up a way to new probabilistic extensions that can treat the modeling assumptions and uncertainties in the data in a more explicit and robust manner (Archambeau et al., 2006; Klami and Kaski, 2008; Klami et al., 2010).

The general formulation provides a flexible modeling framework, where different modeling assumptions can be used to adapt the models in different applications. The connection to classical CCA assumes full covariances for the dataset-specific effects. Simpler models for the dataset-specific effects will not distinguish between the shared and marginal effects as effectively, but they have fewer model parameters that can potentially reduce overlearning and speed up computation. It is also possible to tune the dimensionality of the shared latent signal. Learning of lower-dimensional models can be faster and potentially less prone to overfitting. Interpretation of simpler models is also more straightforward in many applications. The probabilistic formulation allows rigorous treatment of uncertainties in the data and model parameters also with small sample sizes that are common in biomedical studies, and allows the incorporation of prior information through Bayesian priors, as in the regularized dependency detection framework introduced in Publication 4.

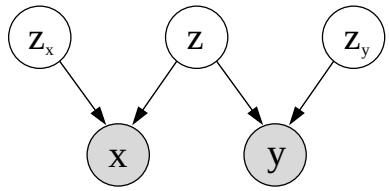


Figure 6.1: A graphical representation of the generative shared latent variable model in Equation (6.1). The latent source $\mathbf{z}$ is shared by observations $\mathbf{x}$ and $\mathbf{y}$. The other effects that are specific to each observation are characterized by $\mathbf{z}_x$ and $\mathbf{z}_y$, respectively. Gray shading indicates observed variables.

Contingency table analysis

Contingency table analysis is a classical approach used to study associations between co-occurring categorical observations. The co-occurrences are represented by cross-tabulating them on a *contingency table*, the rows and columns of which correspond to the first and second set of features, respectively. Various tests are available for measuring dependency between the rows and columns of the table Yates (1934); Agresti (1992), including the classical Fisher test (Fisher, 1934), a standard tool for measuring statistical enrichment of functional categories in gene cluster analysis (Hosack et al., 2003). While the classical contingency table analysis is used to measure dependency between co-occurring variables, more recent approaches use contingency tables to derive objective functions for dependency exploration tasks. The associative clustering principle introduced in Publications 5-6 is an example of such approach.

Other approaches that use contingency table dependencies as objective functions include the *information bottleneck (IB)* principle (Tishby et al., 1999) and *discriminative clustering (DC)* (Sinkkonen et al., 2002; Kaski et al., 2005). These are asymmetric, dependency-seeking approaches that can be used to discover cluster structure in a primary data such that it is maximally informative of another, discrete auxiliary variable. The dependency is represented on a contingency table, and maximization of contingency table dependencies provides the objective function for clustering. While the standard IB operates on discrete data, DC is used to discover cluster structure in continuous-valued data. The two approaches also employ different objective functions. In classical IB, a discrete variable $\mathcal{X}$ is clustered in such a way that the cluster assignments become maximally informative of another discrete variable $\mathcal{Y}$. The complexity of the cluster assignments is controlled by minimizing the mutual information between the cluster indices and the original variables. The task is to find a partitioning $\tilde{\mathbf{X}}$ that minimizes the cost $\mathcal{L}(p(\tilde{\mathbf{X}}|\mathbf{X})) = I(\tilde{\mathbf{X}};\mathbf{X}) - \beta I(\tilde{\mathbf{X}};\mathbf{Y})$, where β controls clustering resolution. In DC, mutual information is replaced by a Bayes factor between the two hypotheses of dependent and independent margins. The Bayes factor is asymptotically consistent with mutual information, but provides an unbiased estimate for limited sample size (see e.g. Sinkkonen et al., 2005). The standard information bottleneck and discriminative clustering are asymmetric methods that treat one of the data sources as the primary target of analysis.

In contrast, the dependency maximization approaches considered in this thesis, the associative clustering (AC) and regularized versions of canonical correlation analysis are symmetric and they operate exclusively on continuous-valued data. CCA is not based on contingency table analysis, but it has close connections to the Gaussian IB (Chechik et al., 2005) that seeks maximal dependency between two sets of normally distributed variables. The Gaussian IB retrieves the same subspace as CCA for one of the data sets. However, in contrast to the symmetric CCA model, Gaussian IB is a directed method that finds dependency-maximizing projections for only one of the two data sets. The second dependency detection approach considered in this thesis, the associative clustering, is particularly related to the symmetric IB that finds two sets of clusters, one for each variable, which are optimally compressed presentations of the original data, and at the same time maximally informative of each other (Friedman et al., 2001). While the objective function in IB is derived from mutual information, AC uses the Bayes factor as an objective function in a similar manner as it is used in the asymmetric discriminative clustering. Another key difference is that while the symmetric IB operates on discrete data, AC employs contingency table analysis in order to discover cluster structure in continuous-valued data spaces.

6.2 Regularized dependency detection

Standard unsupervised methods for dependency detection, such as the canonical correlation analysis or the symmetric information bottleneck, seek maximal dependency between two data sets with minimal assumptions about the dependencies. The unconstrained models involve high degrees of freedom when applied to high-dimensional genomic observations. Such flexibility can easily lead to overfitting, which is even worse for more flexible nonparametric or nonlinear, kernel-based dependency discovery methods. Several ways to regularize the solution have been

suggested to overcome associated problems, for instance by imposing sparsity constraints on the solution space (Bie and Moor, 2003; Vinod, 1976).

In many applications prior information of the dependencies is available, or particular types of dependency are relevant for the analysis task. Such prior information can be used to reduce the degrees of freedom in the model, and to regularize dependency detection. In the cancer gene discovery application of Publication 4, DNA mutations are systematically correlated with transcriptional activity of the genes within the affected region, and identification of such regions is a biomedically relevant research task. Prior knowledge of chromosomal distances between the observations can improve the detection of the relevant spatial dependencies. However, principled approaches to incorporate such prior information in dependency modeling have been missing. Publication 4 introduces regularized models for dependency detection based on classical canonical correlation analysis (Hotelling, 1936) and its probabilistic formulation (Bach and Jordan, 2005). The models are extended by incorporating appropriate prior terms, which are then used to reduce the degrees of freedom based on prior biological knowledge.

Correlation-based variant

In order to introduce the regularized dependency detection framework of Publication 4, let us start by considering regularization of the classical correlation-based CCA. This searches for arbitrary linear projection vectors $\mathbf{v}_x, \mathbf{v}_y$ that maximize the correlation between the projections of the data sets $\mathbf{X}, \mathbf{Y}$. Multiple projection components can be obtained iteratively, by finding the most correlating projection first, and then consecutively the next ones after removing the dependencies explained by the previous CCA components. The procedure will identify maximally dependent linear subspaces of the investigated data sets. To regularize the solution, Publication 4 couples the projections through a transformation matrix $\mathbf{T}$ in such a way that $\mathbf{v}_y = \mathbf{T}\mathbf{v}_x$. With a completely unconstrained $\mathbf{T}$ the model reduces to the classical unconstrained CCA; suitable constraints on can be used to regularize dependency detection.

To enforce regularization one could for instance prefer solutions for $\mathbf{T}$ that are close to a given transformation matrix, $\mathbf{T} \sim \mathbf{M}$, or impose more general constraints on the structure of the transformation matrix that would prefer particular rotational or other linear relationships. Suitable constraints depend on the particular applications; the solutions can be made to prefer particular types of dependency in a soft manner by appropriate penalty terms. In Publication 4 the completely unconstrained CCA model has been compared with a fully regularized model with $\mathbf{T} = \mathbf{I}$; this encodes the biological assumption that probes with small chromosomal distances tend to capture more similar signal between gene expression and copy number measurements than probes with a larger chromosomal distance; the projection vectors characterize this relationship, and are therefore expected to have similar form, $\mathbf{v}_x \sim \mathbf{v}_y$. Utilization of other, more general constraints in related data integration tasks provides a promising topic for future studies.

The correlation-based treatment provides an intuitive and easily implementable formulation for regularized dependency detection. However, it lacks an explicit model for the shared and data-specific effects, and it is likely that some of the dataset-specific effects are captured by the correlation-maximizing projections. This is suboptimal for characterizing the shared effects, and motivates the probabilistic treatment.

Probabilistic dependency detection with similarity constraints

The probabilistic approach for regularized dependency detection in Publication 4 is based on an explicit model of the data-generating process formulated in Equation (6.1). In this model, the transformation matrices $\mathbf{W}_x$, $\mathbf{W}_y$ specify how the shared latent variable $\mathbf{Z}$ is manifested in each data set $\mathbf{X}$, $\mathbf{Y}$, respectively. In the standard model, the relationship between the transformation matrices is not constrained, and the algorithm searches for arbitrary linear transformations that maximize the likelihood of the observations in Equation (6.2). The probabilistic formulation opens up possibilities to guide dependency search through Bayesian priors.

In Publication 4, the standard probabilistic CCA model is extended by incorporating additional prior terms that regularize the relationship by reparameterizing the transformation matrices as $\mathbf{W}_y = \mathbf{T}\mathbf{W}_x$, and setting a prior on $\mathbf{T}$. The treatment is analogous to the correlation-based variant, but now the transformation matrices operate on the latent components, rather than the observations. This allows to distinguish the shared and dataset-specific effects more explicitly in the model. The task is then to learn the optimal parameter matrix $\mathbf{W} = [\mathbf{W}_x; \mathbf{W}_y]$, given the constraint $\mathbf{W}_y = \mathbf{T}\mathbf{W}_x$. The Bayes' rule gives the model likelihood

$$p(\mathbf{X}, \mathbf{Y}, \mathbf{W}, \mathbf{\Psi}) \sim p(\mathbf{X}, \mathbf{Y}|\mathbf{W}, \mathbf{\Psi})p(\mathbf{W}, \mathbf{\Psi}). \tag{6.3}$$

The likelihood term $p(\mathbf{X}, \mathbf{Y}|\mathbf{W}, \mathbf{\Psi})$ can be calculated based on the model in Equation (6.1). This defines the objective function for standard probabilistic CCA, which implicitly assumes a flat prior $p(\mathbf{W}, \mathbf{\Psi}) \sim 1$ for the model parameters. The formulation in Equation (6.3) makes the choice of the prior explicit, allowing modifications on the prior term. To obtain a tractable prior, let us assume that the prior factorizes as $p(\mathbf{W}, \mathbf{\Psi}) = p(\mathbf{W})p(\mathbf{\Psi})$. The first term can be further decomposed as $p(\mathbf{W}) \sim p(\mathbf{W}_x)p(\mathbf{T})$, assuming independent priors for $\mathbf{W}_x$ and $\mathbf{T}$. A convenient and tractable prior for $\mathbf{T}$ is provided by the matrix normal distribution:[1]

$$p(\mathbf{T}) = \mathcal{N}_m(\mathbf{T}|\mathbf{M}, \mathbf{U}, \mathbf{V}). \tag{6.4}$$

For computational simplicity, let us assume independent rows and columns with $\mathbf{U} = \mathbf{V} = \sigma_T\mathbf{I}$. The mean matrix $\mathbf{M}$ can be used to emphasize certain types of dependency between $\mathbf{W}_x$ and $\mathbf{W}_y$. Assuming uninformative, flat priors $p(\mathbf{W}_x) \sim 1$ and $p(\mathbf{\Psi}) \sim 1$, as in the standard probabilistic CCA model, and denoting $\mathbf{\Sigma} = \mathbf{W}\mathbf{W}^T + \mathbf{\Psi}$, the negative log-likelihood of the model is

$$-logp(\mathbf{X}, \mathbf{Y}, \mathbf{W}, \mathbf{\Psi}) \sim log|\mathbf{\Sigma}| + Tr\mathbf{\Sigma}^{-1}\tilde{\mathbf{\Sigma}} + \frac{\| \mathbf{T} - \mathbf{M} \|_F^2}{2\sigma_T^2}. \tag{6.5}$$

This is the objective function to minimize. Note that this has the same form as the objective function of the standard probabilistic CCA, except the additional penalty term $\frac{\|\mathbf{T}-\mathbf{M}\|_F^2}{2\sigma_T^2}$ arising from the prior $p(\mathbf{T})$. This yields the cost function employed in Publication 4. In our cancer gene discovery application the choice $\mathbf{M} = \mathbf{I}$ is used to encode the biological prior constrain $\mathbf{T} \approx \mathbf{I}$, which states that the observations with a small chromosomal distance should on average show similar responses in the integrated data sets, i.e., $\mathbf{W}_x \approx \mathbf{W}_y$. The regularization strength can be tuned

[1] $\mathcal{N}_m(\mathbf{T}|\mathbf{M}, \mathbf{U}, \mathbf{V}) \sim exp\left(-\frac{1}{2}Tr\{\mathbf{U}^{-1}(\mathbf{T} - \mathbf{M})\mathbf{V}^{-1}(\mathbf{T} - \mathbf{M})^T\}\right)$ where $\mathbf{M}$ is the mean matrix, and $\mathbf{U}$ and $\mathbf{V}$ denote row and column covariances, respectively.

with σ_T^2. A fully regularized model is obtained with $\sigma_T^2 \to 0$. When $\sigma_T^2 \to \infty$, $\mathbf{W}_x$ and $\mathbf{W}_y$ become independent *a priori*, yielding the ordinary probabilistic CCA. The σ_T^2 can be used to regularize the solution between these two extremes. Note that it is possible to incorporate also other types of prior information concerning the dependencies into the model through $p(\mathbf{T})$.

The model parameters $\mathbf{W}$, $\mathbf{\Psi}$ are estimated with the EM algorithm. The regularized version is not analytically tractable with respect to $\mathbf{W}$ in the general case, but can be optimized with standard gradient-based optimization techniques. Special cases of the model have analytical solutions, which can speed up the model fitting procedure. In particular, the fully regularized and unconstrained models, obtained with $\sigma_T^2 = 0$ and $\sigma_T^2 = \infty$ respectively, have closed-form solutions for $\mathbf{W}$. Note that the current formulation assumes that the regularization parameters $\mathbf{M}, \sigma_T^2$ are defined prior to the analysis. Alternatively, these parameters could be optimized based on external criteria, such as cancer gene detection performance in our application, or learned from the data in a fully Bayesian treatment these parameters would be treated as latent variables. Incorporation of additional prior information of the data set-specific effects through priors on $\mathbf{W}_x$ and $\mathbf{\Psi}$ provides promising lines for further work.

6.2.1 Cancer gene discovery with dependency detection

The regularized models provide a principled framework for studying associations between transcriptional activity and other regulatory layers of the genome. In Publication 4, the models are used to investigate cancer mechanisms. DNA copy number changes are a key mechanism for cancer, and integration of copy number information with mRNA expression measurements can reveal functional effects of the mutations. While causation may be difficult to grasp, study of the dependencies can help to identify functionally active mutations, and to provide candidate biomarkers with potential diagnostic, prognostic and clinical impact in cancer studies.

The modeling task in the cancer gene discovery application of Publication 4 is to identify chromosomal regions that show exceptionally high levels of dependency between gene copy number and transcriptional levels. The model is used to detect dependency within local chromosomal regions that are then compared in order to identify the exceptional regions. The dependency is quantified within a given region by comparing the strength of shared and data set-specific signal. High scores indicate regions where the shared signal is particularly high relative to the data-set-specific effects. A sliding-window approach is used to screen the genome for dependencies. The regions are defined by the d closest probes around each gene. Then the dimensionality of the models stays constant, which allows direct comparison of the dependency measures between the regions without additional adjustment terms that would be otherwise needed to compensate for differences in model complexity.

Prior information of the dependencies is used to regularize cancer gene detection. Chromosomal gains and losses are likely to be positively correlated with the expression levels of the affected genes within the same chromosomal region or its close proximity; copy number gain is likely to increase the expression of the associated genes whereas deletion will block gene expression. The prior information is encoded in the model by setting $\mathbf{M} = \mathbf{I}$ in the prior term $p(\mathbf{T})$. This accounts for the expected positive correlations between gene expression and copy number

within the investigated chromosomal region. Regularization based on such prior information is shown to improve cancer gene detection performance in Publication 4, where the regularized variants outperformed the unconstrained models.

A genome-wide screen of 51 gastric cancer patients (Myllykangas et al., 2008) reveals clear associations between DNA copy number changes and transcriptional activity. The Figure 6.2 illustrates dependency detection on chromosome arm 17q, where the regularized model reveals high dependency between the two data sources in a known cancer-associated region. The regularized and unconstrained models were compared in terms of receiver-operator characteristics calculated by comparing the ordered gene list from the dependency screen to an expert-curated list of known genes associated with gastric cancer (Myllykangas et al., 2008). A large proportion of the most significant findings in the whole-genome analysis were known cancer genes; the remaining findings with no known associations to gastric cancer are promising candidates for further study.

Biomedical interpretation of the model parameters is also straightforward. A ML estimate of the latent variable values $\mathbf{Z}$ characterizes the strength of the shared signal between DNA mutations and transcriptional activity for each patient. This allows robust identification of small, potentially unknown patient subgroups with shared amplification effects. These would remain potentially undetected when comparing patient groups defined based on existing clinical annotations. The parameters in $\mathbf{W}$ can downweigh signal from poorly performing probes in each data set, or probes that measure genes whose transcriptional levels are not functionally affected by the copy number change. This provides tools to distinguish between so-called *driver* mutations having functional effects from less active *passenger* mutations, which is an important task in cancer studies. On the other hand, the model can combine statistical power across the adjacent measurement probes, and it captures the strongest shared signal in the two sets of observations. This is useful since gene expression and copy number data are typically characterized by high levels of biological and measurement variation and small sample size.

Related approaches

Integration of chromosomal aberrations and transcriptional activity is an actively studied data integration task in functional genomics. The first studies with standard statistical tests were carried out by Hyman et al. (2002) and Phillips et al. (2001) when simultaneous genome-wide observations of the two data sources had become available. The modeling approaches utilized in this context can be roughly classified in regression-based, correlation-based and latent variable approaches. The regression-based models (Adler et al., 2006; Bicciato et al., 2009; van Wieringen and van de Wiel, 2009) characterize alterations in gene expression levels based on copy number observations with multivariate regression or closely related models. The correlation-based approaches (González et al., 2009; Schäfer et al., 2009; Soneson et al., 2010) provide symmetric models for dependency detection, based on correlation and related statistical models. Many of these methods also regularize the solutions, typically based on sparsity constraints and non-negativity of the projections (Lê Cao et al., 2009; Waaijenborg et al., 2008; Witten et al., 2009; Parkhomenko et al., 2009). The correlation-based approach in Publication 4 introduces a complementary approach for regularization that constrains the relationship between subspaces where the correlations are estimated. The latent variable models by Berger et al. (2006); Shen et al. (2009); Vaske et al. (2010),

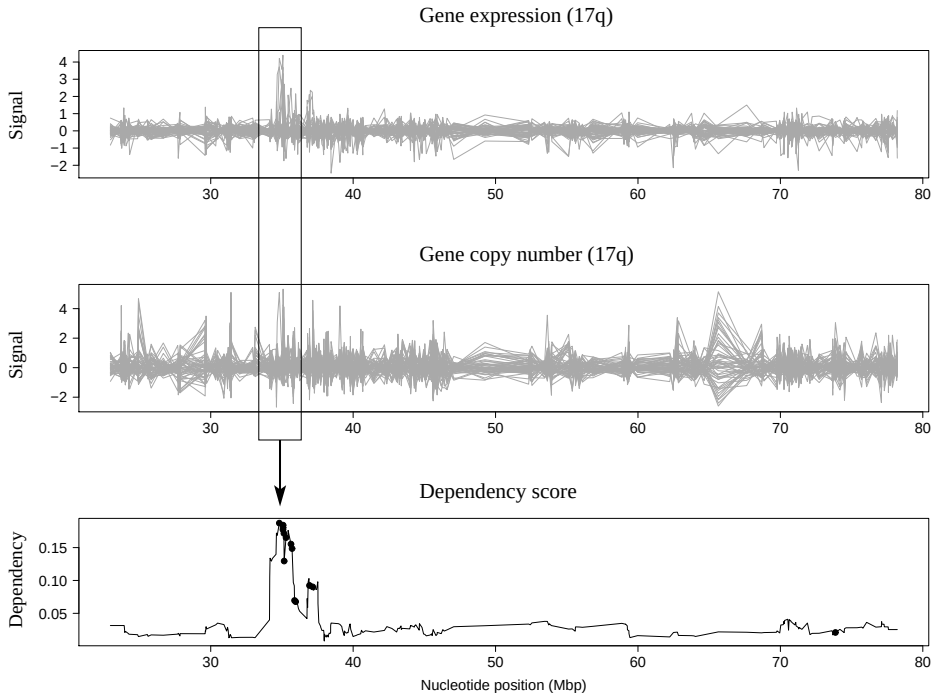


Figure 6.2: Gene expression, copy number signal, and the dependency score along the chromosome arm 17q obtained with the regularized latent variable framework in Equation 6.5. Known cancer-associated genes from an expert-curated list are marked with black dots.

and Publication 4 are based on explicit modeling assumptions concerning the data-generating processes. The iCluster algorithm (Shen et al., 2009) is closely related to the latent variable model considered in Publication 4. While our model detects continuous dependencies, iCluster uses a discrete latent variable to partition the samples into distinct subgroups. The iCluster model is regularized by sparsity constraints on $\mathbf{W}$, while we tune the relationship between $\mathbf{W}_x$ and $\mathbf{W}_y$. Moreover, the model in Publication 4 utilizes full covariance matrices to model for the dataset-specific effects, whereas iCluster uses diagonal covariances. The more detailed model for dataset-specific effects in our model should help to distinguish the shared signal more accurately. Other latent variable approaches include the iterative method based on generalized singular-value decomposition (Berger et al., 2006), and the probabilistic factor graph model PARADIGM (Vaske et al., 2010), which additionally utilizes pathway topology information in the modeling.

Experimental comparison between the related integrative approaches can be problematic since they target related, but different research questions where the biological ground truth is often unknown. For instance, some methods utilize patient class information in order to detect class-specific alterations (Schäfer et al., 2009), other methods perform *de novo* class discovery (Shen et al., 2009), provide tools for gene prioritization (Salari et al., 2010), or guide the analysis with additional functional information of the genes (Vaske et al., 2010). The algorithms introduced in Publication 4 are particularly useful for gene prioritization and class discovery purposes, where the target is to identify the most promising cancer gene

candidates for further validation, or to detect potentially novel cancer subtypes. However, while an increasing number of methods are released as conveniently accessible algorithmic tools (Salari et al., 2010; Shen et al., 2009; Schäfer et al., 2009; Witten et al., 2009), implementations of most models are not available for comparison purposes. Open source implementations of the dependency detection algorithms developed in this thesis have been released to enhance transparency and reproducibility of the computational experiments and to encourage further use of these models (Huovilainen and Lahti, 2010).

6.3 Associative clustering

Functions of human genes are often studied indirectly, by studying model organisms such as the mouse (Davis, 2004; Joyce and Palsson, 2006). Orthologs are genes in different species that originate from a single gene in the last common ancestor of these species. Such genes have often retained identical biological roles in the present-day organisms, and are likely to share the function (Fitch, 1970). Mutations in the genomic DNA sequence are a key mechanism in evolution. Consequently, DNA sequence similarity can provide hypotheses of gene function in poorly annotated species. An exceptional level of conservation may highlight critical physiological similarities between species, whereas divergence can indicate significant evolutionary changes (Jordan et al., 2005). Investigating evolutionary conservation and divergence will potentially lead to a deeper understanding of what makes each species unique. Evolutionary changes primarily target the structure and sequence of genomic DNA. However, not all changes will lead to phenotypic differences. On the other hand, sequence similarity is not a guarantee of functional similarity because small changes in DNA can potentially have remarkable functional implications.

Therefore, in addition to investigating *structural conservation* of the genes at the sequence level, another level of investigation is needed to study *functional conservation* of the genes and their regulation, which is reflected at the transcriptome (Jiménez et al., 2002; Jordan et al., 2005). Transcriptional regulation of the genes is a key regulatory mechanism that can have remarkable phenotypic consequences in highly modular cell-biological systems (Hartwell et al., 1999) even when the original function of the regulated genes would remain intact.

Systematic comparison of transcriptional activity between different species would provide a straightforward strategy for investigating conservation of gene regulation (Bergmann et al., 2004; Enard et al., 2002; Zhou and Gibson, 2004). However, direct comparison of individual genes between species may not be optimal for discovering subtle and complex dependency structures. The associative clustering principle (AC), introduced in Publications 5-6, provides a framework for detecting groups of orthologous genes with exceptional levels of conservation and divergence in transcriptional activity between two species. While standard dependency detection methods for continuous data, such as the generalized singular value decomposition (see e.g. Alter et al., 2003) or canonical correlation analysis (Hotelling, 1936) detect global linear dependencies between observations, AC searches for dependent, local groupings to reveal gene groups with exceptional levels of conservation and divergence in transcriptional activity. The model is free of particular distributional assumptions about the data, which helps to allocate modeling resources to detecting dependent subgroups when variation within each

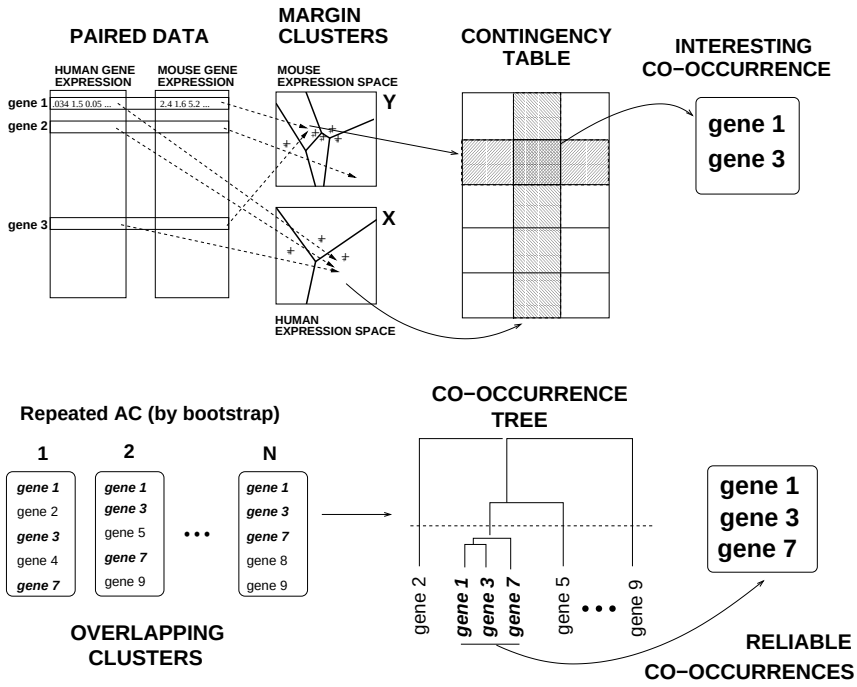


Figure 6.3: Principle of associative clustering (AC). AC performs simultaneous clustering of two data sets, consisting of paired observations, and seeks to maximize the dependency between the two sets of clusters. The clusters are defined by cluster centroids in each data space. The clustering results are represented on a contingency table, where clusters of the two data sets correspond with the rows and columns of the contingency table, respectively. These are called the margin clusters of the contingency table. The table cells are called cross clusters and they contain orthologous genes from the two data sets. The cluster centroids are optimized to produce a contingency table with maximal dependency between the margin cluster counts. Cross clusters that show significant deviation from the null hypothesis of independent margins indicate dependency. In order to enhance the reliability of the results, the clustering is repeated with slightly differing bootstrap samples. Then reliable co-occurrences are identified from a co-occurrence tree with a specified threshold. Frequently co-occurring orthologues are selected for further analyzes.

group is less relevant for the analysis. The remainder of this section provides an overview of the associative clustering principle and its application to studying evolutionary divergence between species.

The associative clustering principle

The principle of associative clustering (AC) is illustrated in Figure 6.3. AC performs simultaneous clustering of two data sets to reveal maximally dependent cluster structure between two sets of observations. The clusters are defined in each data space by *Voronoi parameterization*, where the clusters are defined by cluster centroids to produce connected, internally homogeneous clusters. Let us denote the two sets of clusters by $\{V_i^{(x)}\}_i$, $\{V_j^{(y)}\}_j$. A given data point $\mathbf{x}$ is then assigned to the cluster corresponding to the nearest centroid $\mathbf{m}_i$ in the feature space, with respect to a given distance measure[2] d. This divides the space into non-overlapping *Voronoi regions*. The regions define a clustering for all points of

[2] $\mathbf{x} \in V_i^{(x)}$ if $d(\mathbf{x}, \mathbf{m}_i) \leq d(\mathbf{x}, \mathbf{m}_k)$ for all k.

the data space. The association between the clusters of the two data sets can be represented on a contingency table, where the rows and columns correspond to clusters in the first and second data set, respectively. The clusters in each data set are called *margin clusters*. Each pair of co-occurring observations $(\mathbf{x}_i, \mathbf{y}_i)$ maps to one margin cluster in each data set, and each contingency table cell corresponds to a pair of margin clusters. These are called *cross clusters*.

AC searches for a maximally dependent cluster structure by optimizing the Voronoi centroids in the two data spaces in such a way that the dependency between the contingency table margins is maximized. Let us denote the number of samples in cross cluster i, j by n_{ij}. The corresponding margin cluster counts are $n_{i\cdot} = \sum_j n_{ij}$ and $n_{\cdot j} = \sum_i n_{ij}$. The observed sample frequencies over the contingency table margins and cross-clusters are assumed to follow multinomial distribution with latent parameters $\boldsymbol{\theta}_i, \boldsymbol{\theta}_j$ and $\boldsymbol{\theta}_{ij}$, respectively. Assuming the model M_I of *independent margin clusters*, the expected sample frequency in each cross cluster is given by the outer product of margin cluster frequencies. The model M_d of *dependent margin clusters* deviates from this assumption. The *Bayes factor (BF)* is used to compare the two hypotheses of dependent and independent margins. This is a rigorously justified approach for model comparison, which indicates whether the observations provide superior evidence for either model. Evidence is calculated over all potential values of the model parameters, marginalized over the latent frequencies. In a standard setting, the Bayes factor would be used to compare evidence between the dependent and independent margin cluster models for a given clustering solution. AC uses the Bayes factor in a non-standard manner; as an objective function to maximize by optimizing the cluster centroids in each data space; the centroids define the margin clusters and consequently the margin cluster dependencies.

The centroids are optimized with a conjugate-gradient algorithm after smoothing the cluster borders with continuous parameterization. The hyperparameters $n^{(d)}$, $n^{(x)}$, and $n^{(y)}$ arise from Dirichlet priors of the two multinomial models M_I, M_D of independent and dependent margins, respectively. Setting the hyperparameters to unity yields the classical hypergeometric measure of contingency table dependency (Fisher, 1934; Yates, 1934). With large sample size, the logarithmic Bayes factor approaches mutual information (Sinkkonen et al., 2005). The Bayes factor is a desirable choice especially with a limited sample size since a marginalization over the latent variables makes it robust against uncertainty in the parameter values, and because finite contingency table counts would give a biased estimate of mutual information. The number of clusters in each data space is specified in advance, typically based on the desired level of resolution. Nonparametric extensions, where the number of margin clusters would be inferred automatically from the data form one potential topic for further studies; a closely related approach was recently proposed in Rogers et al. (2010).

Publication 6 introduces an additional, bootstrap-based procedure to assess the reliability of the findings (Figure 6.3). The analysis is repeated with similar, but not identical training data sets obtained by sampling the original data with replacement. The most frequently detected dependencies are then investigated more closely. The analysis will emphasize findings that are not sensitive to small variations in the observed data.

Comparison methods

Associative clustering was compared with two alternative methods: standard K-means on each of the two data sets, and a combination of K-means and information bottleneck (K-IB). K-means (see e.g. Bishop, 2006) is a classical clustering algorithm that provides homogeneous, connected clusters based on Voronoi parameterization. Homogeneity is desirable for interpretation, since the data points within a given cluster can then be conveniently summarized by the cluster centroid. On the other hand, K-means considers each data set independently, which is suboptimal for the dependency modeling task. The two sets of clusters obtained by K-means, one for each data space, can then be presented on a contingency table as in associative clustering. The second comparison method is K-IB introduced in Publication 5. K-IB uses K-means to partition the two co-occurring, continuous-valued data sets into discrete atomic regions where each data point is assigned in its own singleton cluster. This gives two sets of atomic clusters that are mapped on a large contingency table, filled with frequencies of co-occurring data pairs $(\mathbf{x}_k, \mathbf{y}_k)$. The table is then compressed to the desired size by aggregating the margin clusters with the symmetric IB algorithm in order to maximize the dependency between the contingency table margins (Friedman et al., 2001). Aggregating the atomic clusters provides a flexible clustering approach, but the resulting clusters are not necessarily homogeneous and they are therefore difficult to interpret.

AC compared favorably to the other methods. While AC outperformed the standard K-means in dependency modeling, the cluster homogeneity was not significantly reduced in AC. The cross clusters from K-IB (Sinkkonen et al., 2003) were more dependent than in AC. On the other hand, AC produced more easily interpretable localized clusters, as measured by the sum of intra-cluster variances in Publication 6. Homogeneity makes it possible to summarize clusters conveniently, for instance by using the mean expression profiles of the cluster samples, as in Figure 6.4B. While K-means searches for maximally homogeneous clusters and K-IB searches for maximally dependent clusters, AC finds a successful compromise between the goals of dependency and homogeneity.

6.3.1 Exploratory analysis of transcriptional divergence between species

Associative clustering is used in Publications 5 and 6 to investigate conservation and divergence of transcriptional activity of 2818 orthologous human-mouse gene pairs across an organism-wide collection of transcriptional profiling data covering 46 and 45 tissue types in human and mouse, respectively (Su et al., 2002). AC takes as input two gene expression matrices with orthologous genes, one for each species, and returns a dependency-maximizing clustering for the orthologous gene pairs. Interpretation of the results focuses on unexpectedly large or small cross clusters revealed by the contingency table analysis of associative clustering. Compared to plain correlation-based comparisons between the gene expression profiles, AC can reveal additional cluster structure, where genes with similar expression profiles are clustered together, and associations between the two species are investigated at the level of such detected gene groups. The dependency between each pair of margin clusters can be characterized by comparing the respective margin cluster centroids that provide a compact summary of the samples within each cluster.

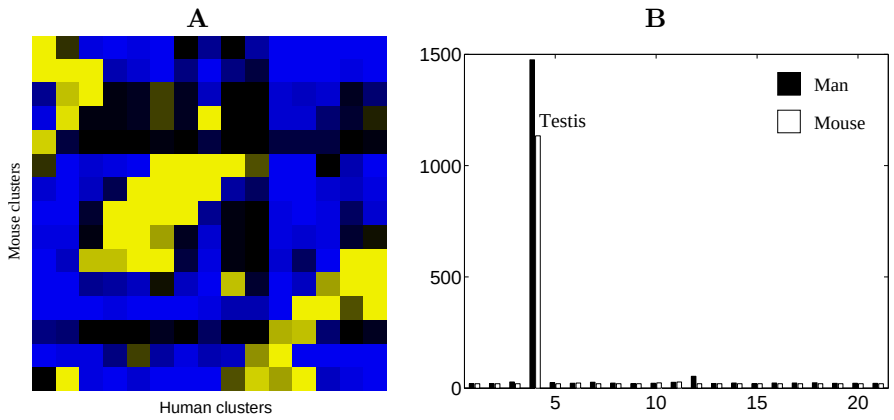


Figure 6.4: **A** The contingency table of associative clustering highlights orthologous gene groups in human (rows) and mouse (columns) with exceptional levels of conservation (yellow) or divergence (blue) in transcriptional activity between the two species. **B** Average expression profiles of a highly conserved group of testis-specific genes across 21 tissues in man and mouse. ©IEEE. Reprinted with permission from Publication 6.

Biological interpretation of the findings, based on enrichment of Gene Ontology (GO) categories (Ashburner et al., 2000), revealed genes with strongly conserved and potentially diverged transcriptional activity. The most highly enriched categories were associated with ribosomal functions, the high conservation of which has also been suggested in earlier studies (Jiménez et al., 2002); ribosomal genes often require coordinated effort of a large group of genes, and they function in cell maintenance tasks that are critical for species survival. An exceptional level of conservation was also observed in a group of testis-specific genes, yielding novel functional hypotheses for certain poorly annotated genes within the same cross-cluster (Figure 6.4). Transcriptional divergence, on the other hand, was detected for instance in genes related to embryonic development.

While general-purpose dependency exploration tools may not be optimal for studying the specific issue of transcriptional conservation, such tools can reveal dependency with minimal prior knowledge about the data. This is useful in functional genomics experiments where little prior knowledge is available. In Publications 5 and 6, associative clustering has been additionally applied in investigating dependencies between transcriptional activity and transcription factor binding, another key regulatory mechanism of the genes.

6.4 Conclusion

The models introduced in Publications 4-6 provide general exploratory tools for the discovery and analysis of statistical dependencies between co-occurring data sources and tools to guide modeling through Bayesian priors. In particular, the models consider linear dependencies (Publication 4) and cluster-based dependency structures (Publications 5-6) between the data sources. The models are readily applicable to data integration tasks in functional genomics. In particular, the models have been applied to investigate dependencies between chromosomal mutations and transcriptional activity in cancer, and evolutionary divergence of transcript-

ional activity between human and mouse. Biomedical studies provide a number of other potential applications for such general-purpose methods. An increasing number of co-occurring observations across the various regulatory layers of the genome are available concerning epigenetic mechanisms, micro-RNAs, polymorphisms and other genomic features (The Cancer Genome Atlas Research Network, 2008). Simultaneous observations provide a valuable resource for investigating the functional properties that emerge from the interactions between the different layers of genomic information. An open source implementation in BioConductor[3] provides accessible computational tools for related data integration tasks, helping to guarantee the utility of the developed models for the computational biology community.

[3] http://www.bioconductor.org/packages/release/bioc/html/pint.html

Chapter 7

Summary and conclusions

Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better.

J.E. Cohen (2004)

Following the initial sequencing of the human genome (International human genome sequencing consortium, 2001; Venter et al., 2001), the understanding of structural and functional organization of genetic information has extended rapidly with the accumulation of research data. This has opened up new challenges and opportunities for making fundamental discoveries about living organisms and creating a holistic picture about genome organization. The increasing need to organize the large volumes of genomic data with minimal human intervention has made computation an increasingly central element in modern scientific inquiry. It is a paradox of our time that the historical scale of data in public and proprietary repositories is only revealing how incomplete our knowledge of the enormous complexity of living systems is. The particular challenges in data-intensive genomics are associated with the complex and poorly characterized nature of living systems, as well as with limited availability of observations. It is possible to solve some of these challenges by combining statistical power across multiple experiments, and utilizing the wealth of background information in public repositories. Exploratory data analysis can help to provide research hypotheses and material for more detailed investigations based on large-scale genomic observations when little prior knowledge is available concerning the underlying phenomena; models that are robust to uncertainty and able to automatically adapt to the data, can facilitate the discovery of novel biological hypotheses. Statistical learning and probabilistic models provide a natural theoretical framework for such analysis.

In this thesis, general-purpose exploratory data analysis methods have been developed for organism-wide analysis of the human transcriptome, a central functional layer of the genome. Integrating evidence across multiple sources of genomic information can help to reveal mechanisms that could not be investigated based on smaller and more targeted experiments; this is a central aspect in all contributions. In particular, methods have been developed (i) in order to improve measurement accuracy of high-throughput observations, (ii) in order to model transcriptional activation patterns and tissue relatedness in genome-wide interaction networks at an organism-wide scale, and (iii) in order to integrate measurements of the human transcriptome with other layers of genomic information. These results contribute

to some of the 'grand challenges' in the genomic era by developing strategies to understand cell-biological systems, genetic contributions to human health and evolutionary variation (Collins et al., 2003). The computational experiments in this thesis have been carried out based on publicly available, anonymized data sets that follow commonly accepted ethical standards in biomedical research. Open access implementations of the key algorithms have been provided to guarantee wide access to these tools and to spark new research beyond the original applications presented in this thesis.

Methodological extensions and application of the developed algorithms to new data integration tasks in functional genomics and in other fields provide a promising line for future studies. The methods developed in this thesis are readily applicable in genome-wide screening studies in cancer and potentially other diseases. Increasing amounts of co-occurring data concerning various aspects of the genome have become available, including gene- and micro-RNA expression, structural variation in the DNA, epigenetic modifications and gene regulatory networks. It is expected that with small modifications the introduced methodology can be applied to study further associations between these and other layers of genome organization, as well as their contributions to human health. The fundamental research challenges in contemporary genome biology provide a wide array of applications for statistical learning and exploratory analysis, and a rich source of ideas for methodological research.

Bibliography

A. S. Adler, M. Lin, H. Horlings, D. S. A. Nuyten, M. J. van de Vijver, and H. Y. Chang. Genetic regulators of large-scale transcriptional signatures in cancer. *Nature Genetics*, 38:421–430, 2006.

A. Agresti. A survey of exact inference for contingency tables. *Statistical Science*, 7: 131–153, 1992.

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

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

D. J. Allocco, I. S. Kohane, and A. J. Butte. Quantifying the relationship between co-expression, co-regulation and gene function. *BMC Bioinformatics*, 5:18, 2004.

O. Alter and G. H. Golub. Reconstructing the pathways of a cellular system from genome-scale signals by using matrix and tensor computations. *Proceedings of the National Academy of Sciences, USA*, 102:17559–17564, 2005.

O. Alter, P. O. Brown, and D. Botstein. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proceedings of the National Academy of Sciences, USA*, 100:3351–3356, 2003.

C. Archambeau and F. Bach. Sparse probabilistic projections. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 73–80. MIT Press, Cambridge, MA, 2008.

C. Archambeau, N. Delannay, and M. Verleysen. Robust probabilistic projections. In W. Cohen and A. Moore, editors, *Proceedings of the 23rd International conference on machine learning*, volume 148, pages 33–40. ACM, Pittsburgh, Pennsylvania, 2006.

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. *Nature Genetics*, 25:25–29, 2000.

H. Auer, S. Lyianarachchi, D. Newsom, M. I. Klisovic, G. Marcucci, and K. Kornacker. Chipping away at the chip bias: RNA degradation in microarray analysis. *Nature Genetics*, 35:292–293, 2003.

F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

F. R. Bach and M. I. Jordan. A probabilistic interpretation of canonical correlation analysis. Technical report, Department of Statistics, University of California, Berkeley, 2005.

L. Badea. Extracting gene expression profiles common to colon and pancreatic adenocarcinoma using simultaneous nonnegative matrix factorization. In R. B. Altman, A. K. Dunker, L. Hunter, T. Murray, and T. E. Klein, editors, *Proceedings of the Pacific Symposium on Biocomputing (PSB'08)*, pages 267–278. World Scientific, USA, 2008.

P. Baldi and S. Brunak. *Bioinformatics: the machine learning approach*. Bradford, London, third edition, 1999.

T. Bammler et al. Standardizing global gene expression analysis between laboratories and across platforms. *Nature Methods*, 2:351–356, 2005.

N. Bannert and R. Kurth. Retroelements and the human genome: New perspectives on an old relation. *Proceedings of the National Academy of Sciences*, 101(S2):14572–14579, 2004.

A.-L. Barabási and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews*, 5:101–113, 2004.

Y. Barash and N. Friedman. Context-specific bayesian clustering for gene expression data. *Journal of Computational Biology*, 9:169–191, 2002.

V. Barbour, B. Cohen, and G. Yamey. Why bigger is not yet better: The problems with huge datasets. *PLoS Medicine*, 2:e55, 2005.

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 Research*, 37:D885–90, 2009.

T. Bayes. Studies in the history of probability and statistics: IX. Thomas Bayes' essay Towards solving a problem in the doctrine of chances. *Biometrika*, 45:296–315, 1763. Printed in 1958.

D. G. Beer, S. L. R. Kardia, C.-C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. Chen, T. G. Gharib, D. G. Thomas, M. L. Lizyness, R. Kuick, S. Hayasaka, J. M. G. Taylor, M. D. Iannettoni, M. B. Orringer, and S. Hanash. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine*, 8: 816–824, 2002.

S. Ben-David and M. Ackerman. Measures of clustering quality: A working set of axioms for clustering. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 121–128. MIT Press, Cambridge, MA, 2008.

D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. GenBank. *Nucleic Acids Research*, 38:D46–51, 2010.

J. A. Berger, S. Hautaniemi, S. K. Mitra, and J. Astola. Jointly analyzing gene expression and copy number data in breast cancer using data reduction models. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3:2–16, 2006.

S. Bergmann, J. Ihmels, and N. Barkai. Similarities and differences in genome-wide expression data of six organisms. *PLoS Biology*, 2:85–93, 2004.

J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. John Wiley & Sons Ltd, Chichester, England, 2000.

R. Beroukhim, C. H. Mermel, D. Porter, G. Wei, S. Raychaudhuri, J. Donovan, J. Barretina, J. S. Boehm, J. Dobson, M. Urashima, K. T. Mc Henry, R. M. Pinchback, A. H. Ligon, Y.-J. Cho, L. Haery, H. Greulich, M. Reich, W. Winckler, M. S. Lawrence, B. A. Weir, K. E. Tanaka, D. Y. Chiang, A. J. Bass, A. Loo, C. Hoffman, J. Prensner, T. Liefeld, Q. Gao, D. Yecies, S. Signoretti, E. Maher, F. J. Kaye, H. Sasaki, J. E. Tepper, J. A. Fletcher, J. Tabernero, J. Baselga, M.-S. Tsao, F. Demichelis, M. A. Rubin, P. A. Janne, M. J. Daly, C. Nucera, R. L. Levine, B. L. Ebert, S. Gabriel, A. K. Rustgi, C. R. Antonescu, M. Ladanyi, A. Letai, L. A. Garraway, M. Loda, D. G. Beer, L. D. True, A. Okamoto, S. L. Pomeroy, S. Singer, T. R. Golub, E. S. Lander, G. Getz, W. R. Sellers, and M. Meyerson. The landscape of somatic copy-number alteration across human cancers. *Nature*, 463:899–905, 2010.

A. Bhattacharjee, W. G. Richards, J. Staunton, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences, USA*, 98:13790–13795, 2001.

S. Bicciato, R. Spinelli, M. Zampieri, E. Mangano, F. Ferrari, L. Beltrame, I. Cifola, C. Peano, A. Solari, and C. Battaglia. A computational procedure to identify significant overlap of differentially expressed and genomic imbalanced regions in cancer datasets. *Nucleic Acids Research*, 37:5057–5070, 2009.

T. D. Bie and B. D. Moor. On the regularization of canonical correlation analysis. In S.-I. Amari, A. Cichocki, S. Makino, and N. Murata, editors, *Proceedings of the International Conference on Independent Component Analysis and Blind Source Separation (ICA2003)*. Nara, Japan, April 1–4 2003.

BioPAX workgroup. *BioPAX - Biological Pathways Exchange Language*, 2005. Level 2, Version 1.0 Documentation.

C. M. Bishop. *Pattern recognition and machine learning*. Springer, Singapore, 2006.

J. Blake. Bio-ontologies – fast and furious. *Nature Biotechnology*, 22:773–774, 2004.

B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19:185–193, 2003.

A.-L. Boulesteix. Over-optimism in bioinformatics research. *Bioinformatics*, 26:437, 2010.

T. Boveri. *Zur Frage der Entstehung maligner Tumoren*. Verlag von Gustav Fischer, Jena, 1914.

J. R. Bradford, Y. Hey, T. Yates, Y. Li, S. D. Pepper, and C. J. Miller. A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. *BMC Genomics*, 11:282, 2010.

U. M. Braga-Neto and E. T. A. Marques. From functional genomics to functional immunomics: New challenges, old problems, big rewards. *PLoS Computational Biology*, 2:e81, 2006.

A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. P. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nature Genetics*, 29:365–371, 2001.

A. Brazma, M. Krestyaninova, and U. Sarkans. Standards for systems biology. *Nature Reviews Genetics*, 7:593–605, 2006.

M. R. Brent. Steady progress and recent breakthroughs in the accuracy of automated genome annotation. *Nature Reviews Genetics*, 9:62–73, 2008.

T. A. Brown. *Genomes*. Garland Science, UK, third edition, 2006.

C. G. Broyden. The convergence of a class of double-rank minimization algorithms, II: The new algorithm. *IMA Journal of Applied Mathematics*, 6:222–231, 1970.

A. Butte. The use and analysis of microarray data. *Nature Reviews*, 1:951–960, 2002.

L. Cabusora, E. Sutton, A. Fulmer, and C. V. Forst. Differential network expression during drug and stress response. *Bioinformatics*, 21:2898–2905, 2005.

G. A. Calin and C. M. Croce. MicroRNA signatures in human cancers. *Nature Reviews Cancer*, 6:857–866, 2006.

V. J. Carey and V. Stodden. Reproducible Research Concepts and Tools for Cancer Bioinformatics. In M. F. Ochs, J. T. Casagrande, and R. V. Davuluri, editors, *Biomedical Informatics for Cancer Research*, pages 149–175. Springer US, Boston, MA, 2010.

P. Carninci. Is sequencing enlightenment ending the dark age of the transcriptome? *Nature Methods*, 6:711–713, 2009.

S. B. Carroll. Genetics and the making of homo sapiens. *Nature*, 422:849–857, 2003.

R. Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997.

C. M. Carvalho, J. Chang, J. E. Lucas, J. R. Nevins, Q. Wang, and M. West. High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association*, 103:1438–1456, 2008.

J. T. Chang, C. Carvalho, S. Mori, A. H. Bild, M. L. Gatza, Q. Wang, J. E. Lucas, A. Potti, P. G. Febbo, M. West, and J. R. Nevins. A genomic strategy to elucidate modules of oncogenic pathway signaling networks. *Molecular Cell*, 34:104–114, 2009.

R. Chari, B. P. Coe, E. A. Vucic, W. W. Lockwood, and W. L. Lam. An integrative multidimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer. *BMC Systems Biology*, 4:67, 2010.

G. Chechik, A. Globerson, N. Tishby, and Y. Weiss. Information Bottleneck for Gaussian variables. *Journal of Machine Learning Research*, 6:165–188, 2005.

R. J. Cho, I. S. Dhillon, Y. Guan, and S. Sra. Minimum sum-squared residue co-clustering of gene expression data. In M. W. Berry, U. Dayal, C. Kamath, and D. Skillicorn, editors, *Proceedings of the 4th SIAM International Conference on Data Mining*, pages 114–125. Florida, USA, 2004.

H. Choi, R. Shen, A. M. Chinnaiyan, and D. Ghosh. A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments. *BMC Bioinformatics*, 8:364, 2007.

J. K. Choi, U. Yu, S. Kim, and O. J. Yoo. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, 19:i84–90, 2003.

G. M. Church. The personal genome project. *Molecular Systems Biology*, 1:30, 2005.

G. R. Cochrane and M. Y. Galperin. The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources. *Nucleic Acids Research*, 38:D1–4, 2010.

B. P. Coe, R. Chari, W. W. Lockwood, and W. L. Lam. Evolving strategies for global gene expression analysis of cancer. *Journal of Cellular Physiology*, 217:590–597, 2008.

J. E. Cohen. Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better. *PLoS Biology*, 2:e439, 2004.

F. S. Collins, E. D. Green, A. E. Guttmacher, and M. S. Guyer. A vision for the future of genomics research. *Nature*, 422:835–847, 2003.

E. Conlon, J. Song, and A. Liu. Bayesian meta-analysis models for microarray data: a comparative study. *BMC Bioinformatics*, 8:80, 2007.

T. F. Consortium. The Transcriptional Landscape of the Mammalian Genome. *Science*, 309:1559–1563, 2005.

R. T. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, 17:1–13, 1946.

F. Crick. Central dogma of molecular biology. *Nature*, 227:561–563, 1970.

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 Research*, 33:e175, 2005.

C. Darwin. *On the Origin of Species by Means of Natural Selection*. Murray, London, 1859.

R. H. Davis. The age of model organisms. *Nature Reviews Genetics*, 5:69–76, 2004.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

J. L. DeRisi, V. R. Iyer, and P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–686, 1997.

J. Downward. Cancer biology: Signatures guide drug choice. *Nature*, 439:274–275, 2006.

S. Draghici, P. Khatri, A. L. Tarca, K. Amin, A. Done, C. Voichita, C. Georgescu, and R. Romero. A systems biology approach for pathway level analysis. *Genome Research*, 17:1537–1545, 2007.

J. T. Dudley, R. Tibshirani, T. Deshpande, and A. J. Butte. Disease signatures are robust across tissues and experiments. *Molecular Systems Biology*, 5:307, 2009.

B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, USA, 1994.

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

M. Eisenstein. More than just 'doing the math'. *Nature Methods*, 3:420–420, 2006.

L. L. Elo, L. Lahti, H. Skottman, M. Kyläniemi, R. Lahesmaa, and T. Aittokallio. Integrating probe-level expression changes across generations of Affymetrix arrays. *Nucleic Acids Research*, 33:e193, 2005.

W. Enard, P. Khaitovich, J. Klose, S. Zöllner, F. Heissig, K. Giavalisco, P. Nieselt-Struwe, E. Muchmore, A. Varki, R. Ravid, G. M. Doxiadis, R. E. Bontrop, and S. Pääbo. Intra- and inter-specific variation of primate gene expression patterns. *Science*, 296:340–343, 2002.

C. Espinosa-Soto and A. Wagner. Specialization Can Drive the Evolution of Modularity. *PLoS Computational Biology*, 6:e1000719, 2010.

D. Evanko. Hacking the genome. *Nature Methods*, 3:495–495, 2006.

D. Evanko. Supplement on visualizing biological data. *Nature Methods*, 7(S1), 2010.

L. Feuk, A. R. Carson, and S. W. Scherer. Structural variation in the human genome. *Nature Reviews Genetics*, 7:85–97, 2006.

R. A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, fifth edition, 1934.

W. M. Fitch. Distinguishing homologous from analogous proteins. *Systematic Zoology*, 19:99–113, 1970.

R. Fletcher. A new approach to variable metric algorithms. *The Computer Journal*, 13: 317–322, 1970.

P. Flicek, B. L. Aken, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, J. Fernandez-Banet, L. Gordon, S. Gräf, S. Haider, M. Hammond, K. Howe, A. Jenkinson, N. Johnson, A. Kähäri, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, G. Koscielny, E. Kulesha, D. Lawson, I. Longden, T. Massingham, W. McLaren, K. Megy, B. Overduin, B. Pritchard, D. Rios, M. Ruffier, M. Schuster, G. Slater, D. Smedley, G. Spudich, Y. A. Tang, S. Trevanion, A. Vilella, J. Vogel, S. White, S. P. Wilder, A. Zadissa, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. M. Fernández-Suarez, J. Herrero, T. J. P. Hubbard, A. Parker, G. Proctor, J. Smith, and S. M. J. Searle. Ensembl's 10th year. *Nucleic Acids Research*, 38:D557–562, 2010.

J. A. Foekens, Y. Wang, J. W. Martens, E. M. Berns, and J. G. Klijn. The use of genomic tools for the molecular understanding of breast cancer and to guide personalized medicine. *Drug Discovery Today*, 13:481–487, 2008.

N. Friedman and D. Koller. Being Bayesian about network structure: A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, 50:95–126, 2003.

N. Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303:799–805, 2004.

N. Friedman, O. Mosenzon, N. Slonim, and N. Tishby. Multivariate information bottleneck. In J. S. Breese and D. Koller, editors, *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 152–161. Morgan Kaufmann Publishers, San Francisco, CA, 2001.

C. Furusawa and K. Kaneko. Zipf's law in gene expression. *Physical Review Letters*, 90:088102, 2003.

G10KCOS consortium. Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *The Journal of Heredity*, 100:659–674, 2009.

G. Gad, E. Levine, and E. Domany. Coupled two-way clustering analysis of gene microarray data. *Proceedings of the National Academy of Sciences, USA*, 97:12079–12084, 2000.

J. Gagneur, H. Sinha, F. Perocchi, R. Bourgon, W. Huber, and L. M. Steinmetz. Genome-wide allele- and strand-specific expression profiling. *Molecular Systems Biology*, 5:274, 2009.

L. Gautier, M. Moller, L. Friis-Hansen, and S. Knudsen. Alternative mapping of probes to genes for Affymetrix chips. *BMC Bioinformatics*, 5:111, 2004.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL, USA, second edition, 2003.

G. K. Gerber, R. D. Dowell, T. S. Jaakkola, and D. K. Gifford. Automated discovery of functional generality of human gene expression programs. *PLoS Computational Biology*, 3:e148, 2007.

D. Gershon. DNA microarrays: More than gene expression. *Nature*, 437:1195–1198, 2005.

E. R. Gibney and C. M. Nolan. Epigenetics and gene expression. *Heredity*, 105:4–13, 2010.

J. J. Goeman and P. Buhlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23:980–987, 2007.

D. Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24:23–26, 1970.

I. González, S. Déjean, P. Martin, O. Gonçalves, P. Besse, and Baccini A. Highlighting relationships between heterogeneous biological data through graphical displays based on regularized canonical correlation analysis. *Journal of Biological Systems*, 17:173–199, 2009.

D. Greco, P. Somervuo, A. D. Lieto, T. Raitila, L. Nitsch, E. Castrén, and P. Auvinen. Physiology, pathology and relatedness of human tissues from gene expression meta-analysis. *PLoS One*, 3:e1880, 2008.

I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer. Co-clustering of biological networks and gene expression data. *Bioinformatics*, 18:S145–154, 2002.

L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402:C47–52, 1999.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, second edition, 2009.

R. D. Hawkins, G. C. Hon, and B. Ren. Next-generation genomics: an integrative approach. *Nature Reviews Genetics*, 11:476–486, 2010.

S. Heber and B. Sick. Quality assessment of Affymetrix GeneChip data. *OMICS: A Journal of Integrative Biology*, 10:358–368, 2006.

A.-M. K. Hein, S. Richardson, H. C. Causton, G. K. Ambler, and P. J. Green. BGX: a fully Bayesian integrated approach to the analysis of Affymetrix GeneChip data. *Biostatistics*, 6:349–373, 2005.

N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, editors. *Bayesian nonparametrics*. Cambridge University Press, USA, 2010.

S. Hochreiter, D.-A. Clevert, and K. Obermayer. A new summarization method for affymetrix probe level data. *Bioinformatics*, 22:943–949, 2006.

J. D. Hoheisel. Microarray technology: beyond transcript profiling and genotype analysis. *Nature Reviews Genetics*, 7:200–210, 2006.

D. Hosack, G. Dennis Jr., B. Sherman, H. Lane, and R. Lempicki. Identifying biological themes within lists of genes with EASE. *Genome Biology*, 4:R70, 2003.

H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.

Z. Hu, C. Fan, D. Oh, J. Marron, X. He, B. Qaqish, C. Livasy, L. Carey, E. Reynolds, L. Dressler, A. Nobel, J. Parker, M. Ewend, L. Sawyer, J. Wu, Y. Liu, R. Nanda, M. Tretiakova, A. Orrico, D. Dreher, J. Palazzo, L. Perreard, E. Nelson, M. Mone, H. Hansen, M. Mullins, J. Quackenbush, M. Ellis, O. Olopade, P. Bernard, and C. Perou. The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics*, 7:96, 2006.

E. Hubbell, W.-M. Liu, and R. Mei. Robust estimators for expression analysis. *Bioinformatics*, 18:1585–1592, 2002.

O.-P. Huovilainen and L. Lahti. pint: Pairwise integration of functional genomics data. Computer program. BioConductor, 2010.

M. E. Hurles, E. T. Dermitzakis, and C. Tyler-Smith. The functional impact of structural variation in humans. *Trends in Genetics*, 24:238–245, 2008.

C. Huttenhower and O. Hofmann. A quick guide to large-scale genomic data mining. *PLoS Computational Biology*, 6:e1000779, 2010.

C. Huttenhower, E. M. Haley, M. A. Hibbs, V. Dumeaux, D. R. Barrett, H. A. Coller, and O. G. Troyanskaya. Exploring the human genome with functional maps. *Genome Research*, 19:1093–1106, 2009.

D. Hwang, A. G. Rust, S. Ramsey, J. J. Smith, D. M. Leslie, A. D. Weston, P. de Atauri, J. D. Aitchison, L. Hood, A. F. Siegel, and H. Bolouri. A data integration methodology for systems biology. *Proceedings of the National Academy of Sciences, USA*, 102:17296–17301, 2005.

K.-B. Hwang, S. W. Kong, S. A. Greenberg, and P. J. Park. Combining gene expression data from different generations of oligonucleotide arrays. *BMC Bioinformatics*, 5:159, 2004.

E. Hyman, P. Kauraniemi, S. Hautaniemi, M. Wolf, S. Mousses, E. Rozenblum, M. Ringner, G. Sauter, O. Monni, A. Elkahloun, O.-P. Kallioniemi, and A. Kallioniemi. Impact of DNA Amplification on Gene Expression Patterns in Breast Cancer. *Cancer Research*, 62:6240–6245, 2002.

T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18:S233–240, 2002.

J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai. Revealing modular organization in the yeast transcriptional network. *Nature Genetics*, 31:370–377, 2002.

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431:931–945, 2004.

International human genome sequencing consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

J. P. A. Ioannidis, D. B. Allison, C. A. Ball, I. Coulibaly, X. Cui, A. C. Culhane, M. Falchi, C. Furlanello, L. Game, G. Jurman, J. Mangion, T. Mehta, M. Nitzberg, G. P. Page, E. Petretto, and V. van Noort. Repeatability of published microarray gene expression analyses. *Nature Genetics*, 41:149–155, 2009.

R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 31:e15, 2003a.

R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249–264, 2003b.

R. A. Irizarry, D. Warren, F. Spencer, I. F. Kim, S. Biswal, B. C. Frank, E. Gabrielson, J. G. N. Garcia, J. Geoghegan, G. Germino, C. Griffin, S. C. Hilmer, E. Hoffman, A. E. Jedlicka, E. Kawasaki, F. Martinez-Murillo, L. Morsberger, H. Lee, D. Petersen, J. Quackenbush, A. Scott, M. Wilson, Y. Yang, S. Q. Ye, and W. Yu. Multiple-laboratory comparison of microarray platforms. *Nature Methods*, 2:345–350, 2005.

R. A. Irizarry, Z. Wu, and H. A. Jaffee. Comparison of Affymetrix GeneChip expression measures. *Bioinformatics*, 22:789–794, 2006.

H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96:161–173, 2001.

A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, New Jersey, 1988.

J. L. Jiménez, M. P. Mitchell, and J. G. Sgouros. Microarray analysis of orthologous genes: conservation of the translational machinery across species at the sequence and expression level. *Genome Biology*, 4:R4, 2002.

J. M. Johnson, S. Edwards, D. Shoemaker, and E. E. Schadt. Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends in Genetics*, 21:93–102, 2005.

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

N. Johnson, V. Speirs, N. J. Curtin, and A. G. Hall. A comparative study of genome-wide SNP, CGH microarray and protein expression analysis to explore genotypic and phenotypic mechanisms of acquired antiestrogen resistance in breast cancer. *Breast Cancer Research and Treatment*, 111:55–63, 2008.

I. K. Jordan, L. Mariño-Ramirez, and E. V. Koonin. Evolutionary significance of gene expression divergence *Gene*, 345:119–126, 2005.

A. R. Joyce and B. O. Palsson. The model organism as a system: integrating 'omics' data sets. *Nature Reviews Molecular Cell Biology*, 7:198–210, 2006.

M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi. KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, 36:D480–484, 2008.

M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, 38:D355–360, 2010.

S. Kaski, J. Sinkkonen, and A. Klami. Discriminative clustering. *Neurocomputing*, 69: 18–41, 2005.

S. Katz, R. A. Irizarry, X. Lin, M. Tripputi, and M. W. Porter. A summarization approach for Affymetrix GeneChip data using a reference training set from a large, biologically diverse database. *BMC Bioinformatics*, 7:464, 2006.

J. Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58:433–451, 1971.

S. Kilpinen, R. Autio, K. Ojala, K. Iljin, E. Bucher, H. Sara, T. Pisto, M. Saarela, R. I. Skotheim, M. Bjorkman, J.-P. Mpindi, S. Haapa-Paananen, P. Vainio, H. Edgren, M. Wolf, J. Astola, M. Nees, S. Hautaniemi, and O. Kallioniemi. Systematic bioinformatic analysis of expression levels of 17,330 human genes across 9,783 samples from 175 types of healthy and pathological tissues. *Genome Biology*, 9:R139, 2008.

A. Klami and S. Kaski. Probabilistic approach to detecting dependencies between data sets. *Neurocomputing*, 72:39–46, 2008.

A. Klami, S. Virtanen, and S. Kaski. Bayesian exponential family projections for coupled data sources. In P. Grunwald and P. Spirtes, editors, *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 286–293. AUAI Press, Corvallis, Oregon, 2010.

J. Kleinberg. An impossibility theorem for clustering. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 446–453. MIT Press, Cambridge, MA, 2002.

T. Kohonen. *Self-Organizing Maps*. Springer, Berlin, third edition, 2001.

T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, USA, 2009.

S. Kullback. *Information Theory and Statistics*. Wiley, New York, 1959.

K. Kurihara, M. Welling, and Y. W. Teh. Collapsed variational dirichlet process mixture models. In *20th International Joint Conference on Artificial Intelligence (IJCAI 2007)*, pages 2796–2801. Morgan Kaufmann Publishers Inc, San Francisco, CA, USA, 2007a.

K. Kurihara, M. Welling, and N. Vlassis. Accelerated variational Dirichlet process mixtures. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 761–768. MIT Press, Cambridge, MA, 2007b.

E. Laajala, T. Aittokallio, R. Lahesmaa, and L. L. Elo. Probe-level estimation improves the detection of differential splicing in Affymetrix exon array studies. *Genome biology*, 10:R77, 2009.

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

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

E. Lander. The new genomics: Global views of biology. *Science*, 274:536–539, 1996.

D. A. Lauffenburger. Cell signaling pathways as control modules: Complexity for simplicity. *Proceedings of the National Academy of Sciences, USA*, 97:5031–5033, 2000.

M. Law, M. Figueiredo, and A. Jain. Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:1154–1166, 2004.

L. Lazzeroni and A. Owen. Plaid models for gene expression data. *Statistica Sinica*, 12: 61–86, 2002.

K.-A. Lê Cao, I. González, and S. Déjean. integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics*, 25:2855–2856, 2009.

H. Ledford. The cancer genome challenge. *Nature*, 464:972–974, 2010.

E. Lee, H.-Y. Chuang, J.-W. Kim, T. Ideker, and D. Lee. Inferring pathway activity toward precise disease classification. *PLoS Computational Biology*, 4:e1000217, 2008.

D. Levine, D. Haynor, J. Castle, S. Stepaniants, M. Pellegrini, M. Mao, and J. Johnson. Pathway and gene-set activation measurement from mRNA expression data: the tissue distribution of human pathways. *Genome Biology*, 7:R93, 2006.

T. R. Lezon, J. R. Banavar, M. Cieplak, A. Maritan, and N. V. Fedoroff. From the Cover: Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proceedings of the National Academy of Sciences, USA*, 103: 19033–19038, 2006.

C. Li and W. H. Wong. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Sciences, USA*, 98:31–36, 2001.

X. Li, Z. He, and J. Zhou. Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation. *Nucleic Acids Research*, 33:6114–6123, 2005.

Y. Li, P. Agarwal, and D. Rajagopalan. A global pathway crosstalk network. *Bioinformatics*, 24:1442–1447, 2008.

S. Liang, Y. Li, X. Be, S. Howes, and W. Liu. Detecting and profiling tissue-selective genes. *Physiological Genomics*, 26:158–162, 2006.

E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekk. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, 326:289–293, 2009.

X. Liu, M. Milo, N. D. Lawrence, and M. Rattray. A tractable probabilistic model for Affymetrix probe-level analysis across multiple chips. *Bioinformatics*, 21:3637–3644, 2005.

D. Lockhart, H. Dong, M. Byrne, M. Follettie, M. Gallo, M. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680, 1996.

J. E. Lucas, C. M. Carvalho, J. L.-Y. Chen, J.-T. Chi, and M. West. Cross-study projections of genomic biomarkers: An evaluation in cancer genomics. *PLoS ONE*, 4: e4523, 2009.

M. Lukk, M. Kapushesky, J. Nikkilä, H. Parkinson, A. Goncalves, W. Huber, E. Ukkonen, and A. Brazma. A global map of human gene expression. *Nature Biotechnology*, 28: 322–324, 2010.

S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1: 24–45, 2004.

MAQC Consortium. The microarray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24:1151–1161, 2006.

E. Mayr. *What makes biology unique?: considerations on the autonomy of a scientific discipline*. Cambridge University Press, New York, 2004.

M. N. McCall, B. M. Bolstad, and R. A. Irizarry. Frozen robust multiarray analysis (fRMA). *Biostatistics*, 11:242–53, 2010.

J. D. McPherson. Next-generation gap. *Nature Methods*, 6(S11):S2–5, 2009.

B. H. Mecham, G. T. Klus, J. Strovel, M. Augustus, D. Byrne, P. Bozso, D. Z. Wetmore, T. J. Mariani, I. S. Kohane, and Z. Szallasi. Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Research*, 32:e74, 2004a.

B. H. Mecham, D. Z. Wetmore, Z. Szallasi, Y. Sadovsky, I. Kohane, and T. J. Mariani. Increased measurement accuracy for sequence-verified microarray probes. *Physiological Genomics*, 18:308–315, 2004b.

R. Mei, E. Hubbell, S. Bekiranov, M. Mittmann, F. C. Christians, M.-M. Shen, G. Lu, J. Fang, W.-M. Liu, T. Ryder, P. Kaplan, D. Kulp, and T. A. Webster. Probe selection for high-density oligonucleotide arrays. *Proceedings of the National Academy of Sciences, USA*, 100:11237–11242, 2003.

M. Milo, A. Fazeli, M. Niranjan, and N. Lawrence. A probabilistic model for the extraction of expression levels from oligonucleotide arrays. *Biochemical Society Transactions*, 31:1510–1512, 2003.

D. Montaner and J. Dopazo. Multidimensional gene set analysis of genomic data. *PLoS One*, 5:e10348, 2010.

D. Montaner, P. Minguez, F. Al-Shahrour, and J. Dopazo. Gene set internal coherence in the context of functional profiling. *BMC Genomics*, 10:197, 2009.

Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562, 2002.

P. Müller and F. A. Quintana. Nonparametric Bayesian Data Analysis. *Statistical Science*, 19:95–110, 2004.

C. Myers, D. Robson, A. Wible, M. Hibbs, C. Chiriac, C. Theesfeld, K. Dolinski, and O. Troyanskaya. Discovery of biological networks from diverse functional genomic data. *Genome Biology*, 6:R114, 2005.

S. Myllykangas, S. Junnila, A. Kokkola, R. Autio, I. Scheinin, T. Kiviluoto, M.-L. Karjalainen-Lindsberg, J. Hollmén, S. Knuutila, P. Puolakkainen, and O. Monni. Integrated gene copy number and expression microarray analysis of gastric cancer highlights potential target genes. *International Journal of Cancer*, 123:817–825, 2008.

S. Nacu, R. Critchley-Thorne, P. Lee, and S. Holmes. Gene expression network analysis and applications to immunology. *Bioinformatics*, 23:850–858, 2007.

F. Naef and M. O. Magnasco. Solving the riddle of the bright mismatches: Labeling and effective binding in oligonucleotide arrays. *Physical Review E*, 68:011906, 2003.

D. Nam and S.-Y. Kim. Gene-set approach for expression pattern analysis. *Briefings in Bioinformatics*, 9:189–197, 2008.

J. Noirel, G. Sanguinetti, and P. C. Wright. Identifying differentially expressed subnetworks with MMG. *Bioinformatics*, 24:2792–2793, 2008.

B. A. Novak and A. N. Jain. Pathway recognition and augmentation by computational analysis of microarray expression data. *Bioinformatics*, 22:233–241, 2006.

D. Nuyten and M. van de Vijver. Using microarray analysis as a prognostic and predictive tool in oncology: focus on breast cancer and normal tissue toxicity. *Seminars in Radiation Oncology*, 18:105–114, 2008.

P. Nymark, P. M. Lindholm, M. V. Korpela, L. Lahti, S. Ruosaari, S. Kaski, J. Hollmén, S. Anttila, V. L. Kinnula, and S. Knuutila. Gene expression profiles in asbestos-exposed epithelial and mesothelial lung cell lines. *BMC Genomics*, 8:62, 2007.

A. Ocana and A. Pandiella. Personalized therapies in the cancer "omics" era. *Molecular Cancer*, 9:202, 2010.

Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40:1413–1415, 2008.

S. C. J. Parker, L. Hansen, H. O. Abaan, T. D. Tullius, and E. H. Margulies. Local DNA topography correlates with functional noncoding regions of the human genome. *Science*, 324:389–392, 2009.

E. Parkhomenko, D. Tritchler, and J. Beyene. Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 8:1, 2009.

H. Parkinson, M. Kapushesky, N. Kolesnikov, G. Rustici, M. Shojatalab, N. Abeygunawardena, H. Berube, M. Dylag, I. Emam, A. Farne, E. Holloway, M. Lukk, J. Malone, R. Mani, E. Pilicheva, T. F. Rayner, F. Rezwan, A. Sharma, E. Williams, X. Z. Bradley, T. Adamusiak, M. Brandizi, T. Burdett, R. Coulson, M. Krestyaninova, P. Kurnosov, E. Maguire, S. G. Neogi, P. Rocca-Serra, S.-A. Sansone, N. Sklyar,

M. Zhao, U. Sarkans, and A. Brazma. ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Research*, 37:D868–872, 2009.

L. Parsons, E. Haque, and H. Liu. Subspace clustering for high dimensional data: A review. *Sigkdd Explorations*, 6:90–105, 2004.

H. Pearson. Genetics: what is a gene? *Nature*, 441:398–401, 2006.

J. L. Phillips, S. W. Hayward, Y. Wang, J. Vasselli, C. Pavlovich, H. Padilla-Nash, J. R. Pezullo, B. M. Ghadimi, G. D. Grossfeld, A. Rivera, W. M. Linehan, G. R. Cunha, and T. Ried. The Consequences of Chromosomal Aneuploidy on Gene Expression Profiles in a Cell Line Model for Prostate Carcinogenesis. *Cancer Research*, 61:8143–8149, 2001.

D. Pinkel and D. G. Albertson. Array comparative genomic hybridization and its applications in cancer. *Nature Genetics*, 37:S11–17, 2005.

A. Polanski and M. Kimmel. *Bioinformatics*. Springer, Germany, 2007.

E. Prak and H. Kazazian Jr. Mobile elements and the human genome. *Nature Reviews Genetics*, 1:134–144, 2000.

T. M. Przytycka, M. Singh, and D. K. Slonim. Toward the dynamic interactome: it's about time. *Briefings in Bioinformatics*, 11:15–29, 2010.

L.-X. Qin. An integrative analysis of microRNA and mRNA Expression - a case study. *Cancer Informatics*, 6:369–379, 2008.

J. Quackenbush. Computational analysis of microarray data. *Nature Reviews Genetics*, 2:418–427, 2001.

J. Rachlin, D. D. Cohen, C. Cantor, and S. Kasif. Biological context networks: a mosaic view of the interactome. *Molecular Systems Biology*, 2:66, 2006.

A. Ramasamy, A. Mondry, C. C. Holmes, and D. G. Altman. Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Medicine*, 5:e184, 2008.

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 12*, pages 554–560. MIT Press, Cambridge, MA, 2000.

J. L. Reed, I. Famili, I. Thiele, and B. O. Palsson. Towards multidimensional genome annotation. *Nature Reviews Genetics*, 7:130–141, 2006.

M. Reimers. Making informed choices about microarray data analysis. *PLoS Computational Biology*, 6:e1000786, 2010.

D. Reiss, N. Baliga, and R. Bonneau. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics*, 7: 280, 2006.

S. Rogers, M. Girolami, C. Campbell, and R. Breitling. The latent process decomposition of cDNA microarray data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2:143–156, 2005.

S. Rogers, A. Klami, J. Sinkkonen, M. Girolami, and S. Kaski. Infinite factorization of multiple non-parametric views. *Machine Learning*, 79:201, 2010.

R. Roth, P. Hevezi, J. Lee, D. Willhite, S. Lechner, A. Foster, and A. Zlotnik. Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. *Neurogenetics*, 7:67–80, 2006.

V. Roth and T. Lange. Feature selection in clustering problems. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 473–480. MIT Press, Cambridge, MA, 2004.

J. Russ and M. E. Futschik. Comparison and consolidation of microarray data sets of human tissue expression. *BMC Genomics*, 11:305, 2010.

B. Sadikovic, M. Yoshimoto, K. Al-Romaih, G. Maire, M. Zielenska, and J. A. Squire. In vitro analysis of integrated global high-resolution DNA methylation profiling with genomic imbalance and gene expression in osteosarcoma. *PLoS One*, 3:e2834, 2008.

Y. Saeys, I. Inza, and P. Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23:2507–2517, 2007.

K. Salari, R. Tibshirani, and J. R. Pollack. DR-Integrator: a new analytic tool for integrating DNA copy number and gene expression data. *Bioinformatics*, 26:414–416, 2010.

H. Sara, O. Kallioniemi, and M. Nees. A decade of cancer gene profiling: from molecular portraits to molecular function. *Methods in Molecular Biology*, 576:61–87, 2010.

R. S. Savage, Z. Ghahramani, J. E. Griffin, B. J. de la Cruz, and D. L. Wild. Discovering transcriptional modules by Bayesian data integration. *Bioinformatics*, 26:i158–167, 2010.

E. E. Schadt. Molecular networks as sensors and drivers of common human diseases. *Nature*, 461:218–223, 2009.

M. Schäfer, H. Schwender, S. Merk, C. Haferlach, K. Ickstadt, and M. Dugas. Integrated analysis of copy number alterations and gene expression: a bivariate assessment of equally directed abnormalities. *Bioinformatics*, 25:3228–3235, 2009.

I. Scheinin, S. Myllykangas, I. Borze, T. Bohling, S. Knuutila, and J. Saharinen. CanGEM: mining gene copy number changes in cancer. *Nucleic Acids Research*, 36:D830–835, 2008.

U. Scherf, D. T. Ross, M. Waltham, L. H. Smith, J. K. Lee, L. Tanabe, K. W. Kohn, W. C. Reinhold, T. G. Myers, D. T. Andrews, D. A. Scudiero, M. B. Eisen, E. A. Sausville, Y. Pommier, D. Botstein, P. O. Brown, and J. N. Weinstein. A gene expression database for the molecular pharmacology of cancer. *Nature Genetics*, 24:236–44, 2000.

B. Schölkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT Press, USA, 2002.

J. W. Schopf. Fossil evidence of Archaean life. *Philosophical Transactions of the Royal Society of London.* Series B, 361:869–885, 2006.

E. Schrödinger. *What is life? Mind and Matter*. Cambridge University Press, 1944.

J. Sebat. Major changes in our dna lead to major changes in our thinking. *Nature Genetics*, 39:S3–5, 2007.

E. Segal, A. Battle, and D. Koller. Decomposing gene expression into cellular processes. In R. B. Altman, A. K. Dunker, L. Hunter, T. A. Jung, and T. E. Klein, editors, *Proceedings of Pacific Symposium on Biocomputing (PSB 2003)*, pages 89–100. World Scientific, Singapore, 2003a.

E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34:166–176, 2003b.

E. Segal, B. Taskar, A. Gasch, N. Friedman, and D. Koller. Rich probabilistic models for gene expression. *Bioinformatics*, 17(S1):i243–252, 2003c.

E. Segal, H. Wang, and D. Koller. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, 19(S1):i264–272, 2003d.

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

D. F. Shanno. Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, 24:647–656, 1970.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

A. J. Sharp, Z. Cheng, and E. E. Eichler. Structural variation of the human genome. *Annual Review of Genomics and Human Genetics*, 7:407–442, 2006.

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

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

C. Sigg, B. Fischer, B. Ommer, V. Roth, and J. Buhmann. Nonnegative CCA for audiovisual source separation. In *Proceedings MLSP'07 IEEE International Workshop on Machine Learning for Signal Processing*, pages 253–258. IEEE Signal Processing Society, Zurich, 2007.

J. Sinkkonen, S. Kaski, and J. Nikkilä. Discriminative clustering: Optimal contingency tables by learning metrics. In T. Elomaa, H. Mannila, and H. Toivonen, editors, *Proceedings of the ECML'02, 13th European Conference on Machine Learning*, pages 418–430. Springer, Berlin, 2002.

J. Sinkkonen, J. Nikkilä, L. Lahti, and S. Kaski. Associative clustering by maximizing a Bayes factor. Technical Report A68, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, 2003.

J. Sinkkonen, S. Kaski, J. Nikkilä, and L. Lahti. Associative Clustering (AC): Technical Details. Technical Report A84, Helsinki University of Technology, Espoo, Finland, 2005.

E. Sliwerska, F. Meng, T. Speed, E. Jones, W. Bunney, H. Akil, S. Watson, and M. Burmeister. SNPs on chips: the hidden genetic code in expression arrays. *Biological Psychiatry*, 61:13–16, 2007.

J. Sommer. The delay in sharing research data is costing lives. *Nature Medicine*, 16:744, 2010.

C. Soneson, H. Lilljebjorn, T. Fioretos, and M. Fontes. Integrative analysis of gene expression and copy number alterations using canonical correlation analysis. *BMC Bioinformatics*, 11:191, 2010.

S. Sonnenburg, M. L. Braun, C. S. Ong, S. Bengio, L. Bottou, G. Holmes, Y. LeCun, K.-R. Müller, F. Pereira, C. E. Rasmussen, G. Rätsch, B. Schölkopf, A. Smola, P. Vincent, J. Weston, and R. Williamson. The need for open source software in machine learning. *The Journal of Machine Learning Research*, 8:2443–2466, 2007.

T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lønning, and A.-L. Børresen-Daleb. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences, USA*, 98:10869–10874, 2001.

J. Stetefeld and M. A. Ruegg. Structural and functional diversity generated by alternative mRNA splicing. *Trends in Biochemical Sciences*, 30:515–521, 2005.

V. Stodden. The scientific method in practice: Reproducibility in the computational sciences. *MIT Sloan Research Paper*, 4773–10, 2010.

M. R. Stratton, P. J. Campbell, and P. A. Futreal. The cancer genome. *Nature*, 458: 719–724, 2009.

L. Strömbäck and P. Lambrix. Representations of molecular pathways: an evaluation of SBML, PSI MI and BioPAX. *Bioinformatics*, 21:4401–4407, 2005.

J. M. Stuart, E. Segal, D. Koller, and S. K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302:249–255, 2003.

A. I. Su, M. P. Cooke, K. A. Ching, Y. Hakak, J. R. Walker, T. Wiltshire, A. P. Orth, R. G. Vega, L. M. Sapinoso, A. Moqrich, A. Patapoutian, G. M. Hampton, P. G. Schultz, and J. B. Hogenesch. Large-scale analysis of the human and mouse transcriptomes. *Proceedings of the National Academy of Sciences, USA*, 99:4465–4470, 2002.

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. *Proceedings of the National Academy of Sciences, USA*, 101:6062–6067, 2004.

J. Su, B.-J. Yoon, and E. R. Dougherty. Accurate and reliable cancer classification based on probabilistic inference of pathway activity. *PLoS ONE*, 4:e8161, 2009.

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences, USA*, 102:15545–15550, 2005.

S. Suthram, J. T. Dudley, A. P. Chiang, R. Chen, T. J. Hastie, and A. J. Butte. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS Computational Biology*, 6:e1000662, 2010.

P. A. C. 't Hoen, Y. Ariyurek, H. H. Thygesen, E. Vreugdenhil, R. H. A. M. Vossen, R. X. de Menezes, J. M. Boer, G.-J. B. van Ommen, and J. T. den Dunnen. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Research*, 36: e141, 2008.

P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrowsky, E. S. Lander, and T. R. Golub. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences, USA*, 96:2907–2912, 1999.

A. Tanay, R. Sharan, and R. Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18:S136–144, 2002.

A. Tanay, R. Sharan, M. Kupiec, and R. Shamir. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proceedings of the National Academy of Sciences*, 101:2981–2986, 2004.

A. Tanay, I. Steinfeld, M. Kupiec, and R. Shamir. Integrative analysis of genome-wide experiments in the context of a large high-throughput data compendium. *Molecular Systems Biology*, 1:0002, 2005.

A. L. Tarca, S. Draghici, P. Khatri, S. S. Hassan, P. Mittal, J.-S. Kim, C. J. Kim, J. P. Kusanovic, and R. Romero. A novel signaling pathway impact analysis. *Bioinformatics*, 25:75–82, 2009.

B. S. Taylor, J. Barretina, N. D. Socci, P. DeCarolis, M. Ladanyi, M. Meyerson, S. Singer, and C. Sander. Functional copy-number alterations in cancer. *PLoS ONE*, 3:e3179, 2008.

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

The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447:799–816, 2007.

R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences, USA*, 99:6567–6572, 2002.

M. M. Tice and D. R. Lowe. Photosynthetic microbial mats in the 3,416-Myr-old ocean. *Nature*, 431:549–552, 2004.

C. Tilstone. Vital statistics. *Nature*, 424:610–612, 2003.

N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *37th Annual Allerton Conference on Communication, Control, and Computing*, pages 368–377. University of Illinois, Urbana, Illinois, 1999.

P. Törönen, M. Kolehmainen, G. Wong, and E. Castrén. Analysis of gene expression data using self-organizing maps. *FEBS Letters*, 451:142–146, 1999.

O. G. Troyanskaya. Putting microarrays in a context: Integrated analysis of diverse biological data. *Briefings in Bioinformatics*, 6:34–43, 2005.

Y. Tu, G. Stolovitzky, and U. Klein. Quantitative noise analysis for gene expression microarray experiments. *Proceedings of the National Academy of Sciences, USA*, 99: 14031–14036, 2002.

J. Tukey. *Exploratory data analysis*. Addison-Wesley, Reading, MA, 1977.

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

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

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

W. N. van Wieringen and M. A. van de Wiel. Nonparametric testing for DNA copy number induced differential mRNA gene expression. *Biometrics*, 65:19–29, 2009.

C. J. Vaske, S. C. Benz, J. Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler, and J. M. Stuart. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, 26:i237–245, 2010.

I. Vastrik, P. D'Eustachio, E. Schmidt, G. Joshi-Tope, G. Gopinath, D. Croft, B. de Bono, M. Gillespie, B. Jassal, S. Lewis, L. Matthews, G. Wu, E. Birney, and L. Stein. Reactome: a knowledge base of biologic pathways and processes. *Genome Biology*, 8:R39, 2007.

J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. D. Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R.-R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Y. Wang, A. Wang, X. Wang, J. Wang, M.-H. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. C. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage, F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y.-H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y.-H. Chiang, M. Coyne, C. Dahlke, A. D. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The Sequence of the Human Genome. *Science*, 291:1304–1351, 2001.

H. Vinod. Canonical ridge and the econometrics of joint production. *Journal of Econometrics*, 4:147–166, 1976.

S. Volinia, M. Galasso, S. Costinean, L. Tagliavini, G. Gamberoni, A. Drusco, J. Marchesini, N. Mascellani, M. E. Sana, R. Abu Jarour, C. Desponts, M. Teitell, R. Baffa, R. Aqeilan, M. V. Iorio, C. Taccioli, R. Garzon, G. Di Leva, M. Fabbri, M. Catozzi, M. Previati, S. Ambs, T. Palumbo, M. Garofalo, A. Veronese, A. Bottoni, P. Gasparini, C. C. Harris, R. Visone, Y. Pekarsky, A. de la Chapelle, M. Bloomston, M. Dillhoff, L. Z. Rassenti, T. J. Kipps, K. Huebner, F. Pichiorri, D. Lenze, S. Cairo, M. A. Buendia, P. Pineau, A. Dejean, N. Zanesi, S. Rossi, G. A. Calin, C. G. Liu, J. Palatini, M. Negrini, A. Vecchione, A. Rosenberg, and C. M. Croce. Reprogramming of miRNA networks in cancer and leukemia. *Genome Research*, 20:589–599, 2010.

S. Waaijenborg, P. C. Verselewel, d. W. Hamer, and A. H. Zwinderman. Quantifying the Association between Gene Expressions and DNA-Markers by Penalized Canonical Correlation Analysis. *Statistical Applications in Genetics and Molecular Biology*, 7:3, 2008.

J. Watson and F. Crick. A structure for deoxyribose nucleic acid. *Nature*, 171:737–738, 1953.

M. West. Bayesian factor regression models in the 'large p, small n' paradigm. *Bayesian statistics*, 7:723–732, 2003.

D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, W. Helmberg, D. L. Kenton, O. Khovayko, D. J. Lipman, T. L. Madden, D. R. Maglott, J. Ostell, J. U. Pontius, K. D. Pruitt, G. D. Schuler, L. M. Schriml, E. Sequeira, S. T. Sherry, K. Sirotkin, G. Starchenko, T. O. Suzek, R. Tatusov, T. A. Tatusova, L. Wagner, , and E. Yaschenko. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 33:D39–45, 2005.

D. J. Wilkinson. Bayesian methods in bioinformatics and computational systems biology. *Briefings in Bioinformatics*, 8:109–116, 2007.

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

C. Wu, R. Carta, and L. Zhang. Sequence dependence of cross-hybridization on short oligo microarrays. *Nucleic Acids Research*, 33:e84, 2005.

Z. Wu and R. Irizarry. Stochastic models inspired by hybridization theory for short oligonucleotide arrays. In P. E. Bourne and D. Gusfield, editors, *Proceedings of the 8th Annual International Conference on Computational Molecular Biology (RECOMB'04)*, pages 98–106. ACM Press, New York, 2004.

V. Wunderlich. Early references to the mutational origin of cancer. *International Journal of Epidemiology*, 36:246–247, 2007.

Y. Yamanishi, J.-P. Vert, A. Nakaya, and M. Kanehisa. Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis. *Bioinformatics*, 19:i323–330, 2003.

Y. Yamanishi, M. Kotera, M. Kanehisa, and S. Goto. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics*, 26:i246–254, 2010.

F. Yates. Contingency tables involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society Supplement*, 1:217–239, 1934.

C. L. Yauk, M. L. Berndt, A. Williams, and G. R. Douglas. Comprehensive comparison of six microarray technologies. *Nucleic Acids Research*, 32:e124, 2004.

J. Zhang, R. P. Finney, R. J. Clifford, L. K. Derr, and K. H. Buetow. Detecting false expression signals in high-density oligonucleotide arrays by an in silico approach. *Genomics*, 85:297–308, 2005.

L. Zhang, L. Wang, A. Ravindranathan, and M. Miles. A new algorithm for analysis of oligonucleotide arrays: Application to expression profiling in mouse brain regions. *Journal of Molecular Biology*, 317:225–235, 2002.

W. Zhang, Q. Morris, R. Chang, O. Shai, M. Bakowski, N. Mitsakakis, N. Mohammad, M. Robinson, R. Zirngibl, E. Somogyi, N. Laurin, E. Eftekharpour, E. Sat, J. Grigull, Q. Pan, W.-T. Peng, N. Krogan, J. Greenblatt, M. Fehlings, D. van der Kooy, J. Aubin, B. Bruneau, J. Rossant, B. Blencowe, B. Frey, and T. Hughes. The functional landscape of mouse gene expression. *Journal of Biology*, 3:21, 2004.

X. Zhou and G. Gibson. Cross-species comparison of genome-wide expression patterns. *Genome Biology*, 5:232, 2004.

D. Zhu, A. O. Hero, H. Cheng, R. Khanna, and A. Swaroop. Network constrained clustering for gene microarray data. *Bioinformatics*, 21:4014–4020, 2005.

I

Publication I

Laura L. Elo, Leo Lahti, Heli Skottman, Minna Kyläniemi, Riitta Lahesmaa, and Tero Aittokallio. Integrating probe-level expression changes across generations of Affymetrix arrays. *Nucleic Acids Research*, 33(22):e193, 2005.

Published online December 14, 2005

Nucleic Acids Research, 2005, Vol. 33, No. 22 ***e193***
doi:10.1093/nar/gni193

Integrating probe-level expression changes across generations of Affymetrix arrays

Laura L. Elo[1,2,*], Leo Lahti[2,3], Heli Skottman[2,4], Minna Kyläniemi[2], Riitta Lahesmaa[2] and Tero Aittokallio[1,2]

[1]Department of Mathematics, FIN-20014 University of Turku, Finland, [2]Turku Centre for Biotechnology, PO Box 123, FIN-20521 Turku, Finland, [3]Laboratory of Computer and Information Science, Helsinki University of Technology, PO Box 5400, FIN-02015 HUT, Finland and [4]Institute for Regenerative Medicine Regea, University of Tampere and Tampere University Hospital, FIN-33520, Tampere, Finland

Received August 16, 2005; Revised October 31, 2005; Accepted November 28, 2005

ABSTRACT

There is an urgent need for bioinformatic methods that allow integrative analysis of multiple microarray data sets. While previous studies have mainly concentrated on reproducibility of gene expression levels within or between different platforms, we propose a novel meta-analytic method that takes into account the vast amount of available probe-level information to combine the expression changes across different studies. We first show that the comparability of relative expression changes and the consistency of differentially expressed genes between different Affymetrix array generations can be considerably improved by determining the expression changes at the probe-level and by considering the latest information on probe-level sequence matching instead of the probe annotations provided by the manufacturer. With the improved probe-level expression change estimates, data from different generations of Affymetrix arrays can be combined more effectively. This will allow for the full exploitation of existing results when designing and analyzing new experiments.

INTRODUCTION

The enormous popularity of gene expression profiling with microarrays in recent years has resulted in a rapid accumulation of data in many laboratories and public databases. As microarray experiments are expensive and often involve biological samples that are difficult to obtain, sample sizes in typical microarray studies are relatively small, leading to several false-positive and false-negative findings. Therefore, methods that can effectively extract information from previous studies are of practical interest for minimizing the number of additional experiments needed without compromising the reliability of the results. However, combining data across studies performed at different times and perhaps in different laboratories is a challenging task where both biological and technical sources of variability must be considered carefully.

A major problem in integrative analysis is that gene expression data generated with different microarray platforms are not directly comparable, and even within the same technique different protocols for sample preparation, array hybridization and data analysis can result in severe variations among data sets. Accordingly, the early cross-platform comparisons often showed poor correlation between their intensity measurements (1,2). More recent studies have showed that implementation of standardized protocols for all steps of the microarray study can markedly increase reproducibility between platforms and even across laboratories (3,4). However, some of the variation can be beyond the capacity of standard normalization techniques if the remaining discrepancies between data sets originate from measuring different splice variants of the same gene (5).

As the compositions of microarrays are regularly updated to incorporate new genes with improved target sequences, it is difficult to combine data even from different generations of the same microarray platform. In particular, Affymetrix high-density oligonucleotide arrays utilize multiple (typically 8–16) 25mer probes, the so-called probe set, to measure the expression level of a transcript target. Although the use of several probes for each target leads to more robust estimates of transcript activity, it is clear that probe qualities may significantly affect the results of a study. It has been noticed that a considerable number of probes on various high-density oligonucleotide arrays do not uniquely match their intended targets (6–9). By matching the probe sequences to the most up-to-date genomic sequence data, it is possible to assess the quality of the probes. Redefinition of probe sets according to the latest probe sequence information can increase their accuracy and cross-platform consistency with other array types (6,8,9).

*To whom correspondence should be addressed. Tel: +358 2 333 8002; Fax: +358 2 333 8000; Email: laliel@utu.fi

Previous works on different generations of Affymetrix arrays have concentrated mainly on the reproducibility of their expression results. In a comparison of two Affymetrix arrays, HuGeneFL and HG-U95A, Nimgaonkar *et al*. (10) concluded that the reproducibility is high only when the corresponding probe sets share many exact probes. Hwang *et al*. (7) advanced the comparison analysis by selecting subsets of probes with overlapping sequence segments and recalculating expression values using the selected probes only. While such probe filtering could significantly improve the reproducibility between Affymetrix HG-U95Av2 and HG-U133A arrays, some useful information from the non-overlapping probes measuring identical targets may be lost. In fact, from the investigator's point of view, the enhanced comparability is of practical importance only when the probes match identical targets.

In the present work, we continue the integrative analysis across generations of Affymetrix arrays by considering explicitly the actual targets of probe sequences rather than their similarities. As most current arrays with an enhanced probe design protocol contain high quality probes that do not share sequence similarity with the older probes, we do not filter probes based on overlap but utilize all available probe-level information across generations. We carry out a thorough examination of two in-house data sets, containing expression data from human HG-U133A and HG-U133Plus2.0 arrays and murine MG-U74Av2 and mouse MOE430 2.0 arrays. Additionally, we consider two publicly available data sets, containing expression data from human HG-U95Av2 and HG-U133A arrays. Each data set contains technical replicates hybridized to two array types, allowing us to isolate the array-effects from the underlying biological variation. Since the technical replicates are assumed to produce the same results on both arrays, the comparability of the arrays can be directly evaluated. We also investigate several different probe set pairing approaches in the comparison studies.

Toward combining results from multiple studies, we propose a novel meta-analytic framework, based on the selected probe set pairing method and our probe-level estimate of expression changes (referred to as PECA). The performance of this procedure is demonstrated on a public data set, which also contains several biological samples hybridized to both HG-U95Av2 and HG-U133A arrays. The meta-analysis method is evaluated in terms of its stability when the sample size is reduced. As agreement between the pure expression measurements do not consider the platform-specific probe-effects, which arise from inherent differences in the hybridization efficiency of different probes, we also use relative expression changes when evaluating the methods. Besides removing the probe-effects, expression changes are often more meaningful for the investigator, as the main interest in most studies is in identifying a set of candidate genes that are differentially expressed between groups of samples instead of their plain expression levels.

MATERIALS AND METHODS

Human embryonic stem cell data (hESC)

Two human embryonic stem cell (hESC) lines, HS306 and HS293, from Karolinska University Hospital (Huddinge, Sweden) were derived and cultured in serum replacement medium on human foreskin fibroblast feeder cells as described previously (11). The total RNA was isolated from 5 to 10 hESC colonies using the RNAeasy mini kit (Qiagen, Valencia, CA). The sample preparation was performed according to the Affymetrix two-cycle GeneChip® Eukaryotic small sample target labeling assay version II (Affymetrix, Santa Clara, CA). The samples were hybridized to human HG-U133A and HG-U133Plus2.0 arrays (Table 1).

Mouse Chlamydia pneumonia infection data (mCPI)

Female inbred Balb/c mice obtained from Harlan Netherlands (Horst, The Netherlands) were infected with *Chlamydia pneumoniae* as described previously (12). The axillary lymph nodes from 12 control mice and the mediastinal lymph nodes from 12 infected and 12 re-infected mice were pooled. The total RNA from CD4+ cells were isolated using the Trizol method (Invitrogen Co., Carlsbad, CA) and further purified with RNAeasy mini kit. The sample preparation was performed according to the Affymetrix two-cycle GeneChip® Eukaryotic small sample target labeling assay version II. The samples were hybridized to murine MG-U74Av2 and mouse MOE430 2.0 arrays (Table 1).

Human acute lymphoblastic leukemia data (ALL)

The public data sets from the microarray studies of Yeoh *et al*. (13) and Ross *et al*. (14) contained expression data from ALL patients with different leukemia subtypes. A total of 360 patient samples were hybridized to HG-U95Av2 arrays and 132 of the same samples were also hybridized to

Table 1. Hybridization scheme

Data set	Condition	Samples	HG-U133Plus2.0	HG-U133A	HG-U95Av2	MOE430 2.0	MG-U74Av2
hESC	HS293	2	1	1	—	—	—
hESC	HS306	2	1	1	—	—	—
mCPI	Control	1	—	—	—	1	2
mCPI	Infected	1	—	—	—	1	1
mCPI	Re-infected	1	—	—	—	1	2
ALL	T-ALL	14	—	1	1	—	—
ALL	E2A-PBX1	18	—	1	1	—	—
IM	Dermatomyositis	5	—	1	1	—	—
IM	Other myopathy	9	—	1	1	—	—

The third column indicates the number of samples in each condition. The rest of the columns are the number of hybridizations per sample in each sample set on different array types.

HG-U133A arrays. We selected for our analyses 32 samples that were hybridized to both array types and represented two genetically distinct leukemia subtypes: 14 T-ALL samples and 18 E2A-PBX1 samples (Table 1). To obtain equal sample sizes in both groups, we randomly excluded 4 E2A-PBX1 samples from the analysis.

Human inflammatory myopathies data (IM)

The publicly available data set from the study of Hwang *et al*. (7) contained muscle tissue samples from 14 patients with inflammatory myopathies. The patients were divided into two groups: five patients had dermatomyositis and nine patients had other inflammatory myopathies. Each sample was hybridized to HG-U95Av2 and HG-U133A arrays. To make the present results directly comparable with the results obtained by Hwang *et al*. we included all the samples into our study.

Probe sequence data

Probe sequences and their 'bestmatch' tables were downloaded from the Affymetrix web pages (www.affymetrix.com). Other array-wise information on probes and probe sets, including GeneID annotations, were provided with annotation data packages of the Bioconductor project (15). Genomic mRNA sequences for alignments were downloaded from Entrez nucleotide (16) for human (March 3, 2005) and mouse (April 29, 2005), excluding EST, STS, GSS, 'working draft' and 'patents' sequences, and sequences with a 'XM_' tag, as in (7). The Entrez mRNA sequences were assigned to GeneID identifiers by using the gene2accession conversion file obtained from NCBI (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA) for human (March 23, 2005) and mouse (April 28, 2005). This resulted in a total of 209 650 and 183 461 mRNA sequences for human and mouse, respectively. The probes in the AFFX-control sets were omitted from the analysis.

Probe verification

To guarantee the quality and comparability of the 25mer probes, we verified them using the Entrez mRNA sequence database (16). Perfect matches of the probes to mRNA sequences were searched with BLAT v. 26 (17). A given probe often matches several mRNA targets. In such cases, it is common that the mRNA sequences are merely separate sequence submissions of the same gene. To distinguish between probes with unique and multiple gene targets, we assigned the Entrez mRNA sequences to GeneID identifiers (18).

The probes were classified according to their manufacturer annotations and our Entrez verifications. *Verified* probes are detected to match Entrez mRNA sequences with a unique GeneID. Probes with no matching GeneID targets are *mistargeted*, and those assigned to several GeneIDs are *non-specific*. A probe is *conflicting* if its verified target is different from the one in the array-wise annotations. *A verified probe set* is a subset of the corresponding original probe set, obtained by masking the mistargeted, non-specific and conflicting probes from the original set. An *alternative probe set* is a collection of probes on a given array that are verified to uniquely measure a given GeneID. An alternative probe set contains verified probes only, but these may include probes from various original probe sets.

Probe set pairing

A common approach to compare different generations of Affymetrix arrays is to use the so-called 'bestmatch' tables provided by the array manufacturer. The best match pairs are based on the similarity between the target sequences of the probe sets. Since the HG-U133Plus2.0 array contains all the probe sets from the HG-U133A and HG-U133B arrays, plus 9921 additional probe sets, the HG-U133A and HG-U133Plus2.0 arrays can be compared by selecting the same probe sets from the two arrays. We consider these pairs as best match pairs as well, although this is a much stricter pairing criterion than the one usually characterizing the best match pairs.

An alternative approach for probe set pairing is to use GeneID identifiers. Original and verified probe sets on both arrays can be assigned to GeneIDs by using the array-wise annotations. As these are not available for alternative probe sets, we used the verified GeneIDs from our Entrez studies. We only considered those GeneIDs for which corresponding probes existed on original, verified and alternative probe sets.

Probe-level expression change averaging (PECA)

We based the selection of genes differentially expressed between two particular groups of samples on probe-level microarray data instead of probe set-level summary intensities obtained with, for instance, robust multi-array average (RMA) (19) or Affymetrix microarray suite (MAS) (www.affymetrix.com). More specifically, we first calculated the selected test statistic separately for each probe in the data and then averaged over the probes within each probe set. In the calculations, we used perfect match (PM) intensities, which were quantile-normalized (20) and log-transformed before the analysis. We refer to this procedure as PECA.

We considered two types of PECA-measures within a microarray study: the signal log-ratio and the Hedges' g, which is a commonly used effect size estimate in meta-analysis (21). Let the normalized logarithmic PM intensities of the probe j in the probe set i under the two conditions within a study be $x_{ij} = (x_{ij1}, \ldots, x_{ijn_1})$ and $y_{ij} = (y_{ij1}, \ldots, y_{ijn_2})$ where the total number of samples within the study is $n = n_1 + n_2$. The signal log-ratio is then defined as $d_{ij} = \bar{x}_{ij} - \bar{y}_{ij}$, and the Hedges' g as $g_{ij} = a(\bar{x}_{ij} - \bar{y}_{ij})/s_{ij}$, where $\bar{x}_{ij}$ and $\bar{y}_{ij}$ are the means of the two groups, s_{ij} is the pooled standard deviation, and $a = 1 - 3/(4n - 9)$ is a correction term that makes the Hedges' g-estimate unbiased. After calculating the probe-level estimates, the probe set-level estimates were formed by averaging over the probes within each probe set. In the present study, the probe sets were defined using the various probe verification criteria and the PECA-estimates were calculated separately within each study on each array generation.

Meta-analysis of effect sizes

Suppose that m studies produce effect size estimates e_k and measures of variability s_k^2, $k = 1, \ldots, m$. Assume that all studies estimate the same parameter μ and any differences between the estimates are due to sampling error $\epsilon_k \sim N(0, s_k^2)$. Then the meta-analysis estimate for μ is the weighted average

over the effect size estimates

$$\hat{\mu} = \frac{\sum_{k=1}^{m} w_k e_k}{\sum_{k=1}^{m} w_k},$$

where the weight w_k is defined as $w_k = s_k^{-2}$. The variance of $\hat{\mu}$ is $s_{\hat{\mu}} = 1/\sum w_k$, and hence the hypothesis H_0:$\mu = 0$ can be considered by using the test statistic $Z = \hat{\mu}/s_{\hat{\mu}}$, which is distributed as $N(0,1)$ under the null hypothesis H_0. For a detailed description of this technique, see (21). Such meta-analytic method was applied in the present study to combine the expression changes across the array types.

TESTING PROCEDURE

We first evaluated the effect of the different probe set pairing and probe verification criteria on the reproducibility of RMA- and MAS-normalized signal intensities between each array pair on which the same sample was hybridized (between-study comparison). We then investigated the comparability of relative expression changes and the agreement of differentially expressed genes between these array pairs using the GeneID matched alternative probe sets (between-study comparison). At this stage, the expression changes were calculated within each array generation (within-study analysis) using the PECA-procedure (probe-level estimation) and the summary intensities from RMA and MAS 5.0 (probe set-level estimation). Finally, we used the meta-analysis approach to combine the expression changes from the different array generations (between-study analysis). The meta-analysis was carried out using PECA-estimated Hedges' *g*-values (probe-level between-study analysis) as well as Hedges'*g*-values calculated from the RMA-derived intensities (probe set-level between-study analysis).

Reproducibility of signal intensities

To assess the level of reproducibility of signal intensities between technical replicates across array generations, we calculated the Pearson correlation coefficient between each array pair from the same sample. The intensity values were obtained with RMA and MAS 5.0. We compared the intensities between the best match pairs of original probe sets and verified probe sets as well as GeneID pairs with three different collections of probes: (i) original Affymetrix probe sets, (ii) verified probe sets and (iii) alternative probe sets. If multiple probe sets corresponded to the same GeneID, their values were averaged (22). On each array, the variability in the intensity values among the probe sets corresponding to the same GeneID was investigated for the 10 GeneIDs with the largest number of probe sets.

Comparability of relative expression changes

The comparability of relative expression changes between alternative probe sets on two array generations was investigated by considering signal log-ratios and Hedges'*g*-values between two particular groups of samples in the hESC, mCPI, ALL and IM data. In addition, we randomly generated 100 subsamples of sizes 2–5 from the ALL data set to study more carefully the performance of the Hedges'*g* with small sample sizes. In each array comparison, two replicate estimates corresponding to the same samples on the different arrays were obtained. We used the Pearson correlation coefficient between these estimates as a measure of comparability between the arrays. The expression changes were calculated using the PECA-approach and the RMA- and MAS-normalized intensities.

Agreement of differentially expressed genes

The agreement of the most differentially expressed genes between the array generations was investigated by ranking the genes according to signal log-ratios and Hedges' *g*-estimates and calculating the proportion of common genes among the top N genes in both array types. If two array generations are comparable, the corresponding lists of differentially expressed genes should contain many overlapping genes (3). Again, we used PECA-estimates and the corresponding estimates obtained using the RMA- and MAS-intensity values in the context of alternative probe sets.

Performance of the meta-analysis

The meta-analysis of PECA-based Hedges'*g*-values was compared with the meta-analysis calculated from the RMA-based summary intensities (23). We also compared the performance of both meta-analyses with the corresponding analyses on the individual data sets. The performance of the methods was evaluated by considering the stability of their results when the number of biological samples was reduced (24). We randomly generated 100 subsamples of sizes 2–5 from the ALL data set and applied each method to them. The results of each subset were then compared with the results obtained from the whole data set by determining the proportion of common genes among the top 100 genes.

RESULTS

While most of the probes on the arrays studied could be confirmed to uniquely match a GeneID, a considerable number of probes were rejected since they were either mistargeted, non-specific or conflicting (Table 2). The number of mistargeted probes was especially high on the HG-U133Plus2.0 and MOE430 2.0 arrays, whereas non-specific and conflicting probes were less common. The high number of mistargeted probes on the two arrays is mainly due to the large number of EST-targeted probe sets on these arrays. Our probe verification did not check probes for matches against EST sequences that often lack GeneID assignment but have been used for

Table 2. Probe verification summary

Array type	Probes	Verified (%)	Mistargeted (%)	Non-specific (%)	Conflicting (%)
HG-U133Plus2.0	604 258	58.2	40.2	1.6	2.6
HG-U133A	247 965	82.5	14.4	3.0	3.1
HG-U95Av2	199 084	82.6	14.4	3.0	2.8
MOE430 2.0	496 468	68.2	30.8	1.1	4.9
MG-U74Av2	197 993	73.1	24.2	2.7	1.3

Probes matched to mRNAs with a unique GeneID in Entrez database are considered verified. Mistargeted probes could not be assigned to a GeneID, whereas non-specific probes have several GeneID targets. If the verified target of a probe is different from the annotations provided by the Bioconductor array packages, the probe is considered conflicting.

Table 3. Numbers of probe sets included into the comparisons

Data set	Array comparison	Best match pairs	GeneID pairs	Multiple original sets (%)	Multiple verified sets (%)
hESC	HG-U133A vs. HG-U133Plus2.0	—	12661	36.7	32.2
mCPI	MG-U74Av2 vs. MOE430 2.0	8595	7735	26.4	14.9
ALL, IM	HG-U95Av2 vs. HG-U133A	8429	8240	25.2	18.9

The best match pairs provided by Affymetrix are based on the similarity of the target sequences of the probe sets. The GeneID pairs were obtained by assigning the probe sets to GeneID identifiers. Only GeneIDs for which probes existed on original, verified and alternative probe sets were considered. If multiple probe sets corresponded to the same GeneID, their values were averaged. The last two columns show the proportion of GeneIDs with multiple probe sets when GeneID pairs of original and verified probe sets were formed.

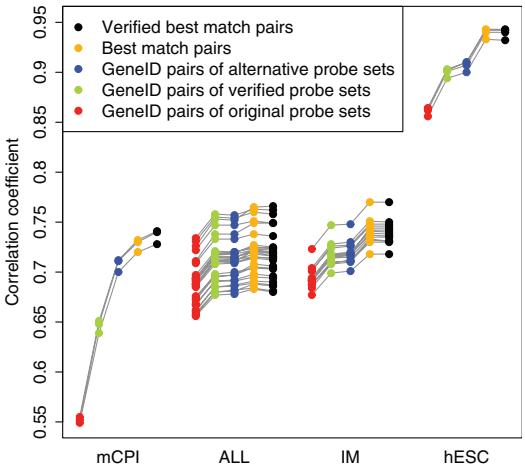


Figure 1. The RMA intensity correlations between technical replicates on two array generations. The Pearson correlation was calculated between each array pair from the same sample. The gray lines show which correlations were obtained from the same array pair with the different probe matching criteria. In the hESC array comparison, the best match probe sets contained exactly the same probes on both array generations, which resulted in very high correlations. The advantages of probe verification and alternative mappings were largest when arrays with different probe collections were compared, as in the mCPI, ALL and IM array comparisons.

the design of several probe sets on these arrays. By simply ignoring the mistargeted and non-specific probes, we were still left with a large number of good-quality probes with a unique GeneID assignment. The typical sizes of the alternative probe sets were approximately the size of the original Affymetrix probe set or its multiplier (see Supplementary Figure 1). The proportion of alternative sets with <5 probes was relatively small, varying between 0.4% (MOE430 2.0) and 2.1% (MG-U74Av2). The numbers of probe sets included into each comparison are listed in Table 3, along with the proportions of GeneIDs with multiple original Affymetrix probe sets.

Effect of probe matching methods on the array reproducibility

Figure 1 illustrates for each array comparison the RMA-based intensity correlations between the pairs of arrays to which the same sample was hybridized. Similar results were obtained with MAS intensities (data not shown). In each array comparison, GeneID pairs of the manufacturer-defined original probe sets performed worst. Probe verification of these sets improved the correlations. Moreover, it was observed that probe verification improved the consistency of the measurements within an array (see Supplementary Figure 2). In the mCPI array comparison, the alternative probe sets produced higher correlations than the verified probe sets, whereas in the ALL, IM and hESC array comparisons the verified sets and the alternative sets performed equally well. In the ALL comparison, also the best match pairs performed similarly, whereas in the mCPI, IM and hESC comparisons, the best match pairs could still improve the correlations. As expected, the improvement was largest in the hESC data, where the best match pairs contained only probes that were the same on both arrays. Interestingly, the verification of the original Affymetrix probe sets used in the best match pairs did not considerably affect the reproducibility of the signal intensities in any of the array comparisons.

Effect of probe-level effect size estimates on the array comparability

The correlations of signal log-ratios and Hedges' *g*-estimates between each replicate study with different array types are shown in Figure 2. In all comparisons, the PECA-estimates showed consistently the best comparability between the array types. The estimates calculated using MAS summary intensities performed generally poorest. With signal log-ratios, the RMA-based estimates usually reduced only slightly the comparability as compared with the PECA-estimates (Figure 2A). With Hedges' *g*-values, however, the benefit from using PECA was considerably higher, especially with small sample sizes (Figure 2B and C). In the hESC data, the correlation increased from below 0.1 with RMA to ~0.7 with PECA (Figure 2B). Similar results were obtained with the ALL data when only two samples from both patient groups were included into the analysis (Figure 2C). As the number of samples increased, the differences between the methods became smaller.

Effect of probe-level effect size estimates on the array agreement

Figure 3 shows the agreement of the most differentially expressed genes between the array types when two groups of samples in the ALL, IM and hESC data were compared. The best agreement was consistently achieved with the PECA-estimates, whereas with MAS-based estimates the correspondence of the top genes between the arrays was poorest. Especially in the hESC array comparison, the superiority of the PECA-method was drastic as compared to the probe set-level methods. For example, with signal log-ratios, the percentage of common genes among top 30 genes was ~25%

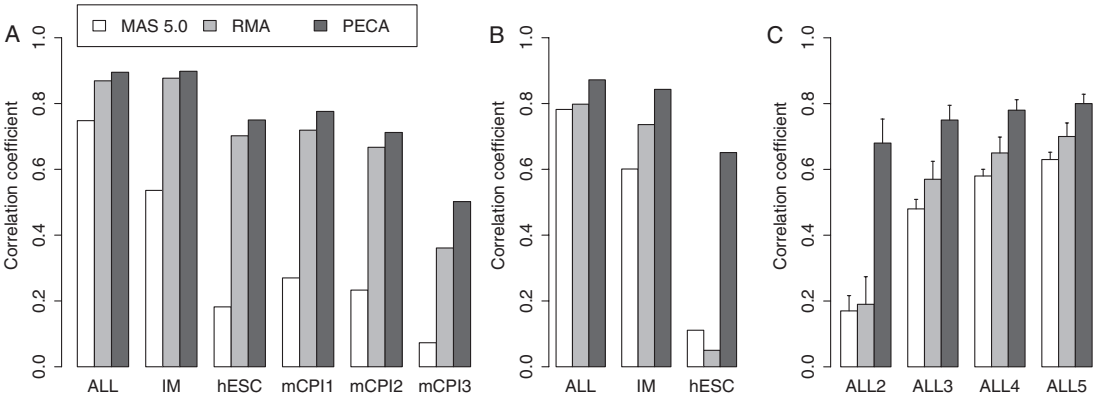


Figure 2. Observed correlations between the expression changes across different arrays as assessed with (**A**) signal log-ratios and (**B** and **C**) Hedges' *g*-estimates. In the ALL array comparison between HG-U95Av2 and HG-U133A arrays, expression changes between two distinct leukemia subtypes (14 samples per group) were calculated. In addition to the whole ALL data set, Hedges' *g*-estimates were calculated for 100 randomly sampled subsets of sizes 2–5 (ALL2–ALL5). Graph C shows the average correlations calculated over these subsets along with their standard deviations. In the IM array comparison between HG-U95Av2 and HG-U133A arrays, expression changes were calculated between patients with dermatomyositis (five samples) and patients with other inflammatory myopathies (nine samples). In the hESC array comparison between HG-U133A and HG-U133Plus2.0 arrays, expression changes between two hESC cell lines (two samples per group) were estimated. In the mCPI array comparison between MG-U74Av2 and MOE430 2.0 arrays, signal log-ratio between an infected and a control sample (mCPI1), between a re-infected and a control sample (mCPI2), and between an infected and a re-infected sample (mCPI3) were calculated. In each two-group comparison, the PECA-estimates of expression changes were compared with the corresponding expression change estimates obtained with RMA- and MAS-based intensity values, which are widely used in microarray data analysis.

with MAS, 60% with RMA and 70% with PECA (Figure 3C). With Hedges' *g*-estimates, there were no common genes among the top 30 genes with either MAS or RMA, whereas the PECA-estimates resulted in ∼50% overlap of the genes (Figure 3F). Within the array generations, the proportion of common genes among the top 100 genes between RMA and PECA was typically ∼80%, while it was 50% or less between MAS and PECA and between MAS and RMA.

Effect of sample sizes on the meta-analysis performance

Figure 4 illustrates the consistency of the 100 most differentially expressed genes identified using 2–5 biological samples from the ALL data as compared with the genes identified from the whole ALL data. As expected, the agreement among the top genes increased when the number of samples increased. The overall agreement of the results obtained with RMA-intensity values was again substantially lower than the agreement of the PECA-based results. The meta-analysis based on the PECA-estimated Hedges' *g*-values was most stable. With two samples, there were on average over 55% of common genes when the PECA-based meta-analysis was applied but only 35% with the RMA-based meta-analysis. When an individual data set of size 2 was considered, the stability of both approaches was reduced as compared with the meta-analysis. In particular, with the RMA-based analysis, the agreement decreased from 35% with the meta-analysis to ∼15% with an individual data set. However, even the meta-analysis could not raise the stability of the RMA-based estimates to the same level as the PECA-estimates. To obtain an agreement of over 50% of genes, the RMA-based meta-analysis typically required four samples, whereas only two samples were needed with the PECA-estimates, even when an individual data set was analyzed.

DISCUSSION

We have introduced a meta-analytic approach, which considers the latest probe-level information when combining the results of multiple Affymetrix microarray studies. We first showed that alternative probe sets provide a good option as compared with the manufacturer-defined probe sets when arrays with different probe collections are compared. Using these alternative sets, we then demonstrated that the comparability of expression changes across different array generations can be considerably improved with PECA-estimation as compared with the estimation based on RMA- or MAS-based summary intensities, especially when the sample sizes are small. The key finding was that by using the PECA-estimates one can more effectively combine the results of individual Affymetrix studies in the context of meta-analysis. In particular, we showed that the consistency of the differentially expressed genes can be improved by integrating PECA-based expression changes across studies. Taken together, these results suggest that available Affymetrix microarray studies of the particular condition can be effectively exploited when analyzing new experiments.

Conventionally, the probe-level expression data are summarized into simple numerical estimates of probe set-level gene expression. A major drawback of this approach is that a substantial amount of probe-level information is discarded. This issue has only lately become a focus of interest. It has been shown that by using probe-level expression data when identifying differentially expressed genes the quality of the resulting gene lists can be improved: Lemon *et al*. (25) and Master *et al*. (26) based their methods on probe-level *t*-tests; Barrera *et al*. (27) applied two-way ANOVA methods to probe-level data; and Chen *et al*. (28) measured probe-level differences in percentiles of ranks. The MAS software also

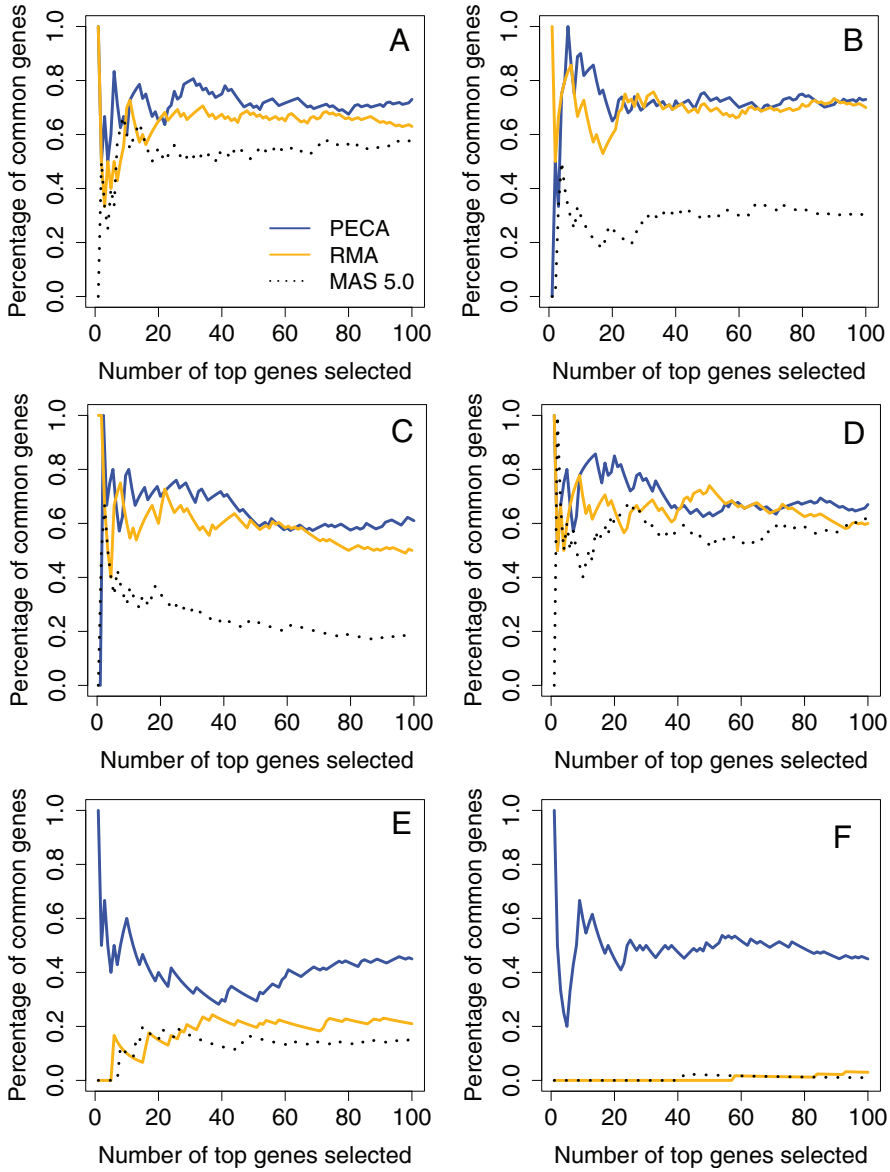


Figure 3. Agreement of differentially expressed genes between technical replicates. The proportion of overlapping genes in the two top *N* lists is plotted as a function of the list size. The genes were ranked with signal log-ratios in (**A**) ALL data (14 samples per group), (**B**) IM data (5 and 9 samples in the groups) and (**C**) hESC data (2 samples per group), and with Hedges' *g*-estimates in (**D**) ALL data, (**E**) IM data and (**F**) hESC data. The observed peaks at the beginning of the curves arise from a single shared top one gene in the two lists.

uses probe-level information in determining differential expression but the algorithm is restricted to comparisons between two arrays only. Our proposed PECA-method can be considered as a generalization of the MAS algorithm and the approaches of (25,26). The method can be used with any number of arrays, and in addition to *t*-statistic, it can improve other measures as well, especially when there are only few samples in the data (see Figure 2). Moreover, the computational burden of PECA is approximately the same as that of RMA- or MAS-normalizations.

We have also carried out an additional study in the Affymetrix spike-in data, where we showed that the

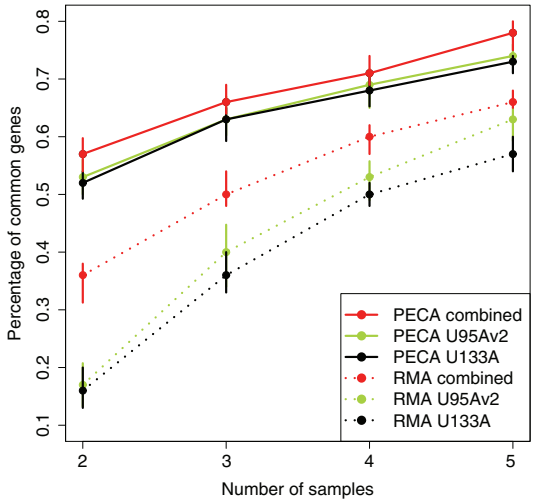


Figure 4. Agreement of differentially expressed genes using Hedges' *g* in the ALL data as the number of samples was varied. The performance was measured by calculating the proportion of common genes among the top 100 genes obtained with the whole data and with randomly selected smaller subsets of sizes 2–5. The results are presented as median percentage over 100 subsets (points) along with the interquartile ranges (error bars). RMA-based (dotted lines) and PECA-based (solid lines) estimation was used with the individual HG-U95Av2 (green) and HG-U133A (black) data sets and with the meta-analysis approaches (red). The RMA-based meta-analysis (RMA combined) represents the meta-analytic approach that has previously been proposed for microarray data (23).

PECA-estimated signal log-ratios and Hedges' *g*-values outperformed the corresponding values calculated from the RMA-normalized intensity values, especially when the sample size was small (see Supplementary Figure 3). In the context of microarray analysis, a common approach to overcome the problem of small sample sizes is to use a modified version of the ordinary *t*-statistic (29). Therefore, we evaluated its performance as well. In general, the PECA-estimated Hedges' *g* performed at least as well as the RMA-based modified *t*-statistic. In particular, with sample sizes 2 and 3, it yielded clearly better AUC-values and the PECA-estimated modified *t*-statistic could not improve its performance further. Although in this study we concentrated on the simple two-group comparisons only, it is possible to generalize the PECA-type analysis to situations, where there are more than two groups to be compared.

In a previous study, Hwang *et al*. (7) suggested that probe filtering could markedly improve the reproducibility of the top ranked genes as assessed with the two-sample *t*-test with unequal variances. After filtering the probe sets according to sequence similarity, they identified 30–40% common genes among the top 20 genes and ~25% common genes among the top 100 genes in the IM data. In our analysis with the PECA-estimated Hedges' *g*, the percentage of commonly identified genes was 40–60% among the top 20 genes and ~45% among the top 100 genes in the same data (see Figure 3E). In general, the percentage of common genes with PECA-estimates was over 40% even when there were only two samples in both groups. With the largest ALL data set, the percentage of common genes increased to 60–80%. These enhanced results clearly demonstrate the importance of the probe-level information in increasing the comparability between array generations. Similar approach could also be used to improve the agreement across different platforms (30).

Similar to (7), we aligned the probes to mRNA sequences with BLAT, which uses heuristics to speed up the search. To evaluate the accuracy of the BLAT search, we aligned the probes of the HG-U133A array also with the Bioconductor matchprobes package, which is based on exact string matching methods. The results obtained with BLAT and matchprobes were virtually the same (BLAT missed 52 of the 241 898 unique probes). The most essential difference between the two methods was in computation time. With an ordinary desktop PC, it took several days to align the HG-U133A probes against human mRNA sequences in Entrez using matchprobes, whereas BLAT made it in hours.

According to our results, the benefits gained from probe verification and alternative mappings are largest when arrays with different probe collections are compared, as in the mCPI, ALL and IM array comparisons (see Figure 1). Although the best match pairs of the original and verified probe sets performed similarly, they rely extensively on manufacturer annotations, including potentially erroneous probes. The alternative probe sets, on the contrary, are expected to refine as the public transcript databases grow in size and improve in accuracy. In the hESC array comparison, correlations between alternative probe sets were somewhat lower than correlations between best match probe sets. This was due to the fact that the original probe sets contained exactly the same probe sequences on both arrays, whereas the alternative probe sets on the HG-U133Plus2.0 array contained also probes that were not included in the HG-U133A array. Also in this case, however, the biological relevance of the alternative probe sets may be higher, since the original probe sets with identical probes would correlate highly even if they were erroneous in biological sense.

Meta-analysis has traditionally been used in medical and social sciences to combine results of different studies (21). Only recently, meta-analysis has also been applied to microarray experiments. Rhodes *et al*. (31) computed gene-specific *P*-values separately for each study and combined them using the Fisher statistic. Choi *et al*. (23) and Stevens and Doerge (32), on the other hand, combined the actual expression data by employing fixed effects and random effects models. In general, a random effects model is more reasonable than a fixed effects model because microarray studies are typically heterogeneous due to, for example, biological variation and differences between experimental methods. However, with only two studies to be combined, which is a typical case with microarrays, we based our integration method on a fixed effects model (33). An analogous approach can be used in the context of a random effects model when there are more studies to be combined.

We showed that the meta-analysis based on the PECA-estimated Hedges' *g*-values was more stable than the Choi *et al*. (23) meta-analysis based on the RMA-estimated summary intensities (see Figure 4). The stability of the methods was evaluated in terms of overlapping top genes obtained when using the whole ALL data set or random subsamples from it. It was assumed that the whole data set provides a

plausible approximation for the true ranking of the genes (24). The spike-in results supported this assumption (see Supplementary Figure 3). Because the reference ranking in the ALL data set was constructed from the same set as the subsamples, the overlap of the top genes might be overestimated with large subsample sizes. As we were interested in the performance of the methods with the smallest sample sizes 2–5, however, such procedure gives valuable information on the stability of the methods. While in this study it was beneficial to have the same samples hybridized to both arrays, the real benefits of the proposed meta-analytic procedure come from combining studies with diverse biological samples.

Previous meta-analysis studies on microarray data have not paid much attention to the quality of the effect size estimates (23,32). With small sample sizes, especially the Hedges' *g*-estimates are prone to unpredictable changes, since gene-specific variability can easily be underestimated resulting in large statistics' values due to chance alone. As only few replications are performed in most microarray experiments, it is critical to improve the effect size estimation with small sample sizes. The general idea of improving the reliability of the microarray results by pooling together results from existing studies is feasible only if the data are properly pre-processed. As probe verification is increasingly used in pre-processing of microarray data or for confirming the final results of a microarray study, it is natural to combine it with other probe-level analysis methods. We demonstrated that summarizing the expression changes over the verified probes only consistently helps in integrating data across studies made with different Affymetrix generations in the same laboratory. The biological findings from the hESC and mCPI data sets are published elsewhere [(34), (Kyläniemi, M., Haveri, A., Vuola, J., Puelakkainen, M. and Lahesman, R., unpublished data)].

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Miina Miller, the Finnish DNA Microarray Centre, Turku Centre for Biotechnology, for technical assistance. The work was supported by the Academy of Finland (grant 203632), the National Technology Agency, Turku University Hospital Research Fund and the Graduate School in Computational Biology, Bioinformatics, and Biometry (ComBi). Funding to pay the Open Access publication charges for this article was provided by the Academy of Finland.

Conflict of interest statement. None declared.

REFERENCES

1. Kothapalli,R., Yoder,S.J., Mane,S. and Loughran,T.P.Jr (2002) Microarray results: how accurate are they? *BMC Bioinformatics*, **3**, 22.
2. Kuo,W.P., Jenssen,T., Butte,A.J., Ohno-Machado,L. and Kohane,I.S. (2002) Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics*, **18**, 405–412.
3. Irizarry,R.A., Warren,D., Spencer,F., Kim,I.F., Biswal,S., Frank,B.C., Gabrielson,E., Garcia,J.G.N., Geoghegan,J., Germino,G. *et al.* (2005) Multiple-laboratory comparison of microarray platforms. *Nature Methods*, **2**, 345–349.
4. Bammler,T., Beyer,R.P., Bhattacharya,S., Boorman,G.A., Boyles,A., Bradford,B.U., Bumgarner,R.E., Bushel,P.R., Chaturvedi,K., Choi,D. *et al.* (2005) Standardizing global gene expression analysis between laboratories and across platforms. *Nature Methods*, **2**, 351–356.
5. Larkin,J.E., Frank,B.C., Gavras,H., Sultana,R. and Quackenbush,J. (2005) Independence and reproducibility across microarray platforms. *Nature Methods*, **2**, 337–343.
6. Gautier,L., Moller,M., Friis-Hansen,L. and Knudsen,S. (2004) Alternative mapping of probes to genes for Affymetrix chips. *BMC Bioinformatics*, **5**, 111.
7. Hwang,K.B., Kong,S.W., Greenberg,S.A. and Park,P.J. (2004) Combining gene expression data from different generations of oligonucleotide arrays. *BMC Bioinformatics*, **5**, 159.
8. Mecham,B.H., Klus,G.T., Strovel,J., Augustus,M., Byrne,D., Bozso,P., Wetmore,D.Z., Mariani,T.J., Kohane,I.S. and Szallasi,Z. (2004) Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Res.*, **32**, e74.
9. Zhang,J., Finney,R.P., Clifford,R.J., Derr,L.K. and Buetow,K.H. (2005) Detecting false expression signals in high-density oligonucleotide arrays by an *in silico* approach. *Genomics*, **85**, 297–308.
10. Nimgaonkar,A., Sanoudou,D., Butte,A.J., Haslett,J.N., Kunkel,L.M., Beggs,A.H. and Kohane,I.S. (2003) Reproducibility of gene expression across generations of Affymetrix microarrays. *BMC Bioinformatics*, **4**, 27.
11. Inzunza,J., Gertow,K., Stromberg,M.A., Matilainen,E., Blennow,E., Skottman,H., Wolbank,S., Ahrlund-Richter,L. and Hovatta,O. (2005) Derivation of human embryonic stem cell lines in serum replacement medium using postnatal human fibroblasts as feeder cells. *Stem Cells*, **23**, 544–549.
12. Penttilä,J.M., Anttila,M., Puolakkainen,M., Laurila,A., Varkila,K., Sarvas,M., Mäkelä,P.H. and Rautonen,N. (1998) Logical immune responses to *Chlamydia Pneumoniae* in the lungs of BALB/c mice during primary infection and reinfection. *Infect. Immun.*, **6**, 5113–5118.
13. Yeoh,E.J., Ross,M.E., Shurtleff,S.A.Williams,W.K., Patel,D., Mahfouz,R., Behm,F.G., Raimondi,S.C., Relling,M.V., Patel,A., Cheng,C. *et al.* (2002) Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, **1**, 133–143.
14. Ross,M.E., Zhou,X., Song,G., Shurtleff,S.A., Girtman,K., Williams,W.K., Liu,H.C., Mahfouz,R., Raimondi,S.C., Lenny,N., Patel,A. and Downing,J.R. (2003) Classification of pediatric lymphoblastic leukemia by gene expression profiling. *Blood*, **102**, 2951–2959.
15. Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al.* (2004) Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
16. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al.* (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **33**, D39–D45.
17. Kent,W.J. (2002) BLAT—The BLAST-Like Alignment Tool. *Genome Res.*, **12**, 656–664.
18. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2005) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **33**, D54–D58.
19. Irizarry,R.A., Hobbs,B., Collin,F., Beazer-Barclay,Y.D., Antonellis,K.J., Scherf,U. and Speed,T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
20. Bolstad,B.M., Irizarry,R.A. and Åstrand, M, Speed,T. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.
21. Hedges,L.V. and Olkin,I. (1985) *Statistical Methods for Meta-analysis*. Academic Press, Orlando, FL.
22. Park,P., Cao,Y.A., Lee,S.Y., Kim,J., Chang,M.S., Hart,R. and Choi,S. (2004) Current issues for DNA microarrays: platform comparison, double

linear amplification, and universal RNA reference. *J. Biotechnol.*, **112**, 225–245.
23. Choi,J.K., Yu,U., Kim,S. and Yoo,O.J. (2003) Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, **19**, i84–i90.
24. Kim,R.D. and Park,P.J. (2004) Improving identification of differentially expressed genes in microarray studies using information from public databases. *Genome Biol.*, **5**, R70.
25. Lemon,W.J., Liyanarachchi,S. and You,M. (2003) A high performance test of differential gene expression for oligonucleotide arrays. *Genome Biol.*, **4**, R67.
26. Master,S.R., Stoddard,A.J., Bailey,L.C., Pan,T.C., Dugan,K.D. and Chodosh,L.A. (2005) Genomic analysis of early murine mammary gland development using novel probe-level algorithms. *Genome Biol.*, **6**, R20.
27. Barrera,L., Benner,C., Tao,Y.C., Winzeler,E. and Zhou,Y. (2004) Leveraging two-way probe-level block design for identifying differential gene expression with high-density oligonucleotide arrays. *BMC Bioinformatics*, **5**, 42.
28. Chen,D.T., Chen,J.J. and Soong,S.J. (2005) Probe rank approaches for gene selection in oligonucleotide arrays with a small number of replicates. *Bioinformatics*, **21**, 2861–2866.
29. Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, 3.
30. Tan,P.K., Downey,T.J., Spitznagel,E.L.Jr, Xu,P., Fu,D., Dimitrov,D.S., Lempicki,R.A., Raaka,B.M. and Cam,M.C. (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.*, **31**, 5676–5684.
31. Rhodes,D.R., Barrette,T.R., Rubin,M.A., Ghosh,D. and Chinnaiyan,A.M. (2002) Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res.*, **62**, 4427–4433.
32. Stevens,J.R. and Doerge,R.W. (2005) Combining Affymetrix microarray results. *BMC Bioinformatics*, **6**, 57.
33. Garrett-Mayer,E., Parmigiani,G., Zhong,X., Cope,L. and Gabrielson,E. Cross-study validation and combined analysis of gene expression microarray data. *Johns Hopkins University, Dept. of Biostatistics Working Papers*, **65**.
34. Skottman,H., Mikkola,M., Lundin,K., Olsson,C., Stromberg,A.M., Tuuri,T., Otonkoski,T., Hovatta,O. and Lahesmaa,R. (2005) Gene expression signatures of seven individual human embryonic stem cell lines. *Stem Cells*, **9**, 1343–1356.

II

Publication II

Leo Lahti, Laura L. Elo, Tero Aittokallio, and Samuel Kaski. Probabilistic analysis of probe reliability in differential gene expression studies with short oligonucleotide arrays. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(1):217–225, 2011.

Probabilistic Analysis of Probe Reliability in Differential Gene Expression Studies with Short Oligonucleotide Arrays

Leo Lahti, Laura L. Elo, Tero Aittokallio, and Samuel Kaski

Abstract—Probe defects are a major source of noise in gene expression studies. While existing approaches detect noisy probes based on external information such as genomic alignments, we introduce and validate a targeted probabilistic method for analyzing probe reliability directly from expression data and independently of the noise source. This provides insights into the various sources of probe-level noise and gives tools to guide probe design.

Index Terms—Applications, biology and genetics, parameter learning, probabilistic algorithms.

1 Introduction

Gene expression profiling is widely used to explore gene function in various biological conditions, and vast collections of microarray data are available in public repositories. These large-scale data sets contain valuable information of both biological and technical aspects of gene expression studies [1], [2], [3], [4]. However, gene expression data are notoriously noisy. A better understanding of the technical aspects of the measurement process could ultimately lead to enhanced measurement techniques and improved analytical procedures, providing more accurate biological results in future studies.

Short oligonucleotide arrays of Affymetrix [5] are one of the most widely used gene expression profiling platforms. These arrays utilize multiple (typically 10-20) 25-mer probes, the so-called probe set, to measure the expression level of each transcript target. The probes within an individual probe set are designed to target the same gene, and ideally they should detect the same gene expression signal. Use of several probes for each target leads to more robust estimates of transcript activity, but the reliability of individual probes is known to vary and may significantly affect the results of a microarray study [6]. For example, it has been noticed that a considerable number of probes on short oligonucleotide arrays do not uniquely match their intended targets [7], [8], [9]. Single-nucleotide polymorphisms, alternative splicing, and nonspecific hybridization add biological variation in the data [10], [11]. Other factors in the measurement process that cause probe-specific effects include RNA extraction and amplification, binding affinities, and experiment-specific variation [12], [13].

Many preprocessing algorithms utilize probe-specific parameters to obtain probeset-level summaries of gene expression. These include MBEI/dChip [14], RMA [15], gcRMA [16], FARMS [17], gMOS [18], and BGX [19]. Despite the importance of probe-specific effects in gene expression analysis and probe design [6], [20], the various sources of probe-level noise are still poorly understood. Only a few studies have systematically analyzed the factors affecting probe reliability. The existing approaches typically rely on external information such as genomic sequence data [8], [9], [11] or physical models [21], [22], [23], and cannot reveal probes that are less reliable due to so far unknown reasons.

We introduce and validate a targeted computational tool for probe reliability analysis. In contrast to previous probe quality studies, the proposed model is independent of external information or physical models. This can advance the understanding of the various factors that affect probe reliability. Our approach is closely related to preprocessing methods that utilize probe-specific parameters to obtain probeset-level summaries of gene expression. A key difference in our work is that we assign an explicit probabilistic measure of reliability to each probe and demonstrate how this information can be used to assess probe performance. Explicit estimates and analysis of probe-specific noise have been missing in preprocessing studies. The method is applied to gene expression data sets from two human genome arrays, HG-U95A/Av2 and HG-U133A, and the results are validated by comparisons to known probe-level error sources: errors in probe-genome alignment, interrogation position of a probe on the target sequence, GC-content, and the presence of SNPs in the target sequences of the probes. Implementation of the method is available in R/BioConductor[1] at http://bioconductor.org/packages/release/bioc/html/RPA.html.

- *L. Lahti and S. Kaski are with the Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University School of Science and Technology, PO Box 15400, FI-00076 Aalto, Finland. E-mail: leo.lahti@iki.fi, samuel.kaski@tkk.fi.*
- *L.L. Elo and T. Aittokallio are with the Department of Mathematics, University of Turku, FI-20014 Turku, Finland, and the Turku Centre for Biotechnology, University of Turku, PO Box 123, FI-20521 Turku, Finland. E-mail: {laliel, teanai}@utu.fi.*

Manuscript received 14 Apr. 2008; revised 29 Jan. 2009; accepted 13 Apr. 2009; published online 22 Apr. 2009.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2008-0-0072. Digital Object Identifier no. 10.1109/TCBB.2009.38.

1. http://www.r-project.org/.

2 Modeling of Probe Reliability

The reliability of a probe is ultimately determined by its ability to measure the expression level of the target transcript. As the true expression level is unknown in most practical situations, the collection of probes measuring the same transcript can provide the ground truth for assessing probe performance (See Supplementary Figure 1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2009.38.). Our model captures the most coherent signal of the probe set, and the reliability of individual probes is estimated with respect to this signal across a large number of arrays. We provide an explicit probabilistic model for probe-level observations, and derive the posterior distribution for the model parameters describing probe reliability and differential gene expression. While probe-level preprocessing algorithms aim at summarizing probe-level measurements [14], [15], [17], [18], [19], we have specifically targeted a more detailed analysis of probe reliability. This avoids certain problems encountered in the preprocessing context as discussed in the next section.

2.1 Model Assumptions

Our approach is based on a Gaussian model for probe effects. This is a reasonable starting point for modeling heterogeneous and partially unknown sources of probe-level noise. The feasibility of related models has already been demonstrated in the preprocessing context [15], [17]. In a nutshell, we assume normally distributed probe effects, and identify probe reliability with its variance over a large number of arrays. In contrast to many probe-level preprocessing methods, where the mean is the important quantity, we use probe-level observations of differential expression. Then the mean cancels out, and the model can focus on estimating the variances (see Section 3 for details).

Variance reflects the noise level of the probe and is the main focus in our analysis. This is different from probe-level preprocessing methods that focus on estimating probe affinities, corresponding to the mean parameter of the Gaussian noise model. For example, the probe-specific parameters in MBEI [14] and RMA [15] preprocessing models describe probe affinities. These are constant shifting factors and as such not informative of probe reliability. Moreover, unidentifiability of probe affinities is a known problem in preprocessing studies [15], [24]. The recently suggested FARMS preprocessing algorithm [17] has a more complex model for probe effects than RMA and contains implicitly a similar probe-specific variance parameter as our model. However, FARMS does not provide explicit estimates of the probe-related parameters and is, therefore, not applicable to probe reliability analysis.

We avoid the modeling of unidentifiable probe affinities by using probe-level observations of differential gene expression. Probe effects are captured in a single probe-specific variance parameter in the resulting model. The number of probe-related parameters in the model is halved, and faster and more robust inferences concerning the parameters of interest can be obtained. Use of a single parameter for probe effects also leads to more straightforward interpretations of probe reliability. Cancelation of the probe affinity parameters in our analysis can partly explain the previous observations that calculating differential expression at probe-level improves the analysis of differential gene expression [25], [26]. However, these methods differ from our approach in that they are nonprobabilistic preprocessing methods that do not aim at quantifying the uncertainty in the probes.

TABLE 1
Gene Expression Data Sets in This Study

Name	Platform	Arrays	Author
ALL-95Av2	HG-U95Av2	37	Yeoh et al. (2002)
GEA-95A	HG-U95A	85	Su et al. (2002)
SPIKE-95Av2	HG-U95Av2	59	Affymetrix
ALL-133A	HG-U133A	37	Ross et al. (2003)
GEA-133A	HG-U133A	158	Su et al. (2004)
SPIKE-133A	HG-U133A	42	Affymetrix

2.2 Comparison to Known Error Sources

The model is applied to six publicly available gene expression data sets, including four large-scale studies on human samples [27], [28], [29], [30], referred to as ALL and GEA data sets, and two spike-in data sets from Affymetrix (www.affymetrix.com), referred to as SPIKE data sets (Table 1). The data sets have been measured using two popular human genome arrays, HG-U95A/Av2 and HG-U133A. To validate our model and to analyze probe reliability on these arrays, we test the overrepresentation of the following probe-level error sources among the least reliable probes predicted by our model.

2.2.1 Probe-Genome Alignments

Ideally, each probe has a unique sequence match to its target gene. In practice, a number of probes do not uniquely match their intended mRNA target. Filtering of probes with erroneous genome alignments has previously been shown to improve the accuracy and comparability of microarray results [8], [9], [11], [26], [31]. A good model for estimating probe reliability should detect such erroneous probes.

2.2.2 Interrogation Position on the Target Sequence

RNA degradation, typically starting from the $5'$ end of the transcript, has been reported to affect the results in microarray studies [32], [33]. Hence, the binding location of the probe on the target sequence, i.e., its *interrogation position*, is likely to affect probe reliability.

2.2.3 GC-Content

Various hybridization effects that are based on the nucleotide content of the probes have been reported [21], [22], [23], [34]. For example, the G/C nucleotides have a higher binding affinity since G-C pairs form three hydrogen bonds whereas the A-T pairs form two. Therefore, the GC-content of a probe is expected to affect its reliability.

2.2.4 SNP Associations

Probes that target sequences with common single-nucleotide polymorphims (SNPs) can produce misleading results in microarray studies [10], [35], [36]. Each probe can measure accurately at most one of the polymorphic target sequences and, therefore, gene expression differences between two individuals can be observed in some probes due to sequence polymorphism rather than real expression changes. This would add noise to microarray data. It is expected that SNPs located in the central region

of the target sequence will have a greater influence on probe reliability than other SNPs due to a larger impact on probe affinity [21], [37].

2.3 Connection to Preprocessing

The reliability of a probe is ultimately measured by its ability to capture the real underlying gene expression signal. This is unknown in most practical situations, however, and needs to be estimated from the probe-level observations. Probe reliability estimates are sensible only if the true signal is estimated accurately in our model. To guarantee this, the performance of the proposed model in estimating relative gene expression changes was compared to four alternative approaches: MAS5.0 (www.affymetrix.com) and RMA [15] are among the most widely applied methods for assessing probe set-level signals (which are then used to calculate the expression changes); FARMS [17] represents the previously introduced probe-level models; and PECA [38] shares the idea of directly utilizing probe-level expression changes. Note that the other methods do not provide explicit estimates of probe reliability, while our method provides only estimates of relative gene expression changes. A general difference between preprocessing algorithms and our method is that preprocessing methods have been designed to summarize probe-level information, whereas our model is specifically targeted at estimating certain probe-specific effects that are then used to analyze probe reliability.

3 Methods

3.1 Probabilistic Model

In the following, we describe a probabilistic model for probe reliability and differential gene expression. In the calculations, we use the logarithmized perfect match (PM) intensities of the Affymetrix arrays, and investigate each probe set separately. Affymetrix arrays also contain so-called mismatch (MM) probes that have an altered nucleotide in the middle (13th) position of the probe. These were originally designed to measure cross-hybridization from unrelated sequences. Some widely used preprocessing algorithms, such as RMA, ignore the MM probes due to the lack of efficient models for utilizing this information [15].

3.1.1 Conditional Likelihood for the Observations

Let us consider a probe set targeted at measuring the expression level of target transcript g. We model probe-level observations as a sum of the true expression signal that is common for all probes, and probe-specific Gaussian noise. A probe-level observation for probe j on array i can then be written as $s_{ij} = g_i + \mu_j + \varepsilon_{ij}$. The mean parameter μ_j describes the systematic probe affinity effect, and the stochastic noise component is distributed as $\varepsilon_{ij} \sim N(0, \tau_j^2)$.

The variance parameters $\{\tau_j^2\}$ are of interest in probe reliability analysis. To focus on these parameters we take advantage of the fact that the unidentifiable probe affinity parameters $\{\mu_j\}$ cancel out when the signal log-ratio between a randomly selected "control" array and the remaining arrays is computed for each probe. The differential expression signal between arrays $t = \{1, \ldots, T\}$ and the control array c for probe j is then $m_{tj} = s_{tj} - s_{cj} = g_t - g_c + \varepsilon_{tj} - \varepsilon_{cj} = d_t + \varepsilon_{tj} - \varepsilon_{cj}$. Using vector notation, the differential gene expression profile of probe j across the arrays $\{t\}$ is now $\mathbf{m}_j = \mathbf{d} + \varepsilon_j$, where the two noise terms have been combined into a single variable ε_j. Note that the control-related noise ε_{cj} is constant across the comparisons whereas the second noise component ε_{tj} depends on the array t.

To identify the probe-specific variance parameter, shared by the two noise components in ε_j for each probe j, we consider the control-related noise ε_{cj} a hidden variable in our model. This can be marginalized out by assuming that the probe-level observations $\mathbf{m}_j$ of the true underlying signal $\mathbf{d}$ are independent given the model parameters. Let us also denote the collection of probe-level signals of a probe set by $\mathbf{m} = \{\mathbf{m}_j\}$. The likelihood for the observations is then

$$P(\mathbf{m}|\mathbf{d}, \boldsymbol{\tau}^2) = \prod_{tj} \int N\left(m_{tj}|d_t - \varepsilon_{cj}, \tau_j^2\right) N\left(\varepsilon_{cj}|0, \tau_j^2\right) d\varepsilon_{cj}$$

$$\sim \prod_j \left(2\pi\tau_j^2\right)^{-\frac{T}{2}} exp\left(-\frac{\sum_t (m_{tj} - d_t)^2 - \frac{\left[\sum_t (m_{tj} - d_t)\right]^2}{T+1}}{2\tau_j^2}\right). \quad (1)$$

3.1.2 Posterior Distribution of the Model Parameters

The posterior density for the model parameters is computed from the conditional likelihood of the data (2) and the prior according to Bayes rule:

$$P(\mathbf{d}, \boldsymbol{\tau}^2|\mathbf{m}) \sim P(\mathbf{m}|\mathbf{d}, \boldsymbol{\tau}^2) P(\mathbf{d}, \boldsymbol{\tau}^2). \quad (2)$$

We use a noninformative prior for $\mathbf{d}$, and conjugate priors for the variance parameters in $\boldsymbol{\tau}^2$ (inverse Gamma distribution, see [39]). Using a standard assumption that $\mathbf{d}$ and $\boldsymbol{\tau}^2$ are independent with $P(\mathbf{d}|\boldsymbol{\tau}^2) \sim 1$, the prior takes the form $P(\mathbf{d}, \boldsymbol{\tau}^2) \sim \prod_j invgam(\tau_j^2; \alpha_j, \beta_j)$, where α_j and β_j are the parameters of the inverse Gamma distribution. These parameters are probe-specific and allow incorporation of prior information about probe reliability into the analysis.

The final model for probe intensities is hence described by two sets of parameters; the vector of underlying differential gene expression signals $\mathbf{d} = [d_1 \ldots d_T]$, and the probe-specific variance parameters $\boldsymbol{\tau}^2 = [\tau_1^2 \ldots \tau_J^2]$. High variance τ_j^2 would indicate that the probe-level observation $\mathbf{m}_j$ is strongly deviated from the estimated true signal $\mathbf{d}$. The Bayesian formulation quantifies the uncertainty in the model parameters and allows incorporation of prior information about probe reliability into the analysis. We refer to this procedure as *Robust Probabilistic Averaging (RPA)*.

3.1.3 Implementation

In this paper, we use the posterior mode as a point estimate for the model parameters. This is searched for by iteratively optimizing $\mathbf{d}$ and $\boldsymbol{\tau}^2$ in (3). The model is initialized to give equal prior weight for each probe by setting $\tau_j^2 = 1$ for each probe j. A mode for $\mathbf{d}$, given $\boldsymbol{\tau}^2$, is searched for by a standard quasi-Newton optimization method [40]. The variance parameters τ_j^2 follow an inverse Gamma distribution with parameters $\hat{\alpha}_j = \alpha_j + \frac{T}{2}$ and

$$\hat{\beta}_j = \beta_j + \frac{1}{2}\left(\sum_t (m_{tj} - d_t)^2 - \frac{\left(\sum_t (m_{tj} - d_t)\right)^2}{T+1}\right)$$

given $\mathbf{d}$. The mode is then given by $\tau_{j,new}^2 := \hat{\beta}_j / (\hat{\alpha}_j + 1)$. We use noninformative priors with $\alpha_j = \beta_j = 10^{-5}$.

3.2 Data

Only the common probe sets of the HG-U95A and HG-U95Av2 platforms were used, referred to as HG-U95A/Av2. Probe intensities were quantile-normalized, and the AFFX control sets excluded before the analysis.

3.2.1 Leukemia Data (ALL)

The public ALL data sets from the microarray studies of Ross et al. [27] and Yeoh et al. [30] contain expression data from patients with various leukemia subtypes. A total of 360 patient samples have been hybridized to HG-U95Av2 arrays and 132 of the same samples are additionally hybridized to HG-U133A arrays. For our analyses, we selected 37 samples that were hybridized to both array types and represent homogeneous patient groups with five distinct leukemia subtypes and control patients (Table 1). We refer to these two data sets as ALL-95Av2 and ALL-133A, respectively.

3.2.2 Gene Expression Atlas (GEA)

The gene expression atlases of Su et al. [28], [29] cover a diverse set of biological conditions measured on the human array platforms HG-U95A and HG-U133A (Table 1). We refer to these two data sets as GEA-95A and GEA-133A, respectively. Some samples in the HG-U95A data were ignored because no biological replicates were available.

3.2.3 Affymetrix Spike-In Data (SPIKE)

The Affymetrix HG-U95Av2 and HG-U133A spike-in data sets were downloaded from the Affymetrix web pages (www.affymetrix.com). We refer to these data sets as SPIKE-95Av2 (59 hybridizations) and SPIKE-133A (42 hybridizations). A total of 14 and 42 genes have been spiked-in at known concentrations on the HG-U95Av2 and HG-U133A arrays, respectively, and arrayed in a Latin Square format. Recently, it has been demonstrated that 22 additional probe sets in the SPIKE-133A data set should also be considered as spiked [41]. Accordingly, we utilized the extended set of 64 spiked probe sets when evaluating the performance of the different analysis approaches in the SPIKE-133A data.

3.2.4 Probe Sequence Data

Probe sequences and their bestmatch tables were downloaded from the Affymetrix web pages (www.affymetrix.com). Other array-wise information on probes and probe sets, including probe locations on the array, were acquired from the annotation data packages of the Bioconductor project [42]. Human genomic mRNA sequences were downloaded from Entrez Nucleotide [43] on 16 August 2006, excluding EST, STS, GSS, working draft and patents sequences, and sequences with a "XM_*" tag, as in [8], [26].

3.2.5 Probe-Genome Alignment

To identify probes having errors in the genomic alignment, all probes on the HG-U95A and HG-U133A arrays were aligned to the nucleotide sequences from Entrez Nucleotide, and assigned GeneIDs according to their matched sequence. Perfect matches of the probes to mRNA sequences were sought with BLAT v. 26 [44], following the same procedure as in [8], [26], but using updated genomic sequence data. The Entrez mRNA sequences were assigned to GeneID identifiers by using the "gene2accession" conversion file obtained from NCBI ftp server (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA, 10 August 2006). The percentage of probes with no GeneID match was 9.4 percent and 10.1 percent for the HG-U95A and HG-U133A arrays, respectively. Multiple GeneID matches were detected for 4.6 percent (HG-U95A) and 4.8 percent (HG-U133A) of the probes.

3.2.6 Single-Nucleotide Polymorphisms

Information about the probe-SNP associations was provided by the CustomCDF BioConductor package [10] that contains SNP mapping for the probes based on data from the dbSNP database [43]. The mappings have been used to investigate SNP effects in microarray data in recent studies [36], [45]. To focus on common SNPs, we considered only SNPs with a minimum population frequency of 5 percent.

4 Results

The RPA algorithm was applied on gene expression data sets from two commonly used microarray platforms to validate the model and to assess the differences between known probe-level noise sources. First, we compared probe reliability estimates to known probe-level error sources. Second, preprocessing comparisons were used to test the preprocessing performance of RPA and, importantly, to guarantee the validity of the probe reliability measures that depend on accurate estimation of the differential gene expression signal.

4.1 Comparison to Known Error Sources

4.1.1 Probe-Genome Alignment

Mistargeted probes that did not uniquely match the GeneID target of the probe set were significantly enriched ($p < 0.05$; hypergeometric test) among the least reliable 1 percent of the probes detected by our model (Fig. 1; Table S1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2009.38.). The mistargeted probes were 1.1-1.7 times more common in the HG-U95A/Av2 data sets than expected, and 2.2-3.1 times more common in HG-U133A. The enrichment of mistargeted probes was the highest for the probes that were consistently unreliable in the independent GEA and ALL data sets. On the HG-U133A array, mistargeted probes could explain 20.4 percent of the least reliable probes while the expected proportion was 6.7 percent. Consistently unreliable probes were detected by using the average rank of the probes obtained in the two experiments. Detection of probes having errors in their genomic alignment was expected because such probes do not necessarily have any correlation with the probe set-level signal. This supports the validity of our model.

4.1.2 Interrogation Position

The interrogation position of a probe on the target sequence was significantly associated with probe reliability ($p < 0.05$; χ^2-test). Probes closest to either end of the target sequence were enriched among the least reliable probes; the observed counts deviated 73-138 percent from the expectation, depending on the interrogation position (Fig. 2a; Table S1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2009. 38.). Enrichment of $5'$-binding probes was expected due to RNA degradation starting from this end of the transcript. Enrichment of 3′ probes is supported by previous findings of Dai et al., who noticed that 3′-focused probe sets have often increased noise levels

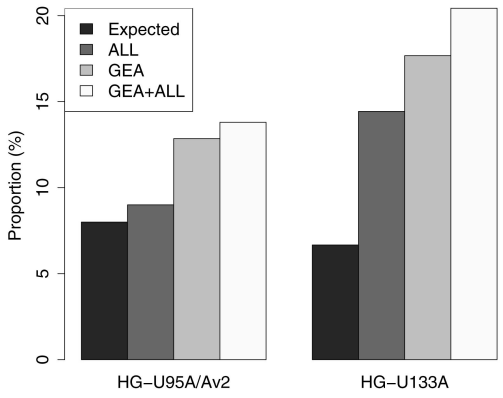


Fig. 1. Genomic alignment and probe reliability. Mistargeted probes that do not uniquely match the GeneID target of the probe set were enriched among the least reliable probes ($p < 0.05$; hypergeometric test). Black bars show the expected proportion of mistargeted probes, i.e., their proportion on the whole array. Gray bars show the proportion of mistargeted probes among the least reliable 1 percent of the probes detected by our model (dark: ALL; light: GEA; white: combined results).

[10]. Probes closer to the 3′ end detect, on average, a higher absolute signal. A higher signal is often associated with higher noise in microarray studies [46], which could explain our observation. Alternative transcription may also cause differences between 3′ probes and the other probes [47], [48].

4.1.3 GC-Content and Probe Reliability

C-rich probe sequences were enriched among the least reliable probes of our model in all data sets except ALL-95Av2 and GEA-95A (Fig. 2b; Table S1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2009.38.). The observed counts for the different GC contents deviated 39-132 percent from the expectation in the investigated data sets ($p < 0.05$; χ^2-test). To guarantee the assumptions of the χ^2-test, probes with most extreme G/C or A/T contents were combined in the test. One explanation for our observation is that high-affinity probes may have higher likelihood of cross-hybridization to nonspecific targets [21]. This would add noise to the probe-level signal.

4.1.4 Single-Nucleotide Polymorphisms

Probes whose target sequences have common SNPs were enriched among the least reliable probes on the HG-U133A platform and in the combined results from HG-U95A/Av2 platform (See Supplementary Figure 3; Table S1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2009.38.). In these data sets, the SNP-associated probes were 1.7-1.9 times more common among the least reliable probes than expected ($p < 0.05$; hypergeometric test). It is interesting to notice that the association between probe reliability and SNPs is observed only when information from the ALL-95Av2 and GEA-95A is combined; a similar observation was made with the GC-rich probes. A likely explanation is that the systematic effects from the SNP-associated, or GC-rich probes are more effectively observed when the data sets are combined and the data set specific noise cancels out. In general, the SNP-associated probes were less reliable than the other probes in all investigated data sets ($p < 0.05$; Wilcoxon test). As expected, probes having a single SNP in the central 13bp region of the 25-mer probe were less reliable than probes with a single SNP in either end of the target sequence on HG-U133A ($p < 0.05$; Wilcoxon test) but, interestingly, not on the HG-U95A/Av2 platform.

4.1.5 Relative Contribution of the Known Error Sources

Probes that are associated with the investigated noise sources had 7-39 percent increase in average variance, detected by RPA, in the studied data sets except ALL95-Av2 (Fig. 3). Mistargeted probes had the highest variances on HG-U133A, whereas probes with the most 5′/3′ interrogation positions

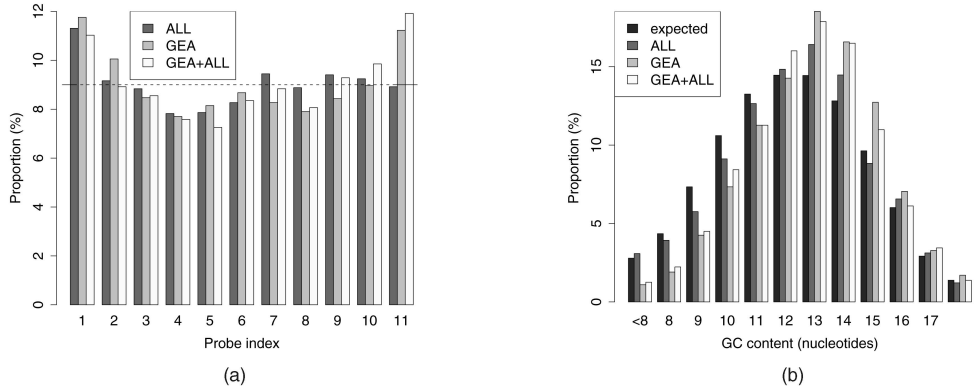


Fig. 2. Probe reliability versus interrogation position and GC-content on the HG-U133A platform. (a) Probes that bind to either the 5′ or the 3′ end of the target transcript were enriched among the least reliable (1 percent) probes ($p < 0.05$; χ^2-test). Probe index indicates the relative interrogation position of the probe on the target sequence, starting from the 5′ end of the transcript. The gray bars show the proportion for each interrogation position among the least reliable probes in the inspected data sets (dark: ALL; light: GEA; white: combined results). The expectation is illustrated by the dashed line. There are 11 probes per probe set on the HG-U133A arrays. (b) GC-rich probes were enriched among the least reliable (1 percent) probes ($p < 0.001$; χ^2-test). The GC-content of a probe is indicated by the number of G/C nucleotides on the 25-mer probes. Gray bars show the proportion of each GC-content among the least reliable probes (dark: ALL; light: GEA; white: combined results). Consistently less reliable probes (GEA+ALL) had the highest deviation from the expectation (black bars). To guarantee the assumptions of the χ^2-test, we combined probes with most extreme G/C or A/T contents for testing. Results for the HG-U95A/Av2 data sets are shown in Supplementary Fig. 2.

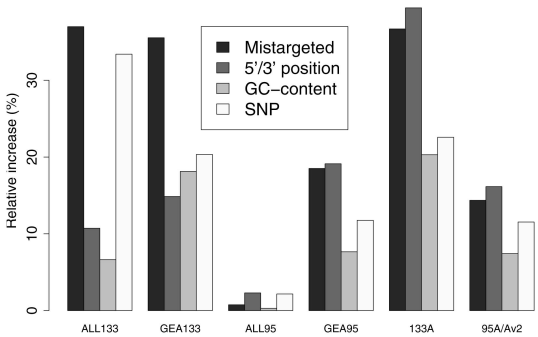


Fig. 3. Increase in the average variance of the probes associated with the investigated noise sources: mistargeted probes having errors in the genomic alignment, most $5'/3'$ probes of each probe set, GC-rich, and SNP-associated probes. The variances were estimated by RPA and describe the noise level of the probes. The results are shown for the individual ALL and GEA data sets, and for their combined results on both platforms (133A and 95A/Av2).

had the highest variances on HG-U95A/Av2. High GC-content led to a more moderate increase in probe-specific variance than the other investigated sources. However, GC-rich probes are more common (28-33 percent of the probes) than mistargeted probes (6-8 percent), probes with common SNPs (3-3.4 percent), or probes in the most $5'/3'$ positions of the target sequence (10-18 percent) and have, therefore, a remarkable contribution to the overall probe-level noise. Interestingly, many (35-60 percent) of the least reliable probes detected by RPA were not associated with the investigated sources, including many probes that have systematically low reliability in independent data sets.

4.2 General Observations of Probe Reliability

Examples of the least reliable probes in the GEA-95A data set are shown in Supplementary Figure 4, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2009.38. Comparison of the results from independent ALL and GEA data sets revealed many probes with consistently poor reliability, although the comparability of the results was affected by data set-specific effects: Spearman correlations of the probe-specific variances $\{\tau_j^2\}$ between the ALL and GEA data sets were 0.28 (HG-U95A/Av2) and 0.52 (HG-U133A). Surprisingly, the least reliable probes in the ALL data sets showed almost identical expression profiles (See Supplementary Figure 5, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2009.38), although they are located in independent probe sets and expected to capture uncorrelated signals. The noise probably originates in the biological samples that have been hybridized on both array types in the ALL-95Av2 and ALL-133A data sets. The specific source of this contamination remains unclear.

4.3 Preprocessing Comparisons

The validity of probe reliability estimates depends on accurate estimation of the probe set-level signal. We compared RPA to other preprocessing methods to test its preprocessing performance and to guarantee the validity of probe reliability estimates.

4.3.1 Spike-In Data

In spike-in data sets, the true expression changes are known and, hence, the different preprocessing approaches can be compared in terms of their receiver operating characteristics (ROC). RPA and PECA were more successful in detecting the spiked genes than MAS5.0 or RMA (Fig. 4). FARMS was found to outperform the other methods when a large number of genes are inspected. The good performance of FARMS in the spike-in data may, however, be favoured by the particular design of the spike-in experiments, in which the expression changes always occur in the same genes. This was supported by the observation that, unlike the other methods, FARMS produced nearly perfect ROC-curves even when replicated samples were compared with each other, although in these comparisons no changes should be detected and the gene rankings should be random (See Supplementary Figure 6, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2009.38.).

4.3.2 Technical Replicates

We also assessed the performance of the different preprocessing methods in real research settings using the ALL and GEA data sets. Since in these data sets the true expression changes were not known, the performance of the different methods was evaluated in terms of their consistency across replicated measurements for both genes and biological samples. Following the approach of Reverter et al. [49], we first measured the consistency of the expression changes within each data set (See Supplementary Figure 7, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2009.38.). Specifically, for each GeneID represented by at least two probe sets on an array, the average Pearson correlation of the expression profiles between all the matching probe sets was calculated. Based on our probe-genome alignments, there were 1,470 and 3,774 such GeneIDs on the HG-U95A/HG-U95Av2 and HG-U133A arrays, respectively. In each data set, RPA produced the highest correlations ($p < 0.05$; paired Wilcoxon test), and PECA and RMA also clearly outperformed not only MAS5.0 but, notably, also FARMS.

To further investigate the performance of the methods, we evaluated the consistency of the expression changes across the two separate data sets, ALL-95Av2 and ALL-U133A, in which the same biological samples have been hybridized (Fig. 5). The consistency was measured by the Pearson correlation between the pairs of arrays, to which the same sample was hybridized. This indicates the performance of the methods, as the technical replicates are assumed to produce effectively the same results on both array versions. The so-called "bestmatch" tables, provided by the array manufacturer (www.affymetrix.com), were utilized to combine the data across the arrays. The results from this analysis supported the earlier findings. In particular, RPA and PECA outperformed the other approaches; RMA performed better than MAS5.0 and FARMS; and MAS5.0 showed the poorest performance ($p < 0.05$; paired Wilcoxon test). Interestingly, the simple PECA yielded better consistency between the data sets than RPA ($p < 0.05$). While the main focus of this paper is in probe reliability analysis, the preprocessing comparisons confirmed that RPA compares favourably with the other methods in estimating differential gene expression. This guarantees the validity of probe reliability estimates in our model.

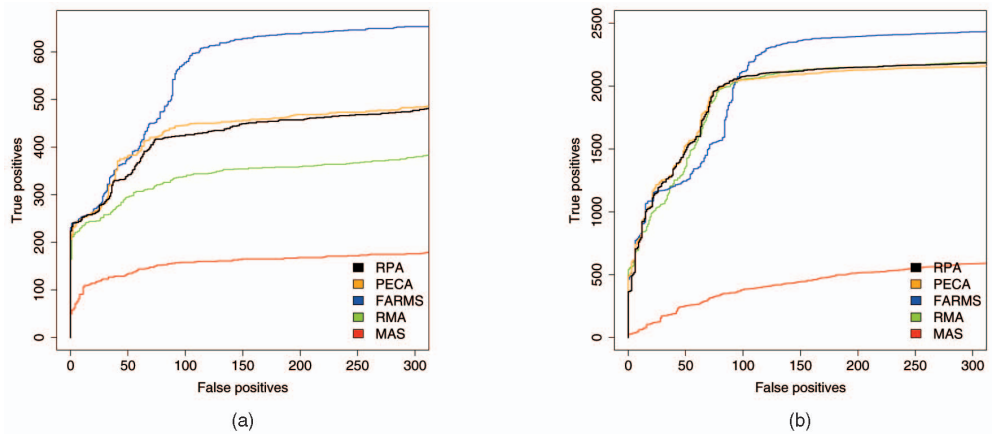


Fig. 4. Preprocessing performance for spike-in data. ROC curves for the various methods that were used to estimate the signal log-ratio: RPA, PECA, RMA, FARMS, and MAS for the two spike-in data sets ((a) Affymetrix HG-U95Av2 and (b) HG-U133A). For each curve, the results from the investigated spike-in samples within the data sets were pooled. The axes have been truncated to focus on the most relevant area. When comparing the curves, the one closest to the upper left corner shows the best performance.

5 Discussion

Previous probe-level models have focused on preprocessing of gene expression data, whereas we have specifically targeted a more detailed analysis of probe reliability. Enrichment of known probe-level error sources among the less reliable observed probes validates our model; many of the findings were explained by errors in genomic alignment, probe interrogation position, GC-content, or common SNPs. However, any single source of error seems to explain only a fraction of the probes that have consistently poor reliability in independent data sets. Therefore, methods that remove probe-level noise based on external information such as genomic alignments are likely to ignore a large number of the least reliable probes. For example, a probe set designed to measure a certain transcript may additionally detect unknown alternatively spliced transcripts which may have different expression patterns [12], or cross-hybridize with mRNAs having closely similar ($>18/25$ bp) but not perfectly matching sequences [11]. Various laboratory- and experiment-specific effects are also known to add experimental noise in microarray studies [12], [13]. The proposed model can detect poorly performing probes that are susceptible to noise from such sources.

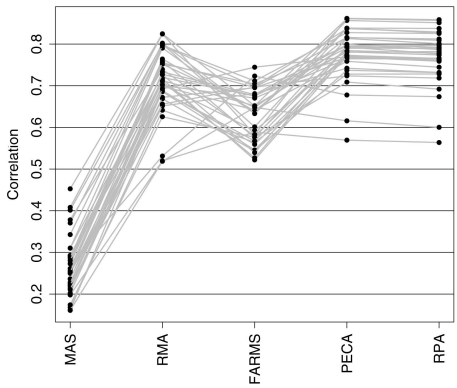


Fig. 5. Reproducibility of signal estimates in real data sets between the technical replicates, i.e., the best match probe sets between the HG-U95Av2 and HG-U133A platforms. The consistency was measured by the Pearson correlation between the pairs of arrays, to which the same sample was hybridized.

A Gaussian model for probe effects is a reasonable starting point for modeling heterogeneous and partially unknown sources of probe-level noise. The feasibility of similar models has already been demonstrated in the preprocessing context. For example, the RMA preprocessing algorithm [15] has a Gaussian model for probe effects with probe-specific mean (affinity) parameters and a shared variance parameter for the probes. We avoid the estimation of probe affinities and instead focus on estimating probe-specific variances. The recently suggested FARMS preprocessing algorithm [17] is closely related to our approach but has a more complex model for probe effects. The model can be written as $s_{ij} = z_i\lambda_j + \mu_j + \varepsilon_{ij}$. Here, z_i captures the underlying gene expression, and the model has three parameters $\{\lambda_j, \mu_j, \varepsilon_{ij}\}$ for each of the 10-20 probes in a probe set. In contrast, our model has a single variance parameter for each probe. The use of a more complex model in FARMS is justified as it aims at summarizing the absolute values of logarithmized PM intensities. This is a hard task since large systematic differences are known to exist between probes [14], [46]. We have shown that by computing differential gene expression at probe-level avoids the need to estimate unidentifiable probe affinity parameters. Use of a single parameter for probe effects leads to more straightforward interpretations about probe reliability and makes the model potentially less prone to overfitting. This is supported by the observation that RPA and PECA compared favourably with other preprocessing methods in the analysis of differential gene expression. The distinguishing feature of the two methods is that they compute differential gene expression at the probe-level. However, only the probabilistic RPA estimates probe reliability.

While for most probe sets, different preprocessing methods give largely consistent results, their differences can be especially large for probe sets containing several

inconsistent probe-level signals. The main contribution of the current study is to introduce and apply a probabilistic model with explicit modeling assumptions to analyze probe reliability on short oligonucleotide arrays. At the same time, the model provides a principled framework for incorporating prior information of the probes in differential gene expression analysis. This is a potential topic for future studies.

6 Conclusion

We have introduced a probabilistic framework for analyzing the reliability of individual probes directly from gene expression data, and validated the model using gene expression data sets from two popular human genome arrays. A major advantage of the proposed approach is its capability to detect unreliable probes independently of physical models or external, constantly updated information such as genomic sequence data. Probe reliability information can be useful in many applications, including evaluation of the end results of gene expression analysis, and recognition of potentially unknown probe-level error sources. It can be used to quantify the uncertainty in the measurements and in designing the probes, and is also utilized by our model to provide robust estimates of differential gene expression. A better understanding of the various probe-level error sources could advance probe design and contribute to reducing probe-related noise in the future generations of gene expression arrays.

Acknowledgments

The work was supported by the Academy of Finland, grants 207467 (LL, SK) and 203632 (LE, TA). SK was additionally funded by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, and LE by the Graduate School in Computational Biology, Bioinformatics, and Biometry (ComBi), and the Academy of Finland, grant 127575.

References

[1] C. Benedict, M. Geisler, J. Trygg, N. Huner, and V. Hurry, "Consensus by Democracy. Using Meta-Analyses of Microarray and Genomic Data to Model the Cold Acclimation Signaling Pathway in Arabidopsis," *Plant Physiology,* vol. 141, no. 4, pp. 1219-1232, 2006.

[2] Z. Hu, C. Fan, D.S. Oh, J. Marron, X. He, B.F. Qaqish, C. Livasy, L.A. Carey, E. Reynolds, L. Dressler, A. Nobel, J. Parker, M.G. Ewend, L.R. Sawyer, J. Wu, Y. Liu, R. Nanda, M. Tretiakova, A.R. Orrico, D. Dreher, J.P. Palazzo, L. Perreard, E. Nelson, M. Mone, H. Hansen, M. Mullins, J.F. Quackenbush, M.J. Ellis, O.I. Olopade, P.S. Bernard, and C.M. Perou, "The Molecular Portraits of Breast Tumors are Conserved Across Microarray Platforms," *BMC Genomics,* vol. 7, p. 96, 2006.

[3] S. Katz, R.A. Irizarry, X. Lin, M. Tripputi, and M.W. Porter, "A Summarization Approach for Affymetrix GeneChip Data Using a Reference Training Set from a Large, Biologically Diverse Database," *BMC Bioinformatics,* vol. 7, p. 464, 2006.

[4] S. Yoon, Y. Yang, J. Choi, and J. Seong, "Large Scale Data Mining Approach for Gene-Specific Standardization of Microarray Gene Expression Data," *Bioinformatics,* vol. 22, pp. 2898-2904, 2006.

[5] D. Lockhart, H. Dong, M. Byrne, M. Follettie, M. Gallo, M. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. Brown, "Expression Monitoring by Hybridization to High-Density Oligonucleotide Arrays," *Nature Biotechnology* vol. 14, pp. 1675-80, 1996.

[6] R.A. Irizarry, D. Warren, F. Spencer, I.F. Kim, S. Biswal, B.C. Frank, E. Gabrielson, J.G.N. Garcia, J. Geoghegan, G. Germino, C. Griffin, S.C. Hilmer, E. Hoffman, A.E. Jedlicka, E. Kawasaki, F. Martinez-Murillo, L. Morsberger, H. Lee, D. Petersen, J. Quackenbush, A. Scott, M. Wilson, Y. Yang, S.Q. Ye, and W. Yu, "Multiple-Laboratory Comparison of Microarray Platforms," *Nature Methods,* vol. 2, pp. 345-350, 2005.

[7] L. Gautier, M. Moller, L. Friis-Hansen, and S. Knudsen, "Alternative Mapping of Probes to Genes for Affymetrix Chips," *BMC Bioinformatics,* vol. 5, p. 111, 2004.

[8] K.-B. Hwang, S.W. Kong, S.A. Greenberg, and P.J. Park, "Combining Gene Expression Data from Different Generations of Oligonucleotide Arrays," *BMC Bioinformatics,* vol. 5, p. 159, 2004.

[9] B.H. Mecham, D.Z. Wetmore, Z. Szallasi, Y. Sadovsky, I. Kohane, and T.J. Mariani, "Increased Measurement Accuracy for Sequence-Verified Microarray Probes," *Physiological Genomics,* vol. 18, pp. 308-315, 2004.

[10] 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 Research,* vol. 33, p. e175, 2005.

[11] J. Zhang, R.P. Finney, R.J. Clifford, L.K. Derr, and K.H. Buetow, "Detecting False Expression Signals in High-Density Oligonucleotide Arrays by an In Silico Approach," *Genomics,* vol. 85, pp. 297-308, 2005.

[12] MAQC Consortium, "The Microarray Quality Control (MAQC) Project Shows Inter- and Intraplatform Reproducibility of Gene Expression Measurements," *Nature Biotechnology,* vol. 24, pp. 1151-1161, 2006.

[13] Y. Tu, G. Stolovitzky, and U. Klein, "Quantitative Noise Analysis for Gene Expression Microarray Experiments," *Proc. Nat'l Academy of Sciences USA,* vol. 99, pp. 14031-14036, 2002.

[14] C. Li and W.H. Wong, "Model-Based Analysis of Oligonucleotide Arrays: Expression Index Computation and Outlier Detection," *Proc. Nat'l Academy of Sciences USA,* vol. 98, pp. 31-36, 2001.

[15] R.A. Irizarry, B.M. Bolstad, F. Collin, L.M. Cope, B. Hobbs, and T.P. Speed, "Summaries of Affymetrix GeneChip Probe Level Data," *Nucleic Acids Research,* vol. 31, p. e15, 2003.

[16] Z. Wu and R. Irizarry, "Stochastic Models Inspired by Hybridization Theory for Short Oligonucleotide Arrays," *Proc. Eight Conf. Research in Computational Molecular Biology (RECOMB '04),* pp. 98-106, 2004.

[17] S. Hochreiter, D.-A. Clevert, and K. Obermayer, "A New Summarization Method for Affymetrix Probe Level Data," *Bioinformatics,* vol. 22, pp. 943-949, 2006.

[18] M. Milo, A. Fazeli, M. Niranjan, and N. Lawrence, "A Probabilistic Model for the Extraction of Expression Levels from Oligonucleotide Arrays," *Biochemical Soc. Trans.,* vol. 31, pp. 1510-1512, 2003.

[19] A.-M.K. Hein, S. Richardson, H.C. Causton, G.K. Ambler, and P.J. Green, "BGX: A Fully Bayesian Integrated Approach to the Analysis of Affymetrix GeneChip Data," *Biostatistics,* vol. 6, pp. 349-373, 2005.

[20] X. Li, Z. He, and J. Zhou, "Selection of Optimal Oligonucleotide Probes for Microarrays Using Multiple Criteria, Global Alignment and Parameter Estimation," *Nucleic Acids Research,* vol. 33, pp. 6114-6123, 2005.

[21] R. Mei, E. Hubbell, S. Bekiranov, M. Mittmann, F.C. Christians, M.-M. Shen, G. Lu, J. Fang, W.-M. Liu, T. Ryder, P. Kaplan, D. Kulp, and T.A. Webster, "Probe Selection for High-Density Oligonucleotide Arrays," *Proc. Nat'l Academy of Sciences USA,* vol. 100, pp. 11237-11242, 2003.

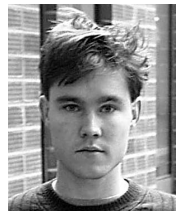
[22] F. Naef and M.O. Magnasco, "Solving the Riddle of the Bright Mismatches: Labeling and Effective Binding in Oligonucleotide Arrays," *Physical Rev. E,* vol. 68, 2003.

[23] C. Wu, R. Carta, and L. Zhang, "Sequence Dependence of Cross-Hybridization on Short Oligo Microarrays," *Nucleic Acids Research,* vol. 33, p. e84, 2005.

[24] A.L. Oberg, D.W. Mahoney, K.V. Ballman, and T.M. Therneau, "Joint Estimation of Calibration and Expression for High-Density Oligonucleotide Arrays," *Bioinformatics,* vol. 22, pp. 2381-2387, 2006.

[25] L. Zhang, L. Wang, A. Ravindranathan, and M. Miles, "A New Algorithm for Analysis of Oligonucleotide Arrays: Application to Expression Profiling in Mouse Brain Regions," *J. Molecular Biology,* vol. 317, pp. 225-235, 2002.

[26] L.L. Elo, L. Lahti, H. Skottman, M. Kyläniemi, R. Lahesmaa, and T. Aittokallio, "Integrating Probe-Level Expression Changes Across Generations of Affymetrix Arrays," *Nucleic Acids Research,* vol. 33, p. e193, 2005.
[27] M.E. Ross, X. Zhou, G. Song, S. Shurtleff, K. Girtman, W. Williams, H.-C. Liu, R. Mahfouz, S. Raimondi, N. Lenny, A. Patel, and J. Downing, "Classification of Pediatric Acute Lymphoblastic Leukemia by Gene Expression Profiling," *Blood,* vol. 102, pp. 2951-2959, 2003.
[28] A.I. Su, M.P. Cooke, K.A. Ching, Y. Hakak, J.R. Walker, T. Wiltshire, A.P. Orth, R.G. Vega, L.M. Sapinoso, A. Moqrich, A. Patapoutian, G.M. Hampton, P.G. Schultz, and J.B. Hogenesch, "Large-Scale Analysis of the Human and Mouse Transcriptomes," *Proc. Nat'l Academy of Sciences USA,* vol. 99, pp. 4465-4470, 2002.
[29] 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," *Proc. Nat'l Academy of Sciences USA,* vol. 101, pp. 6062-6067, 2004.
[30] E.-J. Yeoh et al., "Classification, Subtype Discovery, and Prediction of Outcome in Pediatric Acute Lymphoblastic Leukemia by Gene Expression Profiling," *Cancer Cell,* vol. 1, pp. 133-143, 2002.
[31] H. Yu, F. Wang, K. Tu, L. Xie, Y. Li, and Y. Li, "Transcript-Level Annotation of Affymetrix Probesets Improves the Interpretation of Gene Expression Data," *BMC Bioinformatics,* vol. 8, p. e194, 2007.
[32] H. Auer, S. Lyianarachchi, D. Newsom, M.I. Klisovic, G. Marcucci, and K. Kornacker, "Chipping Away at the Chip Bias: RNA Degradation in Microarray Analysis," *Nature Genetics,* vol. 35, pp. 292-293, 2003.
[33] L. Gautier, L. Cope, B.M. Bolstad, and R.A. Irizarry, "Affy-Analysis of Affymetrix Genechip Data at the Probe Level," *Bioinformatics,* vol. 20, pp. 307-315, 2004.
[34] C.J. Burden, Y. Pittelkow, and S.R. Wilson, "Adsorption Models of Hybridization and Post-Hybridization Behavior on Oligonucleotide Microarrays," *J. Physics: Condensed Matter,* vol. 18, pp. 5545-5565, 2006.
[35] A. Sequeira, F. Meng, B. Rollins, R. Myers, E. Jones, S. Watson, H. Akil, A. Schatzberg, J. Barchas, and W. Bunney, V. M.P., "Coding SNPs Included in Exon Arrays for the Study of Psychiatric Disorders," *Molecular Psychiatry,* vol. 13, pp. 363-365, 2008.
[36] E. Sliwerska, F. Meng, T. Speed, E. Jones, W. Bunney, H. Akil, S. Watson, and M. Burmeister, "SNPs on Chips: The Hidden Genetic Code in Expression Arrays," *Biological Psychiatry,* vol. 61, pp. 13-16, 2007.
[37] I. Lee, A.A. Dombkowski, and B.D. Athey, "Guidelines for Incorporating Non-Perfectly Matched Oligonucleotides into Target-Specific Hybridization Probes for a DNA Microarray," *Nucleic Acids Research,* vol. 32, pp. 681-690, 2004.
[38] L.L. Elo, M. Katajamaa, R. Lund, M. Oresic, R. Lahesmaa, and T. Aittokallio, "Improving Identification of Differentially Expressed Genes by Integrative Analysis of Affymetrix and Illumina Arrays," *OMICS: A J. Integrative Biology,* vol. 10, pp. 369-380, 2006.
[39] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin, *Bayesian Data Analysis,* second ed. Chapman & Hall/CRC, 2003.
[40] D. Goldfarb, "A Family of Variable-Metric Methods Derived by Variational Means," *Math. of Computation,* vol. 24, pp. 23-26, 1970.
[41] M. McGee and Z. Chen, "New Spiked-In Probe Sets for the Affymetrix HG-U133A Latin Square Experiment," COBRA Preprint Series, Article 5, 2006.
[42] R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F.L.C. Li, M. Maechler, A.J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J.Y.H. Yang, and J. Zhang, "Bioconductor: Open Software Development for Computational Biology and Bioinformatics," *Genome Biology,* vol. 5, p. R80, 2004.
[43] D.L. Wheeler, T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. DiCuccio, R. Edgar, S. Federhen, M. Feolo, L.Y. Geer, W. Helmberg, Y. Kapustin, O. Khovayko, D. Landsman, D.J. Lipman, T.L. Madden, D.R. Maglott, V. Miller, J. Ostell, K.D. Pruitt, G.D. Schuler, M. Shumway, E. Sequeira, S.T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R.L. Tatusov, T.A. Tatusova, L. Wagner, and E. Yaschenko, "Database Resources of the National Center for Biotechnology Information," *Nucleic Acids Research,* vol. 36, suppl. 1, pp. D13-D21, 2008.
[44] W.J. Kent, "BLAT-The BLAST-Like Alignment Tool," *Genome Research,* vol. 12, pp. 656-664, 2002.
[45] E.C. Rouchka, A.W. Phatak, and A.V. Singh, "Effect of Single Nucleotide Polymorphisms on Affymetrix Match-Mismatch Probe Pairs," *Bioinformation,* vol. 2, pp. 405-411, 2008.
[46] R.A. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf, and T.P. Speed, "Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data," *Biostatistics,* vol. 4, pp. 249-264, 2003.
[47] Y. Xing, K. Kapur, and W.H. Wong, "Probe Selection and Expression Index Computation of Affymetrix Exon Arrays," *PLoS ONE,* vol. 1, p. e88, 2006.
[48] J. Yan and T.G. Marr, "Computational Analysis of 3-Ends of ESTs Shows Four Classes of Alternative Polyadenylation in Human, Mouse, and Rat," *Genome Research,* vol. 15, pp. 369-375, 2005.
[49] A. Reverter, W. Barris, S. McWilliam, K. Byrne, Y. Wang, S. Tan, N. Hudson, and B. Dalrymple, "Validation of Alternative Methods of Data Normalization in Gene Co-Expression Studies," *Bioinformatics,* vol. 21, pp. 1112-1120, 2005.



Leo Lahti received the MSc (Tech) degree from Helsinki University of Technology, Espoo, Finland, in 2003. He is currently a postgraduate researcher at the Department of Information and Computer Science, Helsinki University of Technology, focusing on development and application of machine learning methods in functional genomics. He is a member of the IEEE.

Laura L. Elo received the PhD degree in applied mathematics from the University of Turku, Finland, in 2007. She is currently a postdoctoral researcher at Turku Centre for Biotechnology and at the Department of Mathematics, University of Turku. Her research interests include computational biology and bioinformatics with applications to computational genomics, proteomics, and systems biology.

Tero Aittokallio received the PhD degree in applied mathematics from the University of Turku, Finland, in 2001. He is currently a research fellow of the Academy of Finland. His research interests include the analysis of biological systems using mathematical modeling and machine learning methods, with special focus on developing data mining approaches to computational problems in systems biology.

Samuel Kaski received the DSc (PhD) degree in computer science from Helsinki University of Technology, Espoo, Finland, in 1997. He is currently a professor of bioinformatics at Helsinki University of Technology, Finland. His main research interests include statistical machine learning and data mining, bioinformatics, and information retrieval. He is a senior member of the IEEE.

Publication III

Leo Lahti, Juha E.A. Knuuttila, and Samuel Kaski. Global modeling of transcriptional responses in interaction networks. *Bioinformatics*, 26(21):2713–2720, 2010.

III

BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 21 2010, pages 2713–2720
doi:10.1093/bioinformatics/btq500

Gene expression

Advance Access publication September 2, 2010

Global modeling of transcriptional responses in interaction networks

Leo Lahti[1,*], Juha E. A. Knuuttila[2] and Samuel Kaski[1,*]

[1]Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University School of Science and Technology, PO Box 15400, FI-00076 Aalto and [2]Neuroscience Center, University of Helsinki, PO Box 54, FI-00014, Helsinki, Finland

Associate Editor: David Rocke

ABSTRACT

Motivation: Cell-biological processes are regulated through a complex network of interactions between genes and their products. The processes, their activating conditions and the associated transcriptional responses are often unknown. Organism-wide modeling of network activation can reveal unique and shared mechanisms between tissues, and potentially as yet unknown processes. The same method can also be applied to cell-biological conditions in one or more tissues.

Results: We introduce a novel approach for organism-wide discovery and analysis of transcriptional responses in interaction networks. The method searches for local, connected regions in a network that exhibit coordinated transcriptional response in a subset of tissues. Known interactions between genes are used to limit the search space and to guide the analysis. Validation on a human pathway network reveals physiologically coherent responses, functional relatedness between tissues and coordinated, context-specific regulation of the genes.

Availability: Implementation is freely available in R and Matlab at http://www.cis.hut.fi/projects/mi/software/NetResponse

Contact: leo.lahti@iki.fi; samuel.kaski@tkk.fi

Supplementary information : Supplementary data are available at *Bioinformatics* online.

Received on February 5, 2010; revised on August 24, 2010; accepted on August 26, 2010

1 INTRODUCTION

Coordinated activation and inactivation of genes through molecular interactions determines cell function. Changes in cell-biological conditions induce changes in the expression levels of co-regulated genes in order to produce specific physiological responses. A huge body of information concerning cell-biological processes is available in public repositories, including gene ontologies (Ashburner *et al.*, 2000), pathway models (Schaefer, 2006), regulatory information (Loots and Ovcharenko, 2007) and protein interactions (Kerrien *et al.*, 2007). Less is known about the contexts in which these processes are activated (Rachlin *et al.*, 2006), and how individual processes are reflected in gene expression (Montaner *et al.*, 2009). Although gene expression measurements provide only an indirect view to physiological processes, their wide availability provides a unique resource for investigating gene co-regulation on a genome- and organism-wide scale. This allows the detection of transcriptional responses that are shared by multiple tissues, suggesting shared physiological mechanism with potential biomedical implications, as demonstrated by the *Connectivity map* (Lamb *et al.*, 2006) where a number of chemical perturbations on a cancer cell line were used to reveal shared transcriptional responses between conditions to enhance screening of therapeutic targets. In this work, we study transcriptional responses of different tissues but the same methods can be directly used for modeling sets of cellular conditions within a single or multiple tissues as well.

Transcriptional responses have been modeled using so-called *gene expression signatures* (Hu *et al.*, 2006). A signature describes a co-expression state of the genes, associated with particular physiological states. Well-characterized signatures have proven to be accurate biomarkers in clinical trials, and hence reliable indicators of cell's physiological state. Disease-associated signatures are often coherent across tissues (Dudley *et al.*, 2009) or platforms (Hu *et al.*, 2006). Commercial signatures are available for routine clinical practice (Nuyten and van de Vijver, 2008), and other applications have been suggested recently (Dudley *et al.*, 2009). The established signatures are typically designed to provide optimal classification performance between two particular conditions. The problem with the classification-based signatures is that their associations to the underlying physiological processes are not well understood (Lucas *et al.*, 2009). Our goal is to enhance the understanding by deriving transcriptional signatures that are explicitly connected to well-characterized processes through the network.

We introduce and validate a novel approach for organism-wide discovery and analysis of transcriptional response patterns in interaction networks. Our algorithm has been designed to detect and model local regions in a network, each of which exhibits coordinated transcriptional response in a subset of measurements. In this study, the method is applied to investigate transcriptional responses of the network across a versatile collection of tissues across normal human body. The algorithm is independent of predefined classifications for genes or tissues. Organism-wide analysis can reveal unique and shared mechanisms between disparate tissues (Lage *et al.*, 2008), and potentially as yet unknown processes (Nacu *et al.*, 2007). The proposed NetResponse algorithm provides an efficient model-based tool for simultaneous feature selection and class discovery that utilizes known interactions between genes to guide the analysis. Related approaches include cMonkey (Reiss *et al.*, 2006) and a modified version of SAMBA biclustering (Tanay *et al.*, 2004).

*To whom correspondence should be addressed.

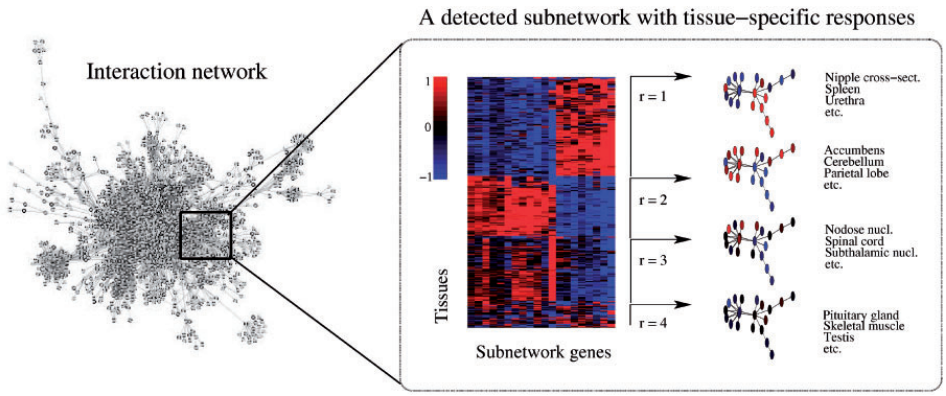


Fig. 1. Organism-wide analysis of transcriptional responses in a human pathway interaction network reveals physiologically coherent activation patterns and tissue-specific regulation. One of the subnetworks and its tissue-specific responses, as detected by the NetResponse algorithm is shown. The expression of each gene is visualized with respect to its mean level of expression across all samples.

However, these are application-oriented tools that rely on additional, organism-specific information, and their implementation is currently not available for most organisms, including human. We provide a general-purpose algorithm whose applicability is not limited to particular organisms.

NetResponse makes it possible to perform data-driven identification of functionally coherent network components and their tissue-specific responses. This is useful since the commonly used alternatives, predefined gene sets or pathways, are collections of intertwined processes rather than coherent functional entities (Nacu *et al.*, 2007). This has complicated their use in gene expression analysis, and methods have consequently been suggested to identify the 'key condition-responsive genes' of predefined gene sets (Lee *et al.*, 2008), or to decompose predefined pathways into smaller functional modules represented by gene expression signatures (Chang *et al.*, 2009). Our network-based search procedure detects the coordinately regulated gene sets in a data-driven manner. Gene expression provides functional information of the network that is missing in purely graph-oriented approaches for studying cell-biological networks (Aittokallio and Schwikowski, 2006). The network brings in prior information of gene function and connects the responses more closely to known processes than purely gene expression-based methods such as biclustering (Madeira and Oliveira, 2004), subspace clustering or other feature selection approaches (Law *et al.*, 2004; Roth and Lange, 2004). A key difference to previous network-based clustering methods, including MATISSE (Ulitsky and Shamir, 2007) and related approaches (Hanisch *et al.*, 2002; Shiga *et al.*, 2007) is that they assume a single correlated response between all genes in a module. NetResponse additionally models tissue-specific responses of the network. This allows a more expressive definition of a functional module, or a signature.

We validate the algorithm by modeling transcriptional responses in a human pathway interaction network across an organism-wide collection of tissues in normal human body. The results highlight functional relatedness between tissues, providing a global view on cell-biological network activation patterns.

2 METHODS

2.1 The NetResponse algorithm

We introduce a new approach for global detection and characterization of transcriptional responses in genome-scale interaction networks. NetResponse searches for local, connected *subnetworks* where joint modeling of gene expression reveals coordinated transcriptional response in particular tissues (Fig. 1). More generally, it is a new algorithm for simultaneous feature selection (for genes) and class discovery (for tissues) that utilizes known interactions between genes to limit the search space and to guide the analysis.

2.1.1 Gene expression signatures Subnetworks are the functional units of the interaction network in our model; transcriptional responses are described in terms of subnetwork activation. Given a physiological state, the underlying assumption is that gene expression in subnetwork n is regulated at particular levels to ensure proper functioning of the relevant processes. This can involve simultaneous activation and repression of the genes: sufficient amounts of mRNA for key proteins has to be available while interfering genes may need to be silenced. This regulation is reflected in a unique expression signature $\boldsymbol{s}^{(n)}$, a vector describing the associated expression levels of the subnetwork genes. The level of regulation varies from gene to gene; expression of some genes is regulated at precise levels whereas other genes fluctuate more freely. Given the physiological state, we assume that the distribution of observed gene expression is Gaussian, $\boldsymbol{x}^{(n)} \sim N(\boldsymbol{s}^{(n)}, \Sigma^{(n)})$.

2.1.2 Modeling tissue-specific transcriptional responses Each subnetwork is potentially associated with alternative transcriptional states, activated in different tissues and corresponding to unique combinations of processes. Since individual processes and their transcriptional responses are in general unknown (Lee *et al.*, 2008), detection of tissue-specific responses provides an efficient proxy for identifying functionally distinct states of the network. Our task is to detect and characterize these signatures. We assume that in a specific observation, the subnetwork n can be in any one of $R^{(n)}$ latent physiological states indexed by r. Each state is associated with a unique expression signature $\boldsymbol{s}_r^{(n)}$ over the subnetwork genes. Associations between the observations and the underlying physiological states are unknown, and treated as latent variables. This leads to a mixture model for gene expression

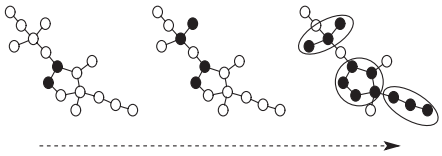


Fig. 2. The agglomerative subnetwork detection procedure. Initially, each gene is assigned in its own singleton subnetwork. Agglomeration proceeds by at each step merging the two neighboring subnetworks that benefit most from joint modeling of their transcriptional responses. This continues until no improvement is obtained by merging the subnetworks.

in the subnetwork n:

$$\boldsymbol{x}^{(n)} \sim \sum_{r=1}^{R^{(n)}} w_r^{(n)} p(\boldsymbol{x}^{(n)} | \boldsymbol{s}_r^{(n)}, \Sigma_r^{(n)}), \quad (1)$$

where each component distribution p is assumed to be Gaussian. In practice, we assume a diagonal covariance matrix $\Sigma_r^{(n)}$.

A particular transcriptional response is characterized by the triple $\{\boldsymbol{s}_r^{(n)}, \Sigma_r^{(n)}, w_r^{(n)}\}$. This defines the shape, fluctuations and frequency of the associated gene expression signature in subnetwork n. The feasibility of the Gaussian modeling assumption is supported by the previous observations of Kong *et al.* (2006), where predefined gene sets were used to investigate differences in gene expression between two predefined sample groups. In our model, the subnetworks, transcriptional responses and the activating tissues are learned from data. In one-channel data such as Affymetrix arrays used in this study, the centroids $\boldsymbol{s}_r^{(n)}$ describe absolute expression signals of the preprocessed array data. Relative differences can be investigated by comparing the detected responses. The model is applicable also on two-channel expression data when a common reference sample is used for all arrays since the relative differences are not altered by the choice of comparison baseline when the same baseline is used for all samples.

Now the model has been specified assuming the subnetworks are given. In practice, they are learned from the data. In order to do this, we make two assumptions. First, we rely on the prior information in the global interaction network, and assume that co-regulated gene groups are connected components in this network. Second, we assume that the subnetworks are independent. This allows a well-defined algorithm, and the subnetworks are then interpretable as independent components of transcriptional regulation. In practice the algorithm, described below, is an agglomerative approximation for searching for locally independent subnetworks.

2.2 Implementation

Efficient implementation is crucial for scalability. For fast computation, we use an agglomerative procedure where interacting genes are gradually merged into larger subnetworks (Fig. 2). Joint modeling of dependent genes reveals coordinated responses and improves the likelihood of the data when compared with independent models, giving the first criterion for merging the subnetworks. However, increasing subnetwork size tends to increase model complexity and the possibility of overfitting since the number of samples remains constant while the dimensionality (subnetwork size) increases. To compensate for this effect, we use a Bayesian information criterion (Gelman *et al.*, 2003) to penalize increasing model complexity and to determine optimal subnetwork size.

The cost function for a subnetwork G is $C(G) = -2L + q\log(N)$, where L is the (marginal) log-likelihood of the data, given the mixture model in Equation (1), q is the number of parameters, and N denotes sample size. NetResponse searches for a joint model for the network genes that maximizes the likelihood of observed gene expression, but avoids increasing model complexity through penalizing an increasing number of model parameters. An optimal model is searched for by at each step merging the subnetwork pair that produces the maximal gain in the cost function. More formally, the algorithm merges at each step the subnetwork pair G_i, G_j that minimizes the cost $\Delta\mathcal{C} = -2(L_{i,j} - (L_i + L_j)) + (q_{i,j} - (q_i + q_j))\log(N)$. The agglomerative scheme is as follows:

Initialize: learn univariate Gaussian mixture for the expression values of each gene, and bivariate joint models for all potential gene pairs with a direct link. Assign each gene into its own singleton subnetwork.

Merge: merge the neighboring subnetworks G_i, G_j that have a direct link in the network and minimize the difference $\mathcal{C}$. Compute new joint models between the newly merged subnetwork and its neighbors.

Terminate: continue merging until no improvement is obtained by merging the subnetworks ($\Delta\mathcal{C} \geq 0$).

The number $R^{(n)}$ of distinct transcriptional responses of the subnetwork is unknown, and is estimated with an infinite mixture model. Learning several multivariate Gaussian mixtures between the neighboring subnetworks at each step is a computationally demanding task, in particular when the number of mixture components is unknown. The Gaussian mixtures, including the number of mixture components, are learned with an efficient variational Dirichlet process implementation (Kurihara *et al.*, 2007). The likelihood L in the model is approximated by the lower bound of the variational approximation. The Gaussian mixture detects a particular type of dependency between the genes. In contrast to MATISSE (Ulitsky and Shamir, 2007) and other studies that use correlation or other methods to measure global co-variation, the mixture model detects coordinated responses that can be activated only in a few tissues. Tissue-specific joint regulation indicates functional dependency between the genes but it may have a minor contribution to the overall correlation between gene expression profiles. In principle, we could also model the dependencies in gene fluctuations within each individual response with covariances of the Gaussian components. However, this would heavily increase model complexity, and therefore we leave dependencies in gene-specific fluctuations within each response unmodeled, and focus on modeling differences between the responses. NetResponse provides a full generative model for gene expression, where each subnetwork is described with an independent joint mixture model. The maximum subnetwork size is limited to 20 genes to avoid numerical instabilities in computation. The infinite Gaussian mixture can automatically adapt model complexity to the sample size. We model subnetworks of 1–20 genes across 353 samples; similar dimensionality per sample size has previously been used with variational mixture models (Honkela *et al.*, 2008).

2.3 Data

2.3.1 Pathway interaction network We investigate the pathway interaction network based on the KEGG database of metabolic pathways (Kanehisa *et al.*, 2008) provided by the signaling pathway impact analysis (SPIA) package (Tarca *et al.*, 2009) of BioConductor (www.bioconductor.org). This implements the pathway impact analysis method originally proposed in Draghici *et al.* (2007), which is currently the only pathway analysis tool that considers pathway topology. SPIA provides the data in a readily suitable form for our analysis. Other pathway datasets, commonly provided in the BioPAX format, are not readily available in a suitable pairwise interaction form. Directionality and types of the interactions were not considered. Genes with no expression measurements were removed from the analysis. We investigate the largest connected component of the network with 1800 unique genes, identified by Entrez GeneIDs.

2.3.2 Gene expression data We analyzed a collection of normal human tissue samples from 10 post-mortem donors (Roth *et al.*, 2006), containing gene expression measurements from 65 normal tissues. To ensure sample quality, RNA degradation was minimized in the original study by flash freezing all samples within 8.5 h post-mortem. Only the samples passing Affymetrix quality measures were included. Each tissue has 3–9 biological replicates measured on the Affymetrix HG-U133plus2.0 platform. The reproducibility of our findings is investigated in an independent human

gene expression atlas (Su *et al.*, 2004), measured on the Affymetrix HG-U133A platform, where two biological replicates are available for each measured tissue. In the comparisons, we use the 25 tissues available in both datasets (adrenal gland cortex, amygdala, bone marrow, cerebellum, dorsal root ganglia, hypothalamus, liver, lung, lymph nodes, occipital lobe, ovary, parietal lobe, pituitary gland, prostate gland, salivary gland, skeletal muscle, spinal cord, subthalamic nucleus, temporal lobe, testes, thalamus, thyroid gland, tonsil, trachea and trigeminal ganglia). Both datasets were preprocessed with RMA (Irizarry *et al.*, 2003). Certain genes have multiple probesets, and a standard approach to summarize information across multiple probesets is to use alternative probeset definitions based on probe-genome remapping (Dai *et al.*, 2005). This would provide a single expression measure for each gene. However, since the HG-U133A array represents a subset of probesets on the HG-U133Plus2.0 array, the redefined probesets are not technically identical between the compared datasets. To minimize technical bias in the comparisons, we use probesets that are available on both platforms. Therefore, we rely on manufacturer annotations of the probesets and use an alternative approach (used e.g. by Nymark *et al.*, 2007), where one of the available probesets is selected at random to represent each unique gene. Random selection is used to avoid selection bias. When available, the 'xxxxxx_at' probesets were used because they are more specific by design than the other probe set types (www.affymetrix.com).

2.4 Validation

The NetResponse algorithm is validated with an application on the pathway interaction network of 1800 genes (Tarca *et al.*, 2009) across 353 gene expression samples from 65 tissues in normal human body (Roth *et al.*, 2006). NetResponse is compared with alternative approaches in terms of physiological coherence and reproducibility of the findings.

2.4.1 Comparison methods NetResponse is designed for organism-wide modeling of transcriptional responses in genome-scale interaction networks. Simultaneous detection of the subnetworks and their tissue-specific responses is a key feature of the model. A straightforward alternative would be a two-step approach where the subnetworks and their tissue-specific responses are detected in separate steps, although this can be suboptimal for detecting tissue-specific responses. Various methods are available for detecting subnetworks based on network and gene expression data (Hanisch *et al.*, 2002; Shiga *et al.*, 2007) in the two-step approach. We use MATISSE, a state-of-the-art algorithm described in Ulitsky and Shamir (2007). MATISSE finds connected subgraphs in the network such that each subgraph consists of highly correlated genes. The output is a list of genes for each detected subnetwork. Since MATISSE only clusters the genes, we model transcriptional responses of the detected subnetworks in a separate step by using a similar mixture model to the NetResponse algorithm. This combination is also new, and called MATISSE+ in this article. The second comparison method is the SAMBA biclustering algorithm (Tanay *et al.*, 2002). The output is a list of associated genes and tissues for each identified bicluster. SAMBA detects gene sets with tissue-specific responses, but, unlike NetResponse and MATISSE+, the algorithm does not utilize the network. Influence of the prior network is additionally investigated by randomly shuffling the gene expression vectors, while keeping the network and the within-gene associations intact. Comparisons between the original and shuffled data help to assess relative influence of the prior network on the results. Comparisons to randomly shuffled genes in SAMBA are not included since SAMBA does utilize network information.

2.4.2 Reproducibility in validation data Reproducibility of the findings is investigated in an independent validation dataset in terms of significance and correlation (for details, see Section 2.3). Each comparison method implies a grouping for the tissues in each subnetwork, corresponding to the detected responses. It is expected that physiologically relevant differences between the groups are reproducible in other datasets. We tested this by estimating differential expression between the corresponding tissues in the validation data for each pairwise comparison of the predicted groups using a standard test for gene set analysis (GlobalTest; Goeman *et al.*, 2004). To ensure that the responses are also qualitatively similar in the validation data, we measured Pearson's correlation between the detected responses and those observed in the corresponding tissues in validation data. The responses were characterized by the centroids provided by the model in NetResponse and MATISSE+. For SAMBA, we used the mean expression level of each gene within each group of tissues since SAMBA groups the tissues but does not characterize the responses. In validation data, the mean expression level of each gene is used to characterize the response within each group of tissues. Probesets were available for 75% of the genes in the detected subnetworks in the validation data; transcriptional responses with less than three probesets in the validation data were not considered. Validation data contained corresponding samples for >79%) of the predicted responses in NetResponse, MATISSE+ and SAMBA (Supplementary Table 1).

3 RESULTS

The validation results reported below demonstrate that the NetResponse algorithm is readily applicable for modeling transcriptional responses in interaction networks on an organism-wide scale. While biomedical implications of the findings require further investigation, NetResponse detects a number of physiologically coherent and reproducible transcriptional responses in the network, and highlights functional relatedness between tissues. It also outperformed the comparison methods in terms of reproducibility of the findings.

3.1 Application to human pathway network

In total, NetResponse identified 106 subnetworks with 3–20 genes (Supplementary Material). For each subnetwork, typically (median) three distinct transcriptional responses were detected across the 65 tissues (Supplementary Fig. 1). One of the subnetworks with four distinct responses is illustrated in Figure 1. Each response is associated with a subset of tissues. Statistically significant differences between the corresponding tissues were observed also in the independent validation data ($P < 0.01$; GlobalTest). Three of the four responses were also qualitatively similar (correlation > 0.8; Supplementary Fig. 2). The first response is associated with immune system-related tissues such as spleen and tonsil. Responses 2–3 are associated with neuronal tissues such as subthalamic or nodose nucleus, or with central nervous system, for example accumbens and cerebellum. The fourth group manifests a 'baseline' signature that fluctuates around the mean expression level of the genes. Testis and pituitary gland are examples of tissues in this group. While most tissues are strongly associated with a particular response, samples from amygdala, bone marrow, cerebral cortex, heart atrium and temporal lobe manifested multiple responses. While alternative responses reveal tissue-specific regulation, detection of physiologically coherent and reproducible responses may indicate shared mechanisms between tissues. Although the responses may reflect previously unknown processes, it is likely that some of them reflect the activation patterns of known pathways. Overlapping pathways can provide a starting point for interpretation. The subnetwork in Figure 1 overlaps with various known pathways, most remarkably with the MAPK pathway with 10 genes (detailed gene–pathway associations are provided in the Supplementary Material; see Subnetwork 12). MAPK is a general signal transduction system that participates in a complex, cross-regulated signaling network that is sensitive to cellular stimuli (Wilkinson and Millar, 2000).

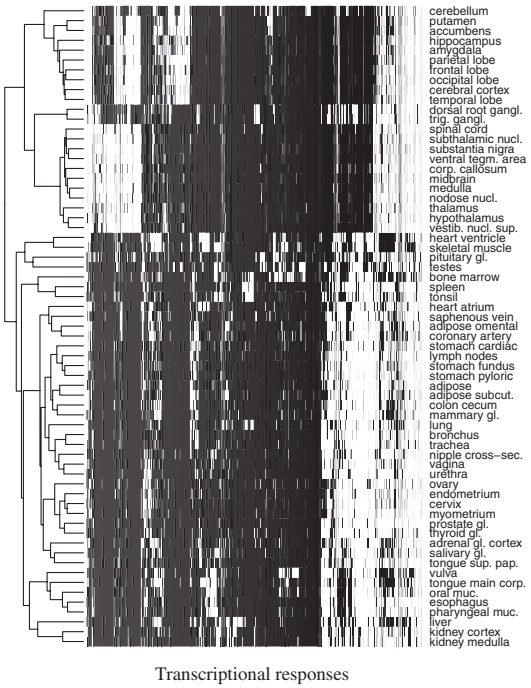


Fig. 3. Associations between 65 tissues (rows) and the detected transcriptional responses of the pathway interaction network of Figure 1. The shade indicates the probability of a particular transcriptional response in each tissue (black: $P=0$; white: $P=1$). Hierarchical clustering based on the signature co-occurrence probabilities between each pair of tissues highlights their relatedness.

Six subnetwork genes participate in the p53 pathway, which is a known regulator of the MAPK signaling pathway. In addition, p53 is known to interact with a number of other pathways, both as an upstream regulator and a downstream target (Wu, 2004). Both MAPK and p53 are associated with processes including cell growth, differentiation and apoptosis, and exhibit diverse cellular responses. Tissue-specific regulation can potentially explain the detection of alternative transcriptional states of the subnetwork.

The detected responses characterize absolute expression signals in our preprocessed one-channel array data. Systematic differences in the expression levels of the individual genes are normalized out in the visualization by showing the relative expression of each gene with respect to its mean expression level across all samples. Note that the choice of a common baseline does not affect the relative differences between the samples.

3.1.1 Tissue-selective network activation Associations between the tissues and the detected transcriptional responses are shown in Figure 3. Some responses are shared by many tissues, while others are more specific to particular contexts such as immune system, muscle or the brain. Related tissues often exhibit similar network activation patterns, which is seen by grouping the tissues according to co-occurrence probabilities of shared transcriptional response.

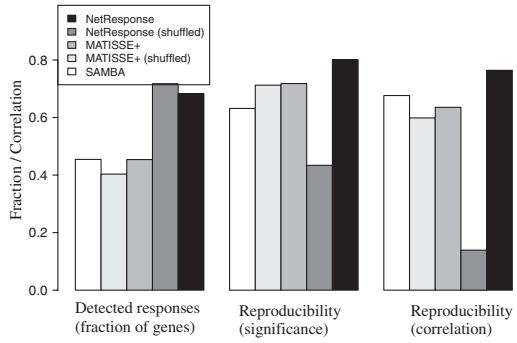


Fig. 4. Comparison between the alternative approaches. *Detected responses:* fraction of genes participating in the detected transcriptional responses. *Reproducibility (significance):* fraction of responses that are reproducible in the validation data in terms of differential expression between the associated tissues ($P<0.05$; GlobalTest). *Reproducibility (correlation):* median correlation between the gene expression levels of the detected responses and the corresponding tissues in the validation data.

This is known as *tissue selectivity* of gene expression (Liang *et al.*, 2006).

3.1.2 Probabilistic tissue connectome Tissue relatedness can be measured in terms of shared transcriptional responses (Supplementary Fig. 3). This is an alternative formulation of the *tissue connectome* map suggested by Greco *et al.* (2008) to highlight functional *connectivity* between tissues based on the number of shared differentially expressed genes at different thresholds. We use shared network responses instead of shared gene count. The use of co-regulated gene groups is expected to be more robust to noise than the use of individual genes. As the overall measure of connectivity between tissues, we use the mean of signature co-occurrence probabilities over the subnetworks, given the model in Equation (1). The analysis reveals functional relatedness between the tissues. In particular, two subcategories of the central nervous system appear distinct from the other tissues. Closer investigation of the observed responses would reveal how the tissues are related at transcriptional level (Supplementary Material).

3.2 Comparison to alternative approaches

NetResponse was compared with the alternative approaches in terms of physiological coherence and reproducibility of the findings (Fig. 4; Supplementary Table 1). NetResponse detected the largest amount of responses; 68% of the network genes were associated with a response, compared with 45% in MATISSE+ and SAMBA. At the same time, NetResponse outperformed the comparison methods in terms of reproducibility of the findings.

3.2.1 Physiological coherence The association between the responses and tissues was measured by normalized mutual information (NMI; Bush *et al.*, 2008) between the sample-response assignments and sample class labels within each subnetwork. The NMI varies from 0 (no association) to 1 (deterministic association). The transcriptional responses detected by NetResponse, MATISSE+ and SAMBA show statistically significant associations to particular

tissues with a significantly higher average NMI (0.46–0.50) than expected based on randomly labeled data (0.26–0.32; $P < 10^{-4}$; Wilcoxon test; Supplementary Table 1). The highest average NMI (0.50) was obtained by NetResponse but differences between NetResponse, MATISSE+ and SAMBA are not significant. NetResponse is significantly physiologically more coherent also when compared with results obtained with shuffled gene expression (NMI 0.22; $P < 10^{-12}$). The observations confirm the potential physiological relevance of the findings in NetResponse, MATISSE+ and SAMBA.

3.2.2 Reproducibility The majority of the detected responses were reproducible both in terms of significance and correlation (Supplementary Fig. 4) as described in Section 2.4. Of the predicted differences between groups of tissues, 80% were significant in validation data with $P < 0.05$ (GlobalTest), compared with 72 and 63% in MATISSE+ and SAMBA, respectively, or 43% obtained for randomly shuffled data with NetResponse (Fig. 4). The changes were also qualitatively similar; in NetResponse the median correlation between the detected responses and corresponding tissues in the validation data is 0.76, which is significantly higher ($P < 0.01$; Wilcoxon test) than in the comparison methods (MATISSE+: 0.64; SAMBA: 0.68), or in randomly shuffled NetResponse data (0.14). NetResponse detected responses for a larger fraction of the genes (68%) than the other methods. This seems an intrinsic property of the algorithm since it detected responses for a similar fraction of the genes also in the network with randomly shuffled genes (72%). However, only the findings from the real data were reproducible.

4 DISCUSSION

Cell-biological networks may cover thousands of genes, but any change in the physiological context typically affects only a small part of the network. While gene function and interactions are often subject to context-specific regulation (Liang *et al.*, 2006), they are typically studied only in particular experimental conditions. Organism-wide analysis could reveal highly specialized functions that are activated only in one or a few tissues. Detection of shared responses between the tissues can reveal previously unknown functional connections and help to formulate novel hypotheses of gene function in previously unexplored contexts. We provide a well-defined algorithm for such analysis.

The results support the validity of the model. NetResponse detected the largest number of responses without compromising physiological coherence or reproducibility of the findings compared with the alternatives. The most highly reproducible results were obtained by NetResponse. Further analysis is needed to establish the physiological role of the findings.

NetResponse is readily applicable for modeling tissue-specific responses in cell-biological networks, including pathways, protein interactions and regulatory networks. The network connects the responses to well-characterized processes, and provides readily interpretable results that are less biased toward known biological phenomena than methods based on predefined gene sets that are routinely used in gene expression studies to bring in prior information of gene function and to increase statistical power. However, these are often collections of intertwined processes rather than coherent functional entities. For example, pathways from KEGG may contain hundreds of genes, while only a small part of a pathway may be affected by changes in physiological conditions (Nacu *et al.*, 2007). This has complicated the use of predefined gene sets in gene expression studies. Draghici *et al.* (2007) demonstrated that taking into account aspects of pathway topology, such as gene and interaction types can improve the estimation of pathway activity. While their SPIA algorithm measures the activity of known pathways between two predefined conditions, our algorithm searches for potentially unknown functional modules, and detects their association to multiple conditions, or tissues, simultaneously. This is useful since biomedical pathways are human-made descriptions of cellular processes, often consisting of smaller, partially independent modules (Chang *et al.*, 2009; Hartwell *et al.*, 1999). Our data-driven search procedure can rigorously identify functionally coherent network modules where the interacting genes show coordinated responses. Joint modeling increases statistical power that is useful since gene expression, and many interaction data types such as protein–protein interactions, have high noise levels. The probabilistic formulation accounts for biological and measurement noise in a principled manner. Certain types of interaction data such as transcription factor binding or protein interactions are directly based on measurements. This can potentially help to discover as yet unknown processes that are not described in the pathway databases (Nacu *et al.*, 2007). False negative interactions form a limitation for the current model because joint responses of co-regulated genes can be modeled only when they form a connected subnetwork.

The need for principled methods for analyzing large-scale collections of gene expression data is increasing with their availability. Versatile gene expression atlases contain valuable information about shared and unique mechanisms between disparate tissues which is not available in smaller and more specific experiments (Lage *et al.*, 2008; Scherf *et al.*, 2000). For example, Lamb *et al.* (2006) demonstrated that large-scale screening of cell lines under diverse conditions can enhance the finding of therapeutic targets. Our model is directly applicable in similar exploratory tasks, providing tools for organism-wide analysis of transcriptional activity in normal human tissues (Roth *et al.*, 2006; Su *et al.*, 2004), cancer and other diseases (Kilpinen *et al.*, 2008; Lukk *et al.*, 2010) in a genome- and organism-wide scale. Similar collections are available for several model organisms including mouse (Su *et al.*, 2004), yeast (Granovskaia *et al.*, 2010) and plants (Schmid *et al.*, 2005). A key advantage of our approach compared with methods that perform targeted comparisons between predefined conditions (Ideker *et al.*, 2002; Sanguinetti *et al.*, 2008) is that it allows systematic organism-wide investigation when the responses and the associated tissues are unknown. The motivation is similar to SAMBA and other biclustering approaches that detect groups of genes that show coordinated response in a subset of tissues (Madeira and Oliveira, 2004), but the network ties the findings more tightly to cell-biological processes in our model. This can focus the analysis and improve interpretability. Since the non-parametric mixture model adjusts model complexity with sample size, our algorithm is potentially applicable also in smaller and more targeted datasets. For example, it could potentially advance disease subtype discovery by revealing differential network activation in subsets of patients.

Many large-scale collections are continuously updated with new measurements. Our algorithm provides no integration technique for new experiments yet; on-line extensions that could directly

integrate data from new experiments provide an interesting topic for further study. Another potential extension would be a fully Bayesian treatment that would provide confidence intervals, removing the need to assess significance of the results in a separate step. While our model provides a model-based criterion for detecting the responses without prior knowledge of the activating tissues, the statistical significance of the findings has to be verified in further experiments. The majority of the responses in our experiments could be verified in an independent dataset. Other potential extensions include adding more structure to address the directionality, relevance and probabilities of the interactions. Not all cell-biological processes have clear manifestations at transcriptome level. Hence, information of transcript and interaction types, as in SPIA, could potentially help to improve the sensitivity of our approach. We could also seek to loosen the constraints imposed by the prior network. However, such extensions would come with an increased computational cost. The simple and efficient implementation is a key advantage.

NetResponse is closely related to subspace clustering methods such as agglomerative independent variable component analysis (AIVGA; Honkela *et al.*, 2008). However, AIVGA and other model-based feature selection techniques (Law *et al.*, 2004; Roth and Lange, 2004) consider all potential connections between the features, which leads to more limited scalability. Finding a global optimum in our model would require exhaustive combinatorial search over all potential subnetworks. Since the complexity depends on the topology of the network, finding a general formulation for the model complexity is problematic. The number of potential solutions grows faster than exponentially with the number of features (genes) and links between them, making exhaustive search in genome-scale interaction networks infeasible. Approximative solutions are needed, and are often sufficient in practice. A combination of techniques is used to achieve an efficient algorithm compared with the model complexity. First, we focus the analysis on those parts of the data that are supported by known interactions. This increases modeling power and considerably limits the search space. Second, the agglomerative scheme finds an approximative solution where at each step the subnetwork pair that leads to the highest improvement in cost function is merged. This finds a solution relatively fast compared with the complexity of the task. Note that the order in which the subnetworks become merged may affect the solution. Finally, the variational implementation considerably speeds up mixture modeling (Kurihara *et al.*, 2007). The running time of our application was 248 min on a standard desktop computer (Intel 2.83GHz; Supplementary Fig. 5).

Investigation of a human pathway interaction network revealed tissue-specific regulation in the network, that is, groups of interacting genes whose joint response differs between tissues. This highlights the context-dependent nature of network activation, and emphasizes an important shortcoming in the current gene set-based testing methods (Nam and Kim, 2008): simply measuring gene set 'activation' is often not sufficient; it is also crucial to characterize *how* the expression changes, and in which conditions. Organism-wide modeling can provide quantitative information about these connections.

5 CONCLUSIONS

We have introduced and validated a general-purpose algorithm for global identification and characterization of transcriptional responses in genome-scale interaction networks across a diverse collection of tissues, applicable also to cell-biological conditions within and between tissues. An organism-wide analysis of a human pathway interaction network validated the model, and provided a global view on cell-biological network activation. The results revealed unique and shared mechanisms between tissues, and potentially help to formulate novel hypotheses of gene function in previously unexplored contexts.

Funding: Academy of Finland (207467); IST Programme of the European Community, under the PASCAL2 Network of Excellence (ICT-216886); Finnish Center of Excellence on Adaptive Informatics Research (AIRC; to L.L. and S.K.).

Conflict of Interest: none declared.

REFERENCES

Aittokallio,T. and Schwikowski,B. (2006) Graph-based methods for analysing networks in cell biology. *Brief. Bioinform.*, **7**, 243–255.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Bush,W. *et al.* (2008) Alternative contingency table measures improve the power and detection of multifactor dimensionality reduction. *BMC Bioinformatics*, **9**, 238.

Chang,J.T. *et al.* (2009) A genomic strategy to elucidate modules of oncogenic pathway signaling networks. *Mol. Cell*, **34**, 104–114.

Dai,M. *et al.* (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.*, **33**, e175.

Draghici,S. *et al.* (2007) A systems biology approach for pathway level analysis. *Genome Res.*, **17**, 1537–1545.

Dudley,J.T. *et al.* (2009) Disease signatures are robust across tissues and experiments. *Mol. Syst. Biol.*, **5**, 307.

Gelman,A. *et al.* (2003) *Bayesian Data Analysis*, 2nd edn. Chapman & Hall/CRC, Boca Raton, FL.

Goeman,J.J. *et al.* (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**, 93–99.

Granovskaia,M.V. *et al.* (2010) High-resolution transcription atlas of the mitotic cell cycle in budding yeast. *Genome Biol.*, **11**, R24

Greco,D. *et al.* (2008) Physiology, pathology and relatedness of human tissues from gene expression meta-analysis. *PLoS ONE*, **3**, e1880.

Hanisch,D. *et al.* (2002) Co-clustering of biological networks and gene expression data. *Bioinformatics*, **18**, 145–154.

Hartwell,L.H. *et al.* (1999) From molecular to modular cell biology. *Nature*, **402**, C47–C52.

Honkela,A. *et al.* (2008) Agglomerative independent variable group analysis. *Neurocomputing*, **71**, 1311–1320.

Hu,Z. *et al.* (2006) The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics*, **7**, 96.

Ideker,T. *et al.* (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18** (Suppl. 1), S233–S240.

Irizarry,R.A. *et al.* (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, **31**, e15.

Kanehisa,M. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36** (Suppl. 1), D480–D484.

Kerrien,S. *et al.* (2007) IntAct–open source resource for molecular interaction data. *Nucleic Acids Res.*, **35** (Suppl. 1), D561–D565.

Kilpinen,S. *et al.* (2008) Systematic bioinformatic analysis of expression levels of 17,330 human genes across 9,783 samples from 175 types of healthy and pathological tissues. *Genome Biol.*, **9**, R139.

Kong,S.W. *et al.* (2006) A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics*, **22**, 2373–2380.

Kurihara,K. *et al.* (2007) Accelerated variational Dirichlet process mixtures. In Schölkopf,B. *et al.* (eds) *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, pp. 761–768.

Lage,K. *et al.* (2008) A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proc. Natl Acad. Sci. USA*, **105**, 20870–20875.

Lamb,J. *et al.* (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.

Law,M. *et al.* (2004) Simultaneous feature selection and clustering using mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.*, **26**, 1154–1166.
Lee,E. *et al.* (2008) Inferring pathway activity toward precise disease classification. *PLoS Comput. Biol.*, **4**, e1000217.
Liang,S. *et al.* (2006) Detecting and profiling tissue-selective genes. *Physiol. Genomics*, **26**, 158–162.
Loots,G. and Ovcharenko,I. (2007) ECRbase: database of evolutionary conserved regions, promoters, and transcription factor binding sites in vertebrate genomes. *Bioinformatics*, **23**, 122–124.
Lucas,J.E. *et al.* (2009) Cross-study projections of genomic biomarkers: an evaluation in cancer genomics. *PLoS ONE*, **4**, e4523.
Lukk,M. *et al.* (2010) A global map of human gene expression. *Nat. Biotechnol.*, **28**, 322–324.
Madeira,S.C. and Oliveira,A.L. (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE Trans. Comput. Biol. Bioinformatics*, **1**, 24–45.
Montaner,D. *et al.* (2009) Gene set internal coherence in the context of functional profiling. *BMC Genomics*, **10**, 197.
Nacu,S. *et al.* (2007) Gene expression network analysis and applications to immunology. *Bioinformatics*, **23**, 850–858.
Nam,D. and Kim,S.-Y. (2008) Gene-set approach for expression pattern analysis. *Brief. Bioinform.*, **9**, 189–197.
Nuyten,D. and van de Vijver,M. (2008) Using microarray analysis as a prognostic and predictive tool in oncology: focus on breast cancer and normal tissue toxicity. *Semin. Radiat. Oncol.*, **18**, 105–114.
Nymark,P. *et al.* (2007) Gene expression profiles in asbestos-exposed epithelial and mesothelial lung cell lines. *BMC Genomics*, **8**, 62.
Rachlin,J. *et al.* (2006) Biological context networks: a mosaic view of the interactome. *Mol. Syst. Biol.*, **2**, 66.
Reiss,D. *et al.* (2006) Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics*, **7**, 280.
Roth,R. *et al.* (2006) Gene expression analyses reveal molecular relationships among 20 regions of the human CNS. *Neurogenetics*, **7**, 67–80.
Roth,V. and Lange,T. (2004) Feature selection in clustering problems. In Thrun,S. *et al.* (eds) *Advances in Neural Information Processing Systems* **16**. MIT Press, Cambridge, MA, pp. 473–480.
Sanguinetti,G. *et al.* (2008) MMG: a probabilistic tool to identify submodules of metabolic pathways. *Bioinformatics*, **24**, 1078–1084.
Schaefer,C.F. (2006) An Introduction to the NCI Pathway Interaction Database. *NCI-Nature Pathway Interaction Database*. [Epub ahead of print, doi:10.1038/PID.2006.001]
Scherf,U. *et al.* (2000) A gene expression database for the molecular pharmacology of cancer. *Nat. Genet.*, **24**, 236–244.
Schmid,M. *et al.* (2005) A gene expression map of Arabidopsis thaliana development. *Nat. Genet.*, **37**, 501–506.
Shiga,M. *et al.* (2007) Annotating gene function by combining expression data with a modular gene network. *Bioinformatics*, **23**, 468–478.
Su,A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.
Tanay,A. *et al.* (2002) Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, **18**, S136–S144.
Tanay,A. *et al.* (2004) Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. *Proc. Natl Acad. Sci. USA*, **101**, 2981–2986.
Tarca,A.L. *et al.* (2009) A novel signaling pathway impact analysis. *Bioinformatics*, **25**, 75–82.
Ulitsky,I. and Shamir,R. (2007) Identification of functional modules using network topology and high-throughput data. *BMC Syst. Biol.*, **1**, 8.
Wilkinson,M.G. and Millar,J.B. (2000) Control of the eukaryotic cell cycle by MAP kinase signaling pathways. *FASEB J.*, **14**, 2147–2157.
Wu,G.S. (2004) The functional interactions between the MAPK and p53 signaling pathways. *Cancer Biol. Therapy*, **3**, 146–151.

Publication IV

Leo Lahti, Samuel Myllykangas, Sakari Knuutila, and Samuel Kaski. Dependency detection with similarity constraints. In *Proceedings of the 2009 IEEE International Workshop on Machine Learning for Signal Processing XIX*, pages 89–94. IEEE, Piscataway, NJ, 2009.

IV

DEPENDENCY DETECTION WITH SIMILARITY CONSTRAINTS

Leo Lahti[1,2]*, Samuel Myllykangas*[3]*, Sakari Knuutila*[2] *and Samuel Kaski*[1]

1. Helsinki University of Technology, Department of Information and Computer Science
PO Box 5400, FI-02015 TKK, Finland
2. University of Helsinki and Helsinki University Central Hospital
Haartman Institute and HUSLAB, Department of Pathology, Helsinki, Finland
3. Stanford University School of Medicine, Department of Medicine, Division of Oncology, and
Stanford Genome Technology Center, Stanford University, Stanford, USA

ABSTRACT

Unsupervised two-view learning, or detection of dependencies between two paired data sets, is typically done by some variant of canonical correlation analysis (CCA). CCA searches for a linear projection for each view, such that the correlations between the projections are maximized. The solution is invariant to any linear transformation of either or both of the views; for tasks with small sample size such flexibility implies overfitting, which is even worse for more flexible nonparametric or kernel-based dependency discovery methods. We develop variants which reduce the degrees of freedom by assuming constraints on similarity of the projections in the two views. A particular example is provided by a cancer gene discovery application where chromosomal distance affects the dependencies between gene copy number and activity levels. Similarity constraints are shown to improve detection performance of known cancer genes.

1. INTRODUCTION

We develop methods for the task of detecting statistical dependencies between multiple sources of co-occurring data. The sources are assumed to share relevant common information, and additionally contain independent but unknown type of noise. The task is to discover the relevant information; both to detect and analyse or interpret it.

This is a particular type of a data fusion task, shared by *multi-view learning*. In multi-view learning each source is interpreted as a different view to the same items, and the task is to enhance classification performance by combining the views. Our task can be interpreted as unsupervised multi-view learning.

The project was funded by Tekes MultiBio project. LL and SK belong to the Adaptive Informatics Research Centre and Helsinki Institute for Information Technology HIIT. LL is funded by the Graduate School of Computer Science and Engineering. SK is partially supported by EU FP7 NoE PASCAL2, ICT 216886.

The traditional statistical way of finding dependencies between data sources is canonical correlation analysis, CCA, which generalizes correlation to multidimensional sources, retaining some of the nice interpretability of correlation coefficients. While the basic correlation coefficient assumes paired scalar values, canonical correlations assume paired vectorial values. The vectors are projected to scalar components before computing the correlations, using linear projections that maximize the correlations. For multidimensional data there will be many correlation coefficients; the second components are constrained to be uncorrelated with the first, and so on.

CCA is known to have two nice properties: the result is invariant to linear transformations of the data spaces, and the solution for any fixed number of components maximizes mutual information between linear projections for Gaussian data. These insights can be interpreted as motivations for generalizing using nonparametric methods [1, 2] and kernel CCA [3, 4].

The flexibility of CCA can cause overfitting problems that are specifically harmful with small sample sizes that abound in biomedical studies, for instance. When the views are high-dimensional, the completely unconstrained linear projections involve high degrees of freedom; several ways to regularize the CCA solution have been suggested to overcome some of the associated problems [5, 6, 7]. We introduce a complementary approach that is based on bringing in prior knowledge to constrain the model family.

Assuming the dimensions of the different views are not completely unrelated but instead are formed of related pairs, it makes sense to search for more constrained projections. In our application, the views are different measurements made on the same locations of the genome, and the dimensions correspond to these particular locations. Constraining the projections to be the same or at least similar in the different views will additionally enhance interpretability of the results, given that relationships between the same compo-

nents in the two views are natural.

Correlation-based CCA has been shown to correspond to the maximum likelihood solution of a simple generative model [4], where the two views are assumed to stem from a shared Gaussian latent variable and normally distributed data-set-specific noise. This has opened up the road to probabilistic and Bayesian formulations [8, 9] which make it possible to deal rigorously with uncertainty in small sample sizes and to include prior knowledge as Bayesian priors.

We suggest also a probabilistic version for constrained dependency search that provides a robust alternative for direct maximization of correlations. While the probabilistic version is slower to compute, it is the recommended choice when prior information of the types of dependency is available, or sample size is small.

The methods will be applied in a very promising application setup for knowledge discovery with dependency detection. The task is to find potential cancer genes by studying the relationship between changes caused by cancer in gene expression and gene copy numbers, that is, amplifications or deletions caused by mutations in cancer samples. Copy number changes are a key mechanism for cancer, and combination of copy number information with gene expression measurements can reveal functional effects of the mutations; gene expression data is informative of gene activity. The rationale goes as follows: Mutations having no functional effect will not cause cancer, and cancer-related gene expression changes may be side effects. Gene expression changes caused by mutations would be strong candidates for cancer mechanisms, and they contribute to the dependencies between the two data sources. While causation can be difficult to grasp, study of the dependencies can provide an efficient proxy for such effects.

2. CANONICAL CORRELATIONS WITH SIMILARITY CONSTRAINTS

2.1. Correlation-based approach

Correlation-based CCA searches for a maximally correlated linear projection of the original data sets with paired samples X and Y. It maximizes the correlation between the projections, $cor(X\mathbf{v}_x, Y\mathbf{v}_y)$, with respect to arbitrary projection vectors $\mathbf{v}_x, \mathbf{v}_y$. However, this flexibility easily leads to overfitting as demonstrated by the case study in Section 3.

In many applications prior information of the potential relationships between the features of the investigated data sets is available. Constraining the projections accordingly can potentially reduce overfitting and help to focus on specific types of dependencies between the two data sets. A particular example of such a model is provided by our cancer gene discovery application, where gene copy number changes are systematically correlated with the gene expression measurements from the same genes.

The relationship between the projections can be parametrized with a transformation matrix T such that $\mathbf{v}_y = T\mathbf{v}_x$. Maximization of the correlations between the projections leads to the following optimization problem:

$$\arg\max_{\mathbf{v},T} = \frac{\mathbf{v}^T\tilde{\Sigma}_{xy}T\mathbf{v}}{\sqrt{\mathbf{v}^T\tilde{\Sigma}_{xx}\mathbf{v}}\sqrt{(T\mathbf{v})^T\tilde{\Sigma}_{yy}T\mathbf{v}}}, \quad (1)$$

where the observed covariances of the two data sets are denoted by the $\tilde{\Sigma}$. Constraints on T can be used to guide the dependency search. We refer to this model as Similarity-constrained CCA (*SimCCA*). Suitable constraints depend on the particular applications; the solutions can be made to prefer particular types of dependencies in a soft manner with an appropriate penalty term on T.

While we consider only one-dimensional projections in the case study, multidimensional projection matrices are also possible. The optimal projection vectors can be sought iteratively as in ordinary CCA. Direct optimization of the correlations provides a simple and computationally efficient way to detect dependencies between data sources but it lacks an explicit model to deal with the uncertainty in the data and model parameters.

2.2. Probabilistic approach

An explicit model-based approach for the dependency exploration task is provided by the probabilistic modeling framework. We derive a probabilistic approach which should be more robust to small sample sizes. The correlation-based CCA has a direct connection to the maximum likelihood (ML) solution of the generative model [4, 10]:

$$\begin{aligned} X &\sim N(W_x\mathbf{z}, \Psi_x) \\ Y &\sim N(W_y\mathbf{z}, \Psi_y), \end{aligned} \quad (2)$$

assuming normally distributed $\mathbf{z}$, and data-set-specific covariances Ψ_x, Ψ_y. The dependency between the data sets is captured by the shared latent variable $\mathbf{z}$, and W_x, W_y characterize the relationship between the data sets. The covariances Ψ_x, Ψ_y characterize data set-specific effects. Note that while optimal projections $\mathbf{v}$ in the correlation-based CCA (Eq. 1) operate on the observed data, the parameters of interest, W_x, W_y, in probabilistic CCA mediate transformations of the latent variable $\mathbf{z}$.

The solutions of the probabilistic CCA can be constrained analogously to the correlation-based approach in Eq (1), by extending the formulation to include appropriate prior terms. The joint likelihood of the model is given by

$$P(X,Y,W,\Psi) \quad (3)$$

$$\sim P(X,Y|W_x,W_y,\Psi)P(W_y|W_x)P(W_x)P(\Psi) \quad (4)$$

$$= \int P(X,Y|W_x,W_y,\Psi,\mathbf{z}) \quad (5)$$

$$P(W_y|W_x)P(W_x)P(\Psi)P(\mathbf{z})d\mathbf{z}. \quad (6)$$

Here Ψ denotes the block-diagonal matrix of Ψ_x and Ψ_y. While incorporation of prior information of the data set-specific effects through the W_x and Ψ provides promising lines for further work, we focus on the shared latent variables as a probabilistic alternative to the correlation-based SimCCA. The relation between the transformation matrices for the shared latent variable is encoded by the prior term $P(W_y|W_x)$ and can be parametrized with a transformation matrix T such that $W_y = TW_x$. Assuming invertible $W_x^T W_x$, we have $T = W_y(W_x^T W_x)^{-1} W_x^T$.

By setting a prior on T it is possible to emphasize certain types of dependencies. With unconstrained T the solution reduces to ordinary probabilistic CCA. In the other extreme T is an identity matrix, $T = I$, and the two shared components, derived from x and y respectively, would be identical. The formulation would also allow tuning of T between these two extremes.

We consider the following simple prior for T: $P(T) = N_+(\parallel (T - I) \parallel |0, \sigma_T^2) = N_+(\parallel W_y(W_x^T W_x)^{-1} W_x^T) - I \parallel |0, \sigma_T^2)$. This can be plugged into $P(W_y|W_x)$ in Eq. (3). We have used Frobenius norm, and N_+ refers to truncated normal distribution for positive input values.

The σ_T^2 can tune the deviation of T from the identity matrix; a strict version of probabilistic SimCCA (pSimCCA) is obtained with $\sigma_T^2 \to 0$, while $\sigma_T^2 \to \infty$ yields ordinary probabilistic CCA (pCCA). With uninformative priors $P(W), P(\Psi) \sim 1$ and normally distributed shared latent variable $\mathbf{z} \sim N(0, I)$, the model has the negative log-likelihood

$$-logP(X, Y, W, \Psi) \sim log|\Sigma| + tr\Sigma^{-1}\tilde{\Sigma} + \frac{\parallel T - I \parallel}{\hat{\sigma}_T^2}. \quad (7)$$

Here $\Sigma = WW^T + \Psi$ contains the matrices W_x, W_y and data set specific covariances Ψ_x, Ψ_y. We have added the prior for T, which tunes the relationship between W_y and W_x. For other details, see [4, 5].

3. ANALYSIS OF FUNCTIONAL COPY NUMBER CHANGES IN GASTRIC CANCER

A promising biomedical application highlights the potential practical value of our approach. Constraints on the potential dependencies between gene expression and copy number are shown to improve the detection of known cancer genes. The advantages of constrained and probabilistic versions become particularly salient when the dimensionality increases and ordinary correlation-based CCA seriously overfits to the data.

3.1. Background and motivation

Copy number changes in chromosomal regions with tumor-suppressor or other cancer-associated genes have important contribution to cancer development and progression. Chromosomal gains and losses are likely to be positively correlated with the expression levels of the affected genes; copy number gain is likely to increase the expression of some of the associated genes whereas deletion will block gene expression. Identification of cancer-associated regions with functional copy number changes has potential diagnostic, prognostic and clinical impact for cancer studies.

Canonical correlations provide a principled framework for detecting the shared variation in gene expression and copy number data. Systematic copy number changes in a particular chromosomal region are captured by multiple copy number probes, and this is also visible in the expression levels of the genes within the affected region. The dependent signals can be subtle, however, as gene expression and copy number data are affected by high levels of unrelated biological and measurement variation, and the sample sizes are typically small.

Both correlation-based and probabilistic SimCCA combine power over the adjacent genes by capturing the strongest shared signal in gene expression and copy number observations. They can also ignore unrelated signal from poorly performing probes, or probes that measure genes that are not functionally affected by the copy number change. This provides tools to distinguish between so-called driver mutations having functional effects from less active passenger mutations, which is an important task in cancer studies. A further advantage of the probabilistic formulation is that the shared latent variable $\mathbf{z}$ provides a robust measure of the amplification effects in each patient.

3.2. Implementation

SimCCA is used to study the association between gene expression and copy number in a gastric cancer data set with 41 patients and 10 controls [11]. The gene expression and copy number data sets were matched for the analysis such that the closest probe by genomic location in gene expression data was selected for each copy number probe, and probes with no match between gene expression and copy number within 5000 bp interval were discarded. The preprocessed data has gene expression and copy number measurements from 5596 genes from ~ 700 chromosomal regions (cytobands). To satisfy the normality assumptions of our model, the data was log_2-transformed and the mean of the signals for each probe was set to 0 before the analysis.

Ordinary and constrained versions of canonical correlation analysis, CCA/SimCCA, were applied to investigate the dependencies between gene expression and copy numbers. The correlations were computed within a specific chromosomal window around each gene. The observed correlations provide a measure of dependency between gene copy number and expression data for each window, or chromosomal region.

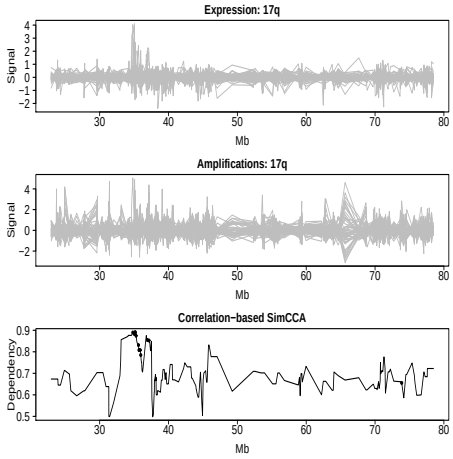


Fig. 1. Gene expression, copy number signal, and the dependency score for a sliding window of 15 genes along the chromosome arm 17q from the SimCCA method of Eq. (1). Known gastric-cancer associated genes from an expert-curated list are marked with black dots.

With unconstrained T, the models defined by Eqs. (1) and (7) reduce to ordinary correlation-based and probabilistic CCA, respectively. We assume that the constraints for T are provided prior to analysis, i.e. the prior parameter σ_T is fixed. Alternatively, σ_T could be optimized based on external criteria such as identification of the known cancer genes in our application. Our empirical results show, however, that already a simple prior for T without an explicit optimization procedure can improve the detection of known cancer genes.

We consider here the two extreme cases of the model where T is (i) completely unconstrained (ordinary CCA; $\sigma_T = \infty$), and (ii) $T = I$ ($\sigma_T = 0$). Point estimates for the model parameters were estimated with EM algorithm in the probabilistic version. Strength of the shared signal versus marginal effects is measured with $Tr(WW^T)/Tr(\Psi)$, where Tr denotes matrix trace. This yields a dependency score between copy number and expression data for the investigated chromosomal neighborghood around each gene. High scores highlight regions where the dependent signal between the two data sets is particularly high relative to the data-set-specific variation.

In addition to the correlation-based and probabilistic SimCCA, we tested a simplified probabilistic version with one-dimensional shared component $\mathbf{z}$ and isotropic covariances for the data-set-specific effects: ($\Psi_x = \sigma_x^2 I; \Psi_y = \sigma_y^2 I$). This is a special case of the full probabilistic model, and it reduces to principal component analysis (PCA) for concatenated data (X, Y). We refer to this method as pSimPCA. The simplified model does not distinguish between the shared and marginal effects as effectively as the full probabilistic CCA but it has fewer model parameters. Low-dimensional latent models are also faster to compute, and interpretation of the results is potentially more straightforward.

3.3. Validation

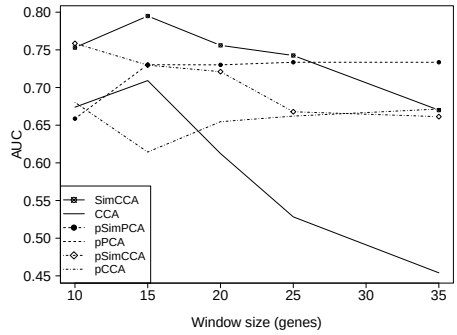


Fig. 2. AUC comparison.

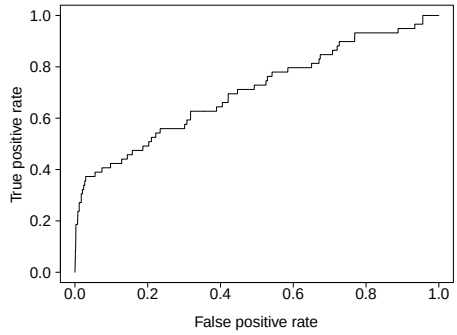


Fig. 3. ROC curve for the results from correlation-based SimCCA with a 15-gene sliding window.

Results from the correlation-based SimCCA are illustrated for chromosome arm 17q in Fig. 1, where SimCCA highlights a known cancer-associated region. The Figure shows the dependency score for the correlation-based SimCCA with a sliding window of 15 genes genes along the chromosome arm. The correlation-based and probabilistic approaches were compared in various window sizes (10,

15, 20, 25, and 35 genes). In each experiment, the gene list ordered by the dependency measure was compared to an expert-curated list of 59 gastric-cancer associated genes in our investigated data set [11].

The correlation-based and probabilistic models were compared with respect to their ability to detect the known cancer genes, measured with the AUC value of the ROC curve for each method. Results are summarized in Fig. 2. The best AUC value (0.79) was obtained with a chromosomal window of 15 genes for the correlation-based SimCCA that directly maximizes the correlations assuming identical projections (Eq. (1)). The corresponding ROC curve is shown in Fig. 3 and presents the tradeoff between true and 'false' positive findings along the ordered gene list. While a large proportion of the most significant findings are in fact known cancer genes, the remaining findings with no known associations to gastric cancer are promising candidates for further studies; among the 100 genes with highest dependencies between gene expression and copy number in their chromosomal neighborghood, 30% of the corresponding regions had previously known association with gastric cancer, while the proportion in the whole data set is 5%.

The constrained dependency detection methods introduced in this paper outperformed the unconstrained models in most cases. The improved detection performance of the constrained models is likely explained by their ability to reduce overfitting. Interestingly, the most constrained probabilistic model, pSimPCA, outperforms the other approaches in the highest-dimensional case. In contrast, the performance of correlation-based CCA decreases steadily with increasing dimensionality (window size) as the number of samples (patients) remains fixed to 51.

In our particular application, gene expression and copy number are expected to have strong linear correlations in cancer-associated chromosomal regions. Correlation-based approach is therefore directly suited for the cancer gene detection task and it has also fewer parameters than the probabilistic versions. However, the performance of correlation-based SimCCA reduces with increasing dimensionality. A likely explanation is that the correlation-based version models also some of the data set-specific effects, which is emphasized in higher-dimensions. The probabilistic formulations provide an alternative way to bring in prior knowledge of the relationships in a principled framework. A potential advantage of the probabilistic approaches is that they have an explicit model for distinguishing the shared signal from data set-specific variation.

3.4. Biomedical interpretation of the findings

The results obtained using the SimCCA algorithm are in general concordant with the output from signal-to-noise statistics and random permutation method that was applied previously to analyze the same data [11, 12]. The advantage of the current method is that it combines the signal across adjacent genes within a particular chromosomal region already in the modeling step. Probabilistic SimCCA estimates the strongest shared signal between the data sets and ignores other variation using explicit modeling assumptions. Probabilistic versions also provide a measure of the amplification effect for each patient which allows robust identification of small patient groups with profound amplification effects that would be missed in previous permutation-based tests due to low event frequency.

In concordance with the previous analyses, the chromosomal area showing the most significant correlation between the gene copy number and expression was 17q12-q21 (Fig. 1). There are a number of potential target genes in that region, including *ERBB2* and *PPP1R1B*, which show clinical and biological relevance. The *ERBB2* gene encodes a transmembrane tyrosine kinase receptor, which is a target of Herceptin. This monoclonal antibody specifically inactivates the overexpressed *ERBB2* protein and is used to treat metastatic breast cancer patients. The expression of *PPP1R1B* has been shown to be associated with repression of programmed cell death and increase the survival of the cancer cells in upper gastrointestinal tract cancers [13].

Another genomic region with correlated gene copy number and expression changes is 10q26, and *FGFR2* was identified as one of the putative target genes of that region. It was recently shown that in a set of gastric cancer cell lines, *FGFR2* amplification is driving the cell proliferation and promoting cancer cell survival. Furthermore, inhibition of the *FGFR2* protein by small molecules retained the growth arresting and apoptotically active phenotype [14]. The detected 1q22 region harbors the *MUC1* gene, whose expression was shown to be associated with the intestinal subtype of gastric cancer [11]. The 20q is one of the most frequently amplified chromosomal regions in gastric cancer. However, despite of high frequency of the amplifications the target genes in that area remain to be described. Our analysis pinpointed the strongest correlating loci to 20q13.12 and significantly narrow the list of putative target genes.

Some of the detected chromosomal regions did not have known association with gastric cancer; we are currently investigating these results more closely. The current application shows promising performance in detecting functional copy number changes, but biomedical studies provide also a number of other potential applications. For example, an increasing number of paired data sets are available in the future for studying the relationships between methylation, single-nucleotide polymorphisms, miRNAs, and other genomic features.

4. DISCUSSION

We have introduced methods that regularize CCA solutions by taking into account similarity constraints. The methods assume that the dependencies between the different views are visible in the same dimensions, that is, the projection matrices are similar. We introduced the constraints to standard CCA, resulting in a quick method that helps in solving the "small n large p problem", where n is the number of samples and p their dimensionality.

If n is very small compared to p, even the constrained CCA may not be sufficient, and we introduced a Bayesian variant into which further prior knowledge can be easily inserted, and which is capable of rigorously handling uncertainty in the data. While we only compare SimCCA and CCA in the present work, the probabilistic formulation allows smooth tradeoff between these two extremes, which is potentially useful in many applications.

Importantly, the constrained approaches for dependency detection can be directly applied in practical tasks in knowledge discovery; good results were obtained in a promising medical application on searching for potential cancer genes by detecting dependencies between gene expression and DNA copy number changes of the genes.

5. REFERENCES

[1] J.W. Fisher III, T. Darrell, W.T. Freeman, and P.A. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *Advances in Neural Information Processing Systems 13*, Cambridge, MA, 2000, pp. 772–778, MIT Press.

[2] A. Klami and S. Kaski, "Non-parametric dependent components," in *Proceedings of ICASSP'05, IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. V–209–V–212. IEEE, 2005.

[3] C. Fyfe and P.L. Lai, "ICA using kernel canonical correlation analysis.," in *Proceedings of the International Workshop on Independent Component Analysis and Blind Signal Separation (ICA 2000)*, 2000, pp. 279–284.

[4] F.R. Bach and M.I. Jordan, "A probabilistic interpretation of canonical correlation analysis," Tech. Rep. 688, Department of Statistics, University of California, Berkeley, 2005.

[5] T. De Bie and B. De Moor, "On the regularization of canonical correlation analysis," in *Proceedings of the International Conference on Independent Component Analysis and Blind Source Separation (ICA2003)*, S.-I. Amari, A. Cichocki, S. Makino, and N. Murata, Eds. 2003.

[6] L. Sun, S. Ji, and J. Ye, "A least squares formulation for canonical correlation analysis," in *ICML '08: Proceedings of the 25th international conference on Machine learning*, New York, NY, USA, 2008, pp. 1024–1031, ACM.

[7] H.D. Vinod, "Canonical ridge and the econometrics of joint production," *J. Econometrics*, vol. 4, no. 2, pp. 147–166, 1976.

[8] A. Klami and S. Kaski, "Generative models that discover dependencies between data sets," in *Machine learning for signal processing XVI*, S. McLoone, T. Adali, J. Larsen, M. Van Hulle, A. Rogers, and S.C. Douglas, Eds., pp. 123–128. IEEE, 2006.

[9] A. Klami and S. Kaski, "Local dependent components," in *Proceedings of ICML 2007, the 24th International Conference on Machine Learning*, Zoubin Ghahramani, Ed., pp. 425–432. Omnipress, 2007.

[10] C. Archambeau, N. Delannay, and M. Verleysen, "Robust probabilistic projections," in *Proceedings of the 23rd International conference on machine learning*, W.W. Cohen and A. Moore, Eds. 2006, pp. 33–40, ACM.

[11] S. Myllykangas, S. Junnila, A. Kokkola, R.Autio, I. Scheinin, T. Kiviluoto, M.L. Karjalainen-Lindsberg, J. Hollmén, S. Knuutila, P. Puolakkainen, and O. Monni, "Integrated gene copy number and expression microarray analysis of gastric cancer highlights potential target genes.," *Int J Cancer*, vol. 123, no. 4, pp. 817–25, 2008.

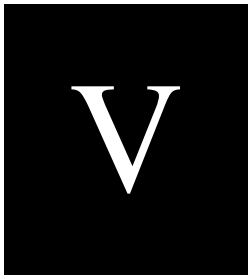
[12] S. Hautaniemi, M. Ringnér, P. Kauraniemi, R. Autio, H. Edgren, O. Yli-Harja, J. Astola, A. Kallioniemi, and O.-P. Kallioniemi, "A strategy for identifying putative causes of gene expression variation in human cancers," *J Franklin Institute*, vol. 341, pp. 77–88, 2004.

[13] A. Belkhiri, A. Zaika, N. Pidkovka, S. Knuutila, C. Moskaluk, and W. El-Rifai, "Darpp-32: a novel anti-apoptotic gene in upper gastrointestinal carcinomas," *Cancer Res*, vol. 65, pp. 6583–92, 2005.

[14] K. Kunii, L. Davis, J. Gorenstein, H. Hatch, M. Yashiro, A. Di Bacco, C. Elbi, and B. Lutterbach, "FGFR2-amplified gastric cancer cell lines require FGFR2 and Erbb3 signaling for growth and survival," *Cancer Res.*, vol. 68, no. 7, pp. 2340–8, 2008.

Publication V

Janne Sinkkonen, Janne Nikkilä, Leo Lahti, and Samuel Kaski. Associative clustering. In Boulicaut, Esposito, Giannotti, and Pedreschi, editors, *Machine Learning: ECML2004 (Proceedings of the ECML'04, 15th European Conference on Machine Learning)*, Lecture Notes in Computer Science 3201, 396–406. Springer, Berlin, 2004.



Associative Clustering

Janne Sinkkonen[1], Janne Nikkilä[1], Leo Lahti[1], and Samuel Kaski[2,1]

[1]Helsinki University of Technology, Neural Networks Research Centre,
P.O. Box 5400, FIN-02015 HUT, Finland
[2]Department of Computer Science,
P.O. Box 68, FIN-00014 University of Helsinki, Finland
{Janne.Sinkkonen,Janne.Nikkila,Leo.Lahti,Samuel.Kaski}@hut.fi
http://www.cis.hut.fi/projects/mi

Abstract. Clustering by maximizing the dependency between two paired, continuous-valued multivariate data sets is studied. The new method, *associative clustering (AC)*, maximizes a Bayes factor between two clustering models differing only in one respect: whether the clusterings of the two data sets are dependent or independent. The model both extends Information Bottleneck (IB)-type dependency modeling to continuous-valued data and offers it a well-founded and asymptotically well-behaving criterion for small data sets: With suitable prior assumptions the Bayes factor becomes equivalent to the hypergeometric probability of a contingency table, while for large data sets it becomes the standard mutual information. An optimization algorithm is introduced, with empirical comparisons to a combination of IB and K-means, and to plain K-means. Two case studies cluster genes 1) to find dependencies between gene expression and transcription factor binding, and 2) to find dependencies between expression in different organisms.

1 Introduction

Distributional clustering by the information bottleneck (IB) principle [21] groups nominal values x of a random variable X by maximizing the dependency of the groups with another, co-occurring discrete variable Y. Clustering documents x by the occurrences of words y in them is an example. For continuous-valued X, the analogue of IB is to *partition* the space of possible values $\boldsymbol{x} \in \mathbb{R}^{d_x}$ by discriminative clustering (DC); then the dependency of the partitions and y is maximized [16].

Both DC and IB maximize dependency between representations of random variables. Their dependency measures are asymptotically equivalent to mutual information (MI)[1]; the empirical mutual information used by IB and some variants of DC is problematic for finite data sets, however. A likelihood interpretation of empirical MI (see [16]) opens a way to probabilistic dependency measures that

[1] Yet another example of dependency maximization is canonical correlation analysis, which uses a second-moment criterion equivalent to mutual information assuming normally distributed data [11].

are asymptotically equivalent to MI but perform better for finite data sets [17]. The current likelihood formulation, however, breaks down when both margins are clustered simultaneously.

In this paper we introduce a novel method, *associative clustering (AC)*, for clustering of paired continuous-valued data by maximizing the dependency between the clusters of X and Y, later called *margin clusters*. A sample application is search for different types of city districts, by partitioning a city into demographically homogeneous regions (Fig. 1**B**). Here the paired data are the coordinates and demographics of the buildings of the city.

As a measure of dependency between the cluster sets, we suggest using a Bayes factor, extended from an optimization criterion for DC [17]. The criterion compares evidence for two models, one assuming independent margin clusters (clusters for $\boldsymbol{x}$ and $\boldsymbol{y}$), and the other allowing more general dependency of the margin clusters in generating data. With suitable prior assumptions the Bayes factor is equivalent to a hypergeometric probability commonly used as a dependency measure for contingency tables. It is well justified for finite data sets, avoiding the problems of empirical mutual information due to sampling uncertainty. Yet it is asymptotically equivalent to mutual information for large data sets. The Bayes factor is also usable as the cost function of IB [14].

AC will be applied for finding dependencies in gene expression data. It will be compared with standard K-means, computed independently for the two margins, which provides a baseline result. The comparison relevals how much is gained by explicit dependency modeling.

AC will additionally be compared with a new variant of IB. IB operates on discrete data, and therefore the continuous multivariates need first to be discretized into atomic regions, for example with K-means. The symmetric IB [5] can then compose discrete representations for the margins as combinations of the atomic regions. Again dependence of the representations is the criterion for clustering, and a dependency-maximized contingency table spanned by the margin clusters results. K-means discretization was chosen because its parameterization is similar to AC and, more importantly, because it is perhaps the most obvious alternative for multidimensional discretization.[2]

In IB, dependency has classically been measured by the (empirical) mutual information. As margin clusters are here combinations of very small Voronoi regions, IB finds dependencies between the data sets well, but on the other hand produces clusters that are potentially less local than those obtained by AC or standard K-means. We will evaluate the average dispersion of the clusters in the empirical tests of Section 4.

Both mixture models for discrete data [2,3,8] and Mixture Discriminant Analysis (MDA)-like [7,13] models for continuous data have common elements with our approach, and can readily be extended for the double-margin case. An im-

[2] Note that discretizing the dimensions independently of each other and using the Cartesian product as the multidimensional partitioning would fail badly for high-dimensional $\boldsymbol{x}$ or $\boldsymbol{y}$. As far as we know, better discretization methods or other comparable methods for co-clustering of continuous data have not been published.

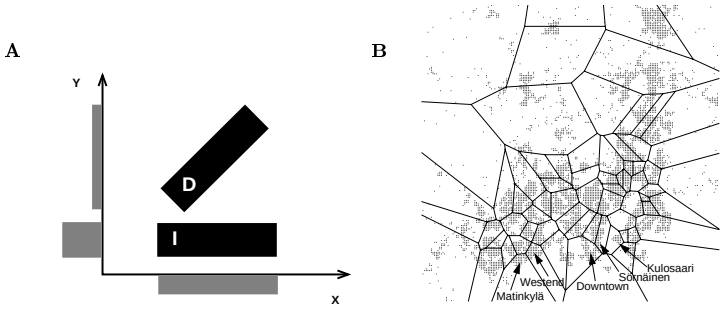


Fig. 1. A Demonstration of the difference between dependency modeling and joint density modeling. The hypothetical joint density of two one-dimensional variables x and y is plotted with black, and the respective marginal densities are depicted as histograms (*grey*). The marginals, here for simplicity univariate, correspond to the paired data of the AC setting. The visualized joint distribution consists of two equally-sized parts: a block in which x and y are independent (denoted by I) and another block (D) where x and y are dependent. Models for the joint distribution would focus equally on both blocks, whereas AC and IB focus on the dependent block D not explainable as products of the marginals, and neglect the independent block I. **B** Partitioning of Helsinki region into demographically homogeneous regions with AC. Here $\boldsymbol{x}$ contains geographic coordinates of buildings and $\boldsymbol{y}$ demographic information about inhabitants indicating social status, family structure, etc. Spatially relatively compact yet demographically homogeneous clusters emerge. For instance downtown and close-by relatively rich (Kulosaari, Westend) areas become separated from less well-off areas

portant difference is that our optimization criterion (as well as that of the Information Bottleneck) focuses only on the *dependencies* between the variables, skipping the parts of the joint distribution representable as a product of margins. Both goals are rigorous but different, as illustrated in Figure 1**A**.

2 Associative Clustering

2.1 Bayes Factor for Maximizing Dependency between Two Sets of Clusters

The dependency between two clusterings, indexed by i and j, for the same set of objects can be measured by mutual information if their joint distribution p_{ij} is known. If only a *contingency table* of co-occurrence frequencies n_{ij} computed from a finite data set is available, the mutual information computed from the empirical distribution would be a biased estimate. A *Bayes factor*, to be introduced below, then has the advantage of properly taking into account the finiteness of the data while still being asymptotically equivalent to mutual information. Bayes factors have been classically used as dependency measures for contingency tables (see, e.g., [6]) by comparing a model of dependent margins to another one

for independent margins. We will use the classical results as building blocks to derive an optimizable criterion for associative clustering; the novelty here is that the Bayes factor is optimized instead of only using it to measure dependency in a fixed table.

In general, frequencies over the cells of a contingency table are multinomially distributed. The model M_i of *independent margins* assumes that the multinomial parameters over cells are outer products of posterior parameters at the margins: $\theta_{ij} = \theta_i \theta_j$. The model M_d of *dependent margins* ignores the structure of the cells as a two-dimensional table and samples cell-wise frequencies directly from a table-wide multinomial distribution θ_{ij}. Dirichlet priors are assumed for both the margin and the table-wide multinomials.

Maximization of the Bayes factor

$$BF = \frac{p(\{n_{ij}\}|M_d)}{p(\{n_{ij}\}|M_i)}$$

with respect to the margin clusters then gives a contingency table where the margins are maximally dependent, that is, which cannot be explained as a product of independent margins. In the associative clustering introduced in this paper, the data counts are defined by the training data set and the parameters that determine how the continuous data spaces are partitioned into margin clusters. Then BF is maximized with respect to the parameters. If this principle were applied to two-way IB, the margins would be determined as groupings of nominal values of the discrete margin variables, and the BF would be maximized with respect to different groupings.

After marginalization over the multinomial parameters, the Bayes factor can be shown to take the form

$$BF = \frac{\prod_{ij} \Gamma(n_{ij} + n^{(d)})}{\prod_i \Gamma(n_{i\cdot} + n^{(x)}) \prod_j \Gamma(n_{\cdot j} + n^{(y)})} \,, \tag{1}$$

with $n_{i\cdot} = \sum_j n_{ij}$ and $n_{\cdot j} = \sum_i n_{ij}$ expressing the margins. The parameters $n^{(d)}$, $n^{(x)}$, and $n^{(y)}$ arise from Dirichlet priors. We have set all three parameters to unity, which makes BF equivalent to the hypergeometric probability classically used as a dependency measure of contingency tables. In the limit of large data sets, (1) becomes mutual information of the margins; [17] outlines the proof for the case of one fixed and one parameterized margin.

2.2 Optimization of AC

For paired data $\{(\boldsymbol{x}_k, \boldsymbol{y}_k)\}$ of real vectors $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, we search for partitionings $\{V_i^{(x)}\}$ for $\boldsymbol{x}$ and $\{V_j^{(y)}\}$ for $\boldsymbol{y}$. The partitions can be interpreted as clusters in the same way as in K-means; they are Voronoi regions parameterized by their centroids $\boldsymbol{m}_i$: $\boldsymbol{x} \in V_i^{(x)}$ if $\|\boldsymbol{x} - \boldsymbol{m}_i\| \leq \|\boldsymbol{x} - \boldsymbol{m}_k\|$ for all k, and correspondingly for $\boldsymbol{y}$. The Bayes factor (1) will be maximized with respect to the Voronoi centroids.

The optimization problem is combinatorial for hard clusters, but gradient methods are applicable after the clusters are smoothed. Gradients for the simpler one-margin problem have been derived in [17], and are analogous here. An extra trick, found to improve the optimization in the fixed-margin case [10], is applied here as well: The denominator of the Bayes factor is given extra weight. A choice of $\lambda^{(\cdot)} > 1$ introduces a regularizing term to the cost function that for large sample sizes approaches margin cluster entropy, and thereby in general favors solutions with uniform margin distributions.

The smoothed *BF*, here called *BF'*, is then optimized with respect to the $\{\boldsymbol{m}\}$ by a conjugate-gradient algorithm (see, for example [1]). We have

$$\log BF' = \sum_{ij} \log\Gamma\left(\sum_k g_i^{(x)}(\boldsymbol{x}_k) g_j^{(y)}(\boldsymbol{y}_k) + n^{(d)}\right)$$
$$- \lambda^{(x)} \sum_i \log\Gamma\left(\sum_k g_i^{(x)}(\boldsymbol{x}_k) + n^{(x)}\right) - \lambda^{(y)} \sum_j \log\Gamma\left(\sum_k g_j^{(y)}(\boldsymbol{y}_k) + n^{(y)}\right) ,$$
$$g_i^{(x)}(\boldsymbol{x}) \equiv Z^{(x)}(\boldsymbol{x})^{-1} \exp\left(-\|\boldsymbol{x} - \boldsymbol{m}_i^{(x)}\|^2 / \sigma_{(x)}^2\right) ,$$

and similarly for $g^{(y)}$. The $g(\cdot)$ are the smoothed Voronoi regions at the margins. The $Z(\cdot)$ is set to normalize $\sum_i g_i^{(x)}(\boldsymbol{x}) = \sum_j g_j^{(y)}(\boldsymbol{y}) = 1$. The parameters σ control the degree of smoothing of the Voronoi regions.

The gradient of $\log BF'$ with respect to an X-prototype $\boldsymbol{m}_i^{(x)}$ is

$$\nabla_{\boldsymbol{m}_i^{(x)}} \log BF' = \frac{1}{\sigma_{(x)}^2} \sum_{k,i'} (\boldsymbol{x}_k - \boldsymbol{m}_i^{(x)}) g_i^{(x)}(\boldsymbol{x}_k) g_{i'}^{(x)}(\boldsymbol{x}_k) \left(L_i^{(x)}(\boldsymbol{y}_k) - L_{i'}^{(x)}(\boldsymbol{y}_k)\right) ,$$

where

$$L_i^{(x)}(\boldsymbol{y}) \equiv \sum_j \Psi\left(\sum_k g_i^{(x)}(\boldsymbol{x}_k) g_j^{(y)}(\boldsymbol{y}_k) + n^{(d)}\right) g_j^{(y)}(\boldsymbol{y}) - \lambda^{(x)} \Psi\left(\sum_k g_i^{(x)}(\boldsymbol{x}_k) + n^{(x)}\right) ,$$

and for y accordingly. In the gradient, $\Psi(\cdot)$ is the digamma function.

Note that the smoothing is used during optimization only. Results are evaluated with hard clusters and the original *BF*.

3 Reference Methods

3.1 Information Bottleneck with K-means (K-IB)

For discrete X and Y, AC-type of clustering translates to grouping the nominal margin values to two sets of clusters that are maximally dependent. The setup is that of the information bottleneck [18,21].

Our continuous data must be discretized before IB can be applied. One approach is to first quantize the vectorial margins $\boldsymbol{x}$ and $\boldsymbol{y}$ separately by, for instance., K-means, without paying attention to possible dependencies between the two margins. This results in two sets of margin partitions which span a large, sparse contingency table that can be filled with frequencies of training data pairs $(\boldsymbol{x}_k, \boldsymbol{y}_k)$. The number of elementary

Voronoi regions is chosen by a validation set, as detailed in Section 4. In the second phase, the large table is compressed by standard IB to the desired size by aggregating the atomic margin clusters. At this stage, joins at the margins are made to explicitly maximize the dependency of margins in the resulting smaller contingency table.

IB algorithms are well described in the literature. We have used the symmetric sequential information bottleneck, described fully in [18]. The algorithm measures dependency of the margins by empirical mutual information, and it is optimized by re-assigning of individual samples (here atomic margin partitions) to clusters until a differential, local version of the cost function does not decrease. Optimization is robust and fast.

The final partitions obtained by the combination of K-means and IB are of a very flexible form, and therefore the method is expected to model the dependencies of the margin variables well—as long as one does not overfit to the data with too many K-means clusters. As a drawback, the final margin clusters will consist of many atomic Voronoi regions, and they are therefore not guaranteed to be especially homogeneous with respect to the original continuous variables ($\boldsymbol{x}$ or $\boldsymbol{y}$). Interpretation of the clusters may then be difficult. Our empirical results support both the good performance of K-IB and the non-localness of the resulting clusters.

3.2 K-means

The data sets will also be clustered by independent K-means clusterings in both data spaces. Results will represent a kind of a baseline, with no attempt to model dependency.

4 Experiments

4.1 Dependencies between Gene Expression Patterns and TF Binding Patterns

We sought gene regulation patterns by exploring dependencies between gene expression on the one hand, and measurement data about potential regulatory interactions on the other. The latter was measurements of binding patterns of putative regulatory proteins, transcription factors (TFs), in the promoter regions of the same genes. Associative clustering, K-IB, and K-means were applied to 6185 genes of the common yeast, *Saccharomyces cerevisiae*. The first margin data ($\boldsymbol{x}$) was 300-dimensional, consisting of expressions after 300 knock-out mutations[3] [9]. The second margin data ($\boldsymbol{y}$) consisted of 113-dimensional patterns of binding intensities of TFs [12]. Margin clusters would then ideally be internally homogeneous sets of expressions and TFs, selected to produce combinations (contingency table cells) with unexpectedly high or low numbers of genes.

For AC, the numbers of margin clusters were chosen to produce cross clusters (contingency table cells) with ten data samples on average. During the cross-validation runs margin clusters were initialized by K-means, and in each fold the best of three AC runs was chosen as the final AC clustering. The parameters $\sigma_{(\cdot)}$ were chosen with

[3] Knocking out means elimination of single genes. In all the data sets, missing values were imputed by gene-wise averages, and variances of dimensions were each separately normalized to unity.

a validation set (half of the data as a training set, and half of the data as validation set), and based on the previous experiments $\lambda^{(\cdot)}$=1.2.

Essentially the same test was conducted for the combination of K-means and information bottleneck (K-IB). Now the number of atomic K-means clusters was chosen with a validation set, resulting in 400 clusters for the expression space and 300 clusters for the transcription factor binding space. In the cross-validation runs, the atomic clusters were computed by K-means from three different random initializations, and for each of these a symmetric IB was sequentially optimized [18]. Of the three runs the best clustering (in the sense of IB cost) was chosen.

K-IB and AC tables were compared to each other and to tables obtained by bare margin K-means (10-fold cross validation, tables evaluated by equation 1, paired t-test). For this data, AC outperformed K-IB ($p<0.01$) and found more dependent clusters. Not surprisingly, significant differences to K-means were found ($p<0.01$) for both AC and K-IB.

The internal dispersion of the margin clusters was measured for all methods by the sum of intra-cluster component-wise variances. As expected, K-IB clusters are more scattered (Figure 2**A**) in both data spaces. Significant difference was found between AC and K-IB, but not between AC and K-means, nor between the random partitioning and K-IB.

Finally, data from the AC cross clusters was studied more closely to find potential biologically interesting gene concentrations, focusing on contingency table cells with the most unexpectedly high data counts. In two of the cells, for example, genes showed a clear and significant bias towards an over-representation of ribosomal protein coding genes. In the one cell, most of the genes coding for constituent proteins of the cellular ribosomal complex are present. In the other cell several genes coding for the mitochondrial ribosomal subunits are present, and also another set of genes coding for cellular ribosomal protein subunits.

4.2 Of Mice and Men

As a second test, we clustered human-mouse expression profiles of putative orthologs, that is, gene pairs sequence-wise similar enough to be suspected to have the same evolutionary origin (see Figure 3). Ideal margin clusters would be internally homogeneous by expression in at least one species. Cross clusters (cells of the contingency table formed from margin clusters) would then be cross-species clusters and will be optimized to detect cross-species regularities in gene expression.

Gene expression from 46 and 45 cell-lines (tissues) of human and mouse were available, respectively [19]. After removing non-expressed genes (Affymetrix AD<200), 4499 putative orthologs from the the HomoloGene [15] data base were available. After experiments analogous to those of Section 4.1, we found the human-mouse orthologs of left-out data to be significantly dependent according to both K-IB and AC (10-fold cross validation against K-means, paired t-test, $p<0.001$). Differences between AC and K-IB were not very clear. AC clusters, however, were probably more condensed ($p<0.05$; Fig. 2**B**) while tables obtained by K-IB were more dependent ($p<0.02$).

To illustrate the use of AC for finding interesting relationships, that is, groups of genes with functional similarity, we picked some cross clusters with significant deviation from the null hypothesis of independent margin clusters (see also Figure 3).

In the first example AC found a gene pair with a rare and potentially interesting functional relationship. This cell had unexpectedly few genes ($p<0.01$), in fact only

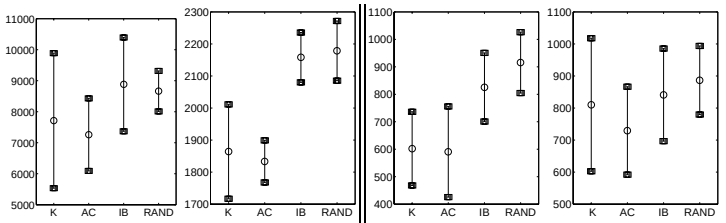


Fig. 2. Average internal dispersion of expression and TF margin clusters obtained by four methods. 'K' denotes independent K-means for the margins, and is supposed to produce very compact clusters. Clusters in 'RAND' are produced by random assignment and therefore represent an upper limit of dispersion. The AC and K-means clusters are more condensed in the expression and in the TF binding space than the IB clusters. Circles denote the average component-wise intra-cluster variances in left-out data of cross-validation folds (n=10), and squares show the approximate 99 percent confidence interval for the means over the cross validation folds. **A**: Yeast expression and TF binding. The differences between neither AC and K-means, nor between IB and RAND are statistically significant ($p>0.1$, 10-fold cross validation, paired t-test), but the difference between IB and AC is significant for both expression and TF binding ($p<0.01$). **B**: Homologous genes of human and mouse. The AC and K-means clusters seem to be more condensed in human and in mouse expression space than the IB or RAND clusters ($p<0.05$; paired t-test). Differences between AC and K-means or between IB and RAND are not statistically significant ($p>0.1$).

a single gene pair (LocusIDs 1808 and 12934). Average margin profiles of the cluster suggested activity in the human brain co-occurring with no activity in the mouse at all. Combining the margin profile information and the fact that only one this kind of gene exists in the contingency table, we may deduce that homologues which are active in human brain but totally silent in the mouse are very rare. Examples of such gene pairs may highlight interesting functional differences between the species. Indeed, the function of the gene was found to be related to embryo-stage brains *and* later brain activity only in humans (see Figure 3B).

In another example, a cross cluster contained unexpectedly many genes ($p<0.01$), most of them testis-specific (see Figure 3D). Due to their tissue specificity and importance for reproduction, they may have sustained their function during evolution.

5 Discussion

We have presented a novel method, associative clustering (AC), for clustering continuous paired data. It maximizes a Bayes factor between two sets of clusters. AC was found to perform better or equally well than a combination of K-means and information bottleneck (IB), and better than standard K-means. AC was also capable of extracting biologically interesting structure from paired gene expression data sets.

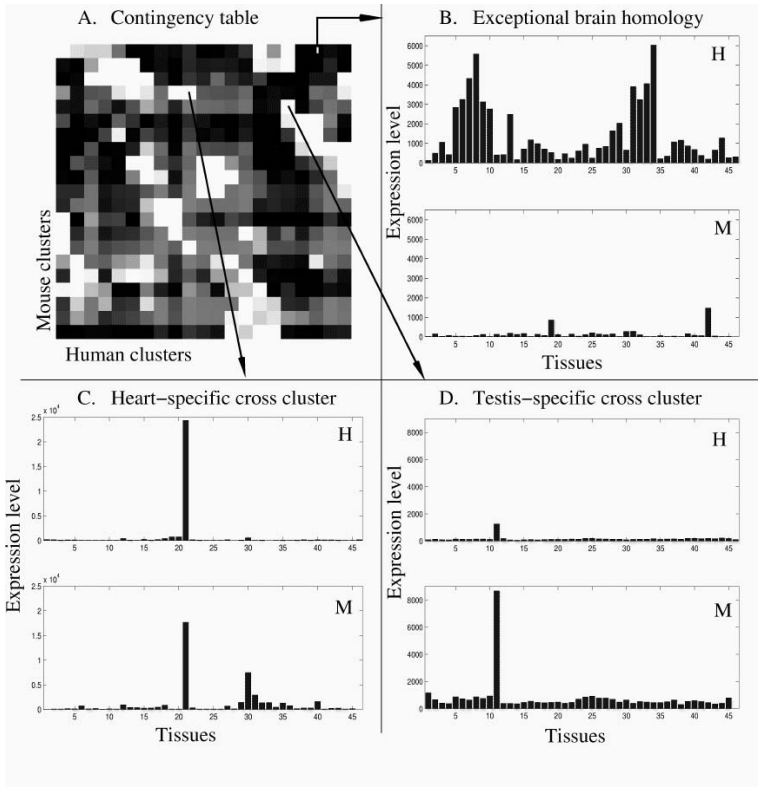


Fig. 3. **A** The contingency table from associative clustering of orthologous human-mouse gene pairs (orthologous genes are supposed or known to have a common evolutionary ancestor gene). *White cross clusters* contain an unexpectedly high number of genes compared to the margin-based expectation. *Black cross clusters* contain examples of exceptional gene pairs. **B** An example of an interesting outlier homology from a black cross cluster: the gene is highly active in most human tissues but is hardly expressed at all in mouse. The first 21 tissues are common for both species in B, C and D. **C** Cluster-wide average profiles reveal activity in heart tissue, and additional strong activity in mouse skeletal muscle. Measuring human skeletal muscle would reveal either a more complete homology or a species difference. **D** A densely populated cross cluster of testis-specific genes.

Maximization of the suggested Bayes factor is asymptotically equivalent to maximization of mutual information, and could therefore be seen as a dependency criterion alternative to empirical mutual information. It additionally gives information bottleneck-type dependency modeling a new justification that is clearly different from joint distribution models but still rigorously probabilistic. The Bayes factor could probably replace mutual informaton in the Information-Theoretic Co-Clustering Algorithm [4] as well.

The work could possibly be extended towards a compromise between strict dependency modeling and a model of the joint density (as has been done for one-sided clustering, [10]). Then the margins could be estimated in part from non-paired data. This would be analogous to "semisupervised learning" from partially labeled data (see e.g. [20]), the labels having been replaced by samples of co-occurring paired data.

Acknowledgments This work has been supported by the Academy of Finland, decisions #79017 and #207467. We thank Jaakko Peltonen for the code for sequential IB, and Juha Knuuttila and Christophe Roos for help with biological interpretation of the results.

References

1. Bazaraa, M.S., Sherali, H.D., Shetty, C.M.: Nonlinear Programming: Theory and Algorithms (1993). Wiley, New York
2. Blei, D., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Machine Learning Res. **3** (2003) 993–1022
3. Buntine, W.: Variational extensions to EM and multinomial PCA. In: Elomaa, T., Mannila, H., Toivonen, H. (eds.): Proc. of the ECML'02, Lecture Notes in Artificial Intelligence, 2430 (2002). Springer, Berlin, pp. 23–34
4. Dhillon, I.S., Mallela, S., Kumar, R.: A divisive information-theoretic feature clustering algorithm for text classification. J. Machine Learning Res. **3** (2003) 1265–1287
5. Friedman, N., Mosenzon, O., Slonim, N., Tishby, N.: Multivariate information bottleneck. In: Proc. of UAI'01, The 17th Conference on Uncertainty in Artificial Intelligence (2001). Morgan Kaufmann Publishers, San Francisco, CA, pp. 152–161
6. Good, I.J.: On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. Annals of Statistics, **4** (1976) 1159–1189
7. Hastie, T., Tibshirani, R.: Discriminant analysis by Gaussian mixtures. J. of the R. Stat. Soc. B **58** (1996) 155–176
8. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Machine Learning **42** (2001) 177–196
9. Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffrey, E., Dai, H., He, Y.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemaker, D.D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M., Friend, S.H.: Functional discovery via a compendium of expression profiles. Cell **102** (2000) 109–126
10. Kaski, S., Sinkkonen, J., and Klami, A.: Regularized discriminative clustering. In: Molina, C., Adali, T., Larsen, J., van Hulle, M., Douglas, S., Rouat, J. (eds.): Neural Networks for Signal Processing XIII (2003). IEEE, New York, NY, pp. 289–298

11. Kay, J.: Feature discovery under contextual supervision using mutual information. In: Proc. of IJCNN'92, International Joint Conference on Neural Networks (1992). IEEE, Piscataway, NJ, pp. 79–84
12. Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Tomphson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.-B., Volkert, T.L., Fraenkel, E., Gifford, D.K., Young, R.A.: Transcriptional regulatory networks in saccharomyces cerevisiae. Science **298** (2002) 799–804
13. Miller, D.J., Uyar, H.S.: A mixture of experts classifier with learning based on both labelled and unlabelled data. In: Mozer. M, Jordan, M., Petsche, T. (eds.): Advances in Neural Information Processing Systems, 9 (1997). MIT Press, Cambridge, MA, pp. 571–577
14. Peltonen, J., Sinkkonen, J., and Kaski, S.. Sequential information bottleneck for finite data. In: Proc. of the International Conference on Machine Learning (to appear)
15. Pruitt, K.D., Maglott, D.R.: RefSeq and LocusLink: NCBI gene-centered resources. Nucleic Acids Research **29** (2001) 137–141
16. Sinkkonen, J. and Kaski, S.: Clustering based on conditional distributions in an auxiliary space. Neural Computation **14** (2002) 217–239
17. Sinkkonen, J., Kaski, S., and Nikkilä, J.: Discriminative clustering: Optimal contingency tables by learning metrics. In Elomaa, T., Mannila, H., Toivonen, H. (eds.): Proc. of the ECML'02, 13th European Conference on Machine Learning (2002). Springer, Berlin, pp. 418–430
18. Slonim, N.: The information bottleneck: theory and applications (2002). PhD thesis, Hebrew University, Jerusalem
19. Su, A.I., Cooke, M.P., Ching, K.A., Hakak, Y., Walker, J.R., Wiltshire, T., Orth, A.P., Vega, R.G., Sapinoso, L.M., Moqrich, A., Patapoutian, A., Hampton, G.M., Schultz, P.G., and Hogenesch, J.B.: Large-scale analysis of the human and mouse transcriptomes. PNAS **99** (2002) 4465–4470
20. Szummer, M. and Jaakkola, T.: Kernel expansions with unlabeled examples. In: Leen, T, Dietterich, T., Tresp, V. (eds.): Advances in Neural Information Processing Systems, 13 (2001). MIT Press, Cambridge, MA, pp. 626–632
21. Tishby, N., Pereira, F.C., Bialek, W.: The information bottleneck method. In: Hajek, B. and Sreenivas, R.S. (eds.): Proc. of The 37th Annual Allerton Conference on Communication, Control, and Computing (1999). University of Illinois, Urbana, Illinois, pp. 368–377

Publication VI

Samuel Kaski, Janne Nikkilä, Janne Sinkkonen, Leo Lahti, Juha E.A. Knuuttila, and Cristophe Roos. Associative clustering for exploring dependencies between functional genomics data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2(3):203–216, 2005. Special Issue on Machine Learning for Bioinformatics – Part 2.

VI

Associative Clustering for Exploring Dependencies between Functional Genomics Data Sets

Samuel Kaski, Janne Nikkilä, Janne Sinkkonen, Leo Lahti, Juha E.A. Knuuttila, and Christophe Roos

Abstract—High-throughput genomic measurements, interpreted as cooccurring data samples from multiple sources, open up a fresh problem for machine learning: What is in common in the different data sets, that is, what kind of statistical dependencies are there between the paired samples from the different sets? We introduce a clustering algorithm for exploring the dependencies. Samples within each data set are grouped such that the dependencies between groups of different sets capture as much of pairwise dependencies between the samples as possible. We formalize this problem in a novel probabilistic way, as optimization of a Bayes factor. The method is applied to reveal commonalities and exceptions in gene expression between organisms and to suggest regulatory interactions in the form of dependencies between gene expression profiles and regulator binding patterns.

Index Terms—Biology and genetics, clustering, contingency table analysis, machine learning, multivariate statistics.

1 Introduction

Assume two data sets with *cooccurring* samples, that is, samples coming in pairs $(\mathbf{x}, \mathbf{y})$, where $\mathbf{x}$ belongs to the first set and $\mathbf{y}$ to the second set. In this paper, both $\mathbf{x}$ and $\mathbf{y}$ are gene expression profiles or other multivariate real-valued genomic measurements about the same gene. The general research problem is to find *common properties* in the set of pairs; statistically speaking, the goal is to find statistical dependencies between the pairs.[1]

In this paper we search for dependencies expressible by clusters. The standard unsupervised clustering methods, reviewed for gene expression clustering for instance in [32], aim at finding clusters where genes have similar expression profiles. Our goal is different: to cluster the $\mathbf{x}$ and the $\mathbf{y}$ separately such that the dependencies between the two clusterings capture as much as possible of the statistical dependencies between the two sets of clusters. In this sense, the clustering is *associative*; it finds associations between samples of different spaces. The research problem will be formalized in Section 2.

1. The fundamental difference from searching for differences between data sets [18], where the relative order of the samples within the two sets is not significant, both sets are within the same space, and the goal is to find differences between data *distributions*, is that our data are paired and we search for commonalities between the pairs of *samples* that can have different variables (attributes) and different dimensionalities.

- *S. Kaski and J. Nikkilä are with the University of Helsinki, Department of Computer Science, PO Box 68, FI-00014 University of Helsinki, Finland. E-mail: samuel.kaski@cs.helsinki.fi, janne.nikkila@hut.fi.*
- *J. Sinkkonen and L. Lahti are with the Helsinki University of Technology, Neural Networks Research Centre, PO Box 5400, FI-02015 HUT, Finland. E-mail: {janne.sinkkonen, leo.lahti}@hut.fi.*
- *J.E.A. Knuuttila is with the Neuroscience Center, PO Box 56, FI-00014 University of Helsinki, Finland. E-mail: Juha.Knuuttila@helsinki.fi.*
- *C. Roos is with Medicel Oy, Huopalahdentie 24, FI-00350 Helsinki, Finland. E-mail: christophe.roos@helsinki.fi.*

Manuscript received 8 Sept. 2004; revised 22 Dec. 2004; accepted 14 Apr. 2005; published online 31 Aug. 2005.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-0140-0904.

The problem of searching for common properties in two or more paired data sets differs from classic machine learning problems, commonly categorized into unsupervised and supervised. Supervised learning targets at finding classes (in classification) or predicted values of a variable (in regression). In probabilistic terms, the goal is to build a good model for the distribution $p(\mathbf{y}|\mathbf{x})$ while, in the kind of dependency modeling discussed in this paper, the goal should be symmetric. Basic unsupervised learning, on the other hand, is symmetric in a trivial sense: All variation of one variable—be it $\mathbf{x}$, $\mathbf{y}$, or the combination $(\mathbf{x}, \mathbf{y})$—is modeled, and there is no mechanism for separating between-data-set variation from within-data-set variation. Common to both kinds of learning, and indeed to all machine learning, is model fitting: A model parameterized by $\boldsymbol{\theta}$ is fitted to the data.

A different kind of problem to be addressed in this paper is modeling only the variation in $\mathbf{x}$ and $\mathbf{y}$ that is *common* to both variables. In other words, we search for *dependencies* between the $\mathbf{x}$ and $\mathbf{y}$. This symmetric goal has traditionally been formalized as maximizing the dependency between two representations, $\hat{\mathbf{x}} \equiv \mathbf{f}_x(\mathbf{x}; \boldsymbol{\theta}^x)$ and $\hat{\mathbf{y}} \equiv \mathbf{f}_y(\mathbf{y}; \boldsymbol{\theta}^y)$, of $\mathbf{x}$ and $\mathbf{y}$, respectively. A familiar example is canonical correlation analysis [24], where both the $\mathbf{f}_x$ and $\mathbf{f}_y$ are linear projections and the data are assumed to be normally distributed. This idea has been generalized to nonlinear functions [4] and to finding clusters of $\mathbf{x}$ informative of a nominal-valued y [3], [37]. It has been formalized in the information bottleneck framework [44], [40], resulting in efficient algorithms for two nominal-valued variables [41], [35].

Symmetric dependency modeling with non or semiparametric methods (such as clustering) is a natural way of formalizing the search for commonalities in cooccurring data sets, when one is not able or willing to postulate a detailed parametric model a priori. Such situations are common in modern data-driven functional genomics: Microarray-based high-throughput measurement techniques make it possible

1545-5963/05/$20.00 © 2005 IEEE Published by the IEEE CS, CI, and EMB Societies & the ACM

to test broad hypotheses, related, for example, to organism-wide differences in response or to functions of a gene over a range of organisms. Mining the data stored in community-resource databanks for new hypotheses is fruitful as well. In data mining, the search for dependencies between data sets is a considerably better-defined target than the common, unsupervised search for clusters and other regularities.

We study two cases of symmetric dependency modeling: search for regularities and differences in expression of orthologous genes in different organisms and search for regulatory interactions between expression and transcription factor binding patterns. More generally, we argue that, once a research goal can be dressed into a search for dependencies between data sets, our approach is a well-defined middle ground between purely hypothesis-driven research for which hypotheses must be available and purely exploratory research, where the task is often ill-defined.

Analogically to the two linear projections in canonical correlation analysis, we use two sets of clusters as the representations in the dependency search. Clusters are more flexible than linear projections and they have a definite role in exploratory data analysis, that is, in "looking at the data:" Clustering reveals outliers, finds groups of similar data, and simply compresses numerous samples into a more manageable and even visualizable summary. Clusters and other kinds of unsupervised models are of particular importance as the first step of microarray data analysis, where data are often noisy and even erroneous, and in general not well-known a priori.[2]

For microarray data, the existing dependency-searching techniques have two deficiencies. First, mutual information, the dependency measure that they maximize, is defined for probability distributions which in turn need to be estimated from samples. The separate estimation stage with its own optimality criteria will introduce uncontrollable errors to the models. The errors are negligible for asymptotically large data sets but nonnegligible for many real-life sets. We will directly define a dependency measure for data instead of distributions and justify it by combinatorial and Bayesian arguments. For asymptotically large data sets, the dependency measure becomes mutual information and can therefore be viewed as a principled alternative to mutual information for finite data sets.

The second shortcoming has been that the models are not applicable to symmetric dependency clustering of *continuous* data. While a trivial extension of existing continuous-data methods may seem sufficient, a conceptual change is actually required. Existing finite-data formulations either maximize the likelihood $p(\mathbf{y}|\mathbf{x})$ of one data set, say $\mathbf{y}$, given $\mathbf{x}$, or maximize the symmetric joint likelihood for $p(\mathbf{x}, \mathbf{y})$. Neither of these approaches is dependency modeling: Conditional models are asymmetric, while joint density models represent all variation in $\mathbf{x}$ and $\mathbf{y}$ instead of common variation and, therefore, do not even asymptotically reduce to mutual information. A solution we present in this paper is to use a hypothesis comparison approach which translates to a Bayes factor cost function.

Bayesian networks, used also as models of expression regulation [16], [36], are models for the joint density of all data sources. In these models, the structure of dependencies between variables is, at least to some extent, fixed in advance. To a degree, dependencies can be learned from data, but learning is hard and data-intensive. Our approach complements Bayesian networks in two ways. First, it is more exploratory and assumption-free because no dependency structure is imposed, except the one implied by cluster parameterization and division of the data set. Second, as joint distribution models, Bayesian networks represent not only the common variation between the data sets but partly also the unique variation within each data set. In this sense, the representations they produce are compromises for the task of modeling the between-set variation.

From the biological perspective, the advantages of clustering by maximizing dependency between two sources of genomic information are at least two-fold. First, the new problem setting makes it possible to formulate new kinds of hypotheses about the dependency of the sources, not possible with conventional one-source clusterings. Such hypotheses are sought in the orthologous genes application in Section 5. Second, mining for regularities in the common properties of two data sets is a more constrained problem that mining for any kinds of regularities within either of them. Hence, assuming the sets are chosen cleverly, the results are potentially better targeted. Our hypothesis is that there will be less false positives in the discovered regulatory interactions when expression and transcription factor binding are combined in a dependency maximizing way, compared to one-source clusterings. We will study the interactions in Section 6.

2 Associative Clustering

The abstract task solved by associative clustering (introduced in the preliminary paper [39]) is the following: cluster two sets of data, with samples $\mathbf{x}$ and $\mathbf{y}$, each separately, such that 1) the clusterings would capture as much as possible of the dependencies between pairs of data samples $(\mathbf{x}, \mathbf{y})$ and 2) the clusters would contain (relatively) similar data points. The latter is roughly a definition of a cluster.

Fig. 1 gives a brief overview of the method. For paired data $\{(\mathbf{x}_k, \mathbf{y}_k)\}$ of real vectors $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$, we search for partitionings $\{V_i^{(x)}\}$ for $\mathbf{x}$ and $\{V_j^{(y)}\}$ for $\mathbf{y}$. The partitions can be interpreted as clusters in the same way as in K-means; they are Voronoi regions parameterized by their prototype vectors $\mathbf{m}_i$. The $\mathbf{x}$ belongs to $V_i^{(x)}$ if $\|\mathbf{x} - \mathbf{m}_i^{(x)}\| \leq \|\mathbf{x} - \mathbf{m}_{i'}^{(x)}\|$ for all i' and correspondingly for $\mathbf{y}$.

2.1 Bayes Factor for Measuring Dependency between Two Sets of Clusters

The dependency between two cluster sets, indexed by i and j, can be measured by mutual information if the joint distribution p_{ij} is known. If only a *contingency table* of cooccurrence frequencies n_{ij} computed from a finite data set is available, mutual information computed from the empirical distribution would be a biased estimate. A *Bayes*

2. This very legitimate and necessary use of clustering in the beginning of the research process should not be confused with the widespread use of clusterings as a general-purpose tool in all possible research tasks, which could better be solved by other methods.

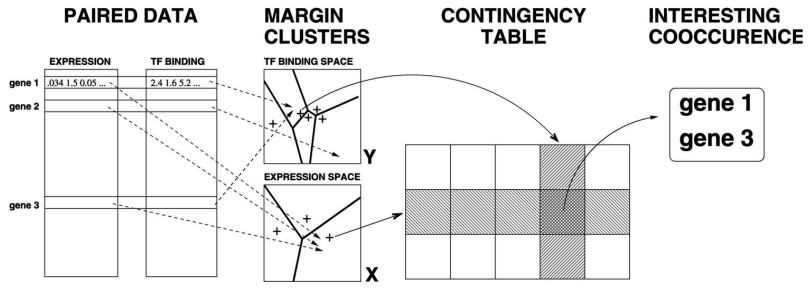


Fig. 1. Associative clustering (AC) in a nutshell. Two data sets are clustered into Voronoi regions. The Voronoi regions are defined in the standard way as sets of points closest to prototype vectors, but the prototypes are not optimized to minimize a quantization error but by the AC algorithm. In this example, the data sets are gene expression profiles and transcription factor (TF) binding profiles. A one-to-one correspondence between the sets exist: Each gene has an expression profile and a TF binding profile. As each gene falls to a TF cluster and to an expression cluster, we get a contingency table by placing the two sets of clusters as rows and columns and by counting genes falling to each combination of an expression cluster and a TF cluster. Rows and columns, that is, the Voronoi regions defined within each data set, respectively, are called *margin clusters*, while the combinations corresponding to the cells of the contingency table are called *cross clusters*. *Associative clustering*, by definition, finds Voronoi prototypes that maximize the dependency seen in the contingency table. Voronoi regions are representations for the data sets just as the linear combinations are in canonical correlation analysis. In both cases, dependency between the two parameterized representations is maximized. Maximization of dependency in a contingency table results in a maximal amount of surprises, counts not explainable by the margin distributions. The most surprising cross clusters with a very high or low number of genes potentially give rise to interesting interpretations. Reliability is assessed by the bootstrap.

factor, to be introduced below, has the advantage of properly taking into account the finite size of the data set while still being asymptotically equivalent to mutual information. Bayes factors have classically been used as dependency measures for contingency tables (see, e.g., [20]) by comparing a model of dependent margins to another model for independent margins. We will use the classical results as building blocks to derive an optimizable criterion for associative clustering; the novelty here is that the Bayes factor is *optimized* instead of only being used to measure dependency in a fixed table. The categorical variables defining the rows and columns of the contingency table are defined by the Voronoi regions. They are parameterized by the cluster prototypes which are optimized to maximize the Bayes factor.

The Bayes factor compares two alternative models, one describing a contingency table where the margins are dependent and the other a table with independent margins. The clusters are then tuned to make the dependent model describe the (contingency table) data better than the independent model, which can be interpreted as maximization of dependency.

In general, frequencies over the cells of a contingency table can be assumed to be multinomially distributed. The model M_I of *independent margins* assumes that the multinomial parameters over cells are outer products of posterior parameters at the margins: $\theta_{ij} = \theta_i \theta_j$. The model M_D of *dependent margins* ignores the structure of the cells as a two-dimensional table and samples cell-wise frequencies directly from a table-wide multinomial distribution θ_{ij}. Dirichlet priors are set for both the margin and the table-wide multinomials.

Maximization of the Bayes factor

$$BF = \frac{p(\{n_{ij}\}|M_D)}{p(\{n_{ij}\}|M_I)}$$

with respect to the margin clusters then gives a contingency table where the margins are maximally dependent, that is, the table is as far from the product of independent margins as possible. In associative clustering, the counts are influenced by the parameters of the Voronoi regions. The BF is maximized with respect to these parameters.

After marginalization over the multinomial parameters, the Bayes factor takes the form (derivation in the technical report [38])

$$BF = \frac{\prod_{ij} \Gamma(n_{ij} + n^{(d)})}{\prod_i \Gamma(n_{i\cdot} + n^{(x)}) \prod_j \Gamma(n_{\cdot j} + n^{(y)})}, \tag{1}$$

where $n_{i\cdot} = \sum_j n_{ij}$ and $n_{\cdot j} = \sum_i n_{ij}$ express the margins. The hyperparameters $n^{(d)}$, $n^{(x)}$, and $n^{(y)}$ arise from Dirichlet priors. We have set all three hyperparameters to unity, which makes the BF equivalent to the hypergeometric probability classically used as a dependency measure of contingency tables. For large data set sizes N, the logarithmic Bayes factor approaches mutual information of the distribution $p_{ij} = n_{ij}/N$ with margins $p_i = n_{i\cdot}/N$ and $p_j = n_{\cdot j}/N$ [38]:

$$\begin{aligned} \frac{1}{N} \log BF &= \sum_{i,j} p_{ij} \log \frac{p_{ij}}{p_i p_j} - \log N + 1 + \mathcal{O}\left(\frac{1}{N} \log N\right) \\ &= \hat{I}(I, J) - \log N + 1 + \mathcal{O}\left(\frac{1}{N} \log N\right), \end{aligned} \tag{2}$$

where $\hat{I}(I, J)$ is the mutual information between the categorical variables I and J having cluster indices as their values.

2.2 Optimization of AC

The Bayes factor (1) will be maximized with respect to the Voronoi prototypes. The optimization problem is combinatorial for hard clusters, but gradient methods are applicable after the clusters are smoothed. Gradients are derived in a technical report [38]. An extra trick, found to improve the optimization in the simpler case where one of the margins is fixed [27], is applied here as well: The denominator of the Bayes factor is given extra weight by introducing constants

$\lambda^{(\cdot)}$. A choice of $\lambda^{(\cdot)} > 1$ introduces a regularizing term to the cost function that, for large sample sizes, approaches margin cluster entropy and, thereby, in general, favors solutions with uniform margin distributions.

The smoothed BF, here denoted by BF', is then optimized with respect to the cluster prototypes $\{\mathbf{m}\}$ by a conjugate-gradient algorithm (for a textbook account, see [2]). We have

$$\begin{aligned}\log BF' = &\sum_{ij} \log\Gamma\left(\sum_k g_i^{(x)}(\mathbf{x}_k) g_j^{(y)}(\mathbf{y}_k) + n^{(d)}\right) \\ &- \lambda^{(x)} \sum_i \log\Gamma\left(\sum_k g_i^{(x)}(\mathbf{x}_k) + n^{(x)}\right) \\ &- \lambda^{(y)} \sum_j \log\Gamma\left(\sum_k g_j^{(y)}(\mathbf{y}_k) + n^{(y)}\right),\end{aligned} \tag{3}$$

where

$$g_i^{(x)}(\mathbf{x}) \equiv Z^{(x)}(\mathbf{x})^{-1} \exp\left(-\|\mathbf{x} - \mathbf{m}_i^{(x)}\|^2 / \sigma_{(x)}^2\right)$$

and similarly for $g^{(y)}$. The $g(\cdot)$ are the smoothed Voronoi regions at the margins. The $Z(\cdot)$ is set to normalize $\sum_i g_i^{(x)}(\mathbf{x}) = \sum_j g_j^{(y)}(\mathbf{y}) = 1$. The parameters σ control the degree of smoothing of the Voronoi regions.

The gradient of $\log BF'$ with respect to an X-space prototype $\mathbf{m}_i^{(x)}$ is

$$\nabla_{\mathbf{m}_i^{(x)}} \log BF' = \frac{1}{\sigma_{(x)}^2} \sum_{k,i'} \left(\mathbf{x}_k - \mathbf{m}_i^{(x)}\right) g_i^{(x)}(\mathbf{x}_k) g_{i'}^{(x)}(\mathbf{x}_k) \left(L_i^{(x)}(\mathbf{y}_k) - L_{i'}^{(x)}(\mathbf{y}_k)\right),$$

where

$$\begin{aligned}L_i^{(x)}(\mathbf{y}) \equiv &\sum_j \Psi\left(\sum_k g_i^{(x)}(\mathbf{x}_k) g_j^{(y)}(\mathbf{y}_k) + n^{(d)}\right) g_j^{(y)}(\mathbf{y}) \\ &- \lambda^{(x)} \Psi\left(\sum_k g_i^{(x)}(\mathbf{x}_k) + n^{(x)}\right)\end{aligned}$$

and for $\mathbf{y}$ accordingly. In the gradient, $\Psi(\cdot)$ is the digamma function.

In summary, the optimization of AC proceeds as follows:

1. Parameters $\{\mathbf{m}^{(x)}\}$ and $\{\mathbf{m}^{(y)}\}$ are independently initialized by choosing the best of several (here: three) K-means runs initialized randomly.
2. On the basis of experience with other data sets, we choose $\lambda^{(\cdot)} = 1.2$.
3. Parameters $\sigma_{(\cdot)}$ are chosen by running the algorithm for half of the data and testing on the rest.
4. The $\{\mathbf{m}^{(x)}\}$ and $\{\mathbf{m}^{(y)}\}$ are optimized with a standard conjugate gradients algorithm, using $\log BF'$ as the target function.

Gradients of the $\mathbf{m}$-parameters plugged into the algorithm are shown above. The reported results are from cross-validation runs.

In one-margin optimization with clusters in the other margin fixed, the smoothing trick performs equivalently to or better than simulated annealing [27]. Also note that smoothing is for optimization only: Results are evaluated with BF, which translates to having crisp clusters.

2.3 Uncertainty in Clustering

Our use of Bayes factors is different from their traditional use in hypothesis testing, cf., [20]. In AC, we do not test any hypotheses but maximize the Bayes factor to explicitly find dependencies. This leaves the uncertainty of the solution open.

A widely used "light-weight" (compared to posterior computation) method to take into account the uncertainty in clustering is bootstrap [12], [21]. As in [29], we use bootstrap to produce several perturbed clusterings. We wish to find cross clusters (contingency table cells) that signify dependencies between the data sets and are reproducible.

Reproducibility of the found dependencies will be estimated from the bootstrap clusterings as follows:

First, we define what we mean by a significantly dependent cross cluster within a given AC-clustering. The optimized AC model provides a way of estimating how unlikely a cross cluster is, *given that the margins are independent*. For this purpose, several (1,000 or more) data sets of the same size as the observed one are generated from the marginals of the contingency table (i.e., under the null hypothesis of independence). The cross clusters with the observed amount of data more extreme than that observed by chance with probability 0.01 or less (Bonferroni corrected with the number of cross clusters) are defined to be *significantly dependent cross clusters*.

Next, the two criteria, dependency and reproducibility, will be combined by evaluating how likely it is for each gene pair to occur within the same significantly dependent cross cluster in bootstrap (this is analogous to [29]). The result, interpreted as a similarity matrix, will finally be summarized by hierarchical clustering.

Please note that we do not expect to find dependencies for all genes in the whole data sets, since, with noisy genomic data, that would hardly be possible. In other words, we are interested in finding the most dependent, robust *subsets of the data*. This is exactly what the final gene clusters from bootstrapped, most dependent cross clusters provide.

2.4 Extremity of the Clusters

In the yeast case studies, we evaluate which cross clusters are exceptional by their expression or TF binding profile. For determining the extremity of the observed within-cluster profiles, for each of them, 10,000 random sets of genes were first sampled, each of the same size as the cluster under study. We then computed within-cluster average profiles for the observed cluster as well as for the simulated ones. A part of the observed profile was denoted as extreme if it was lower or higher in value than all the simulations.

3 Reference Methods

First, we need a baseline method to give a lower bound for the results. For AC, it should not optimize the dependency of the clusters, but only perform conventional clustering while being as similar to AC as possible in other respects. In this

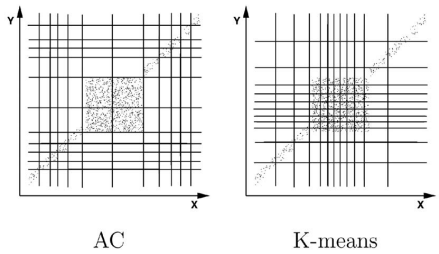


Fig. 2. Associative clustering concentrates on dependent subsets of data. Here, both margin spaces, denoted by **X** and **Y**, are one-dimensional, and the figure shows a scatterplot of the data (dots on the plane where **X** and **Y** are the axes). Cluster borders in the X-space are shown with the vertical lines and cluster borders in the Y-space with horizontal lines. The resulting grid of so-called cross clusters then corresponds to the contingency table; the number of dots within each grid cell gives the amount of data in a contingency table cell. The AC cells are sparse in the bulk of independent data in the middle and denser on the sides where the **X** and **Y** are dependent. K-means, in contrast, focuses on modeling the bulk of the data in the middle. (For this data set, AC has lots of local maxima.)

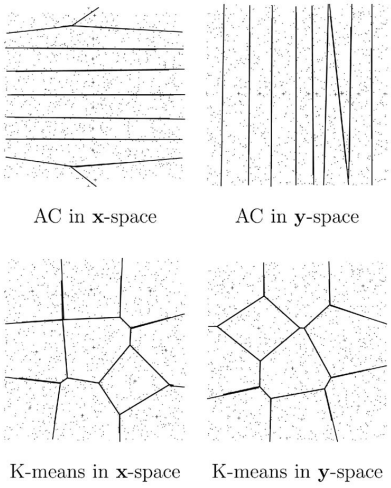


Fig. 3. Associative clustering focuses on modeling the variation that is relevant to dependencies between the data sets. Both of the margin spaces are two-dimensional here, and the data has been constructed such that the vertical dimension of the $\mathbf{x}$-space is dependent on the horizontal dimension in the $\mathbf{y}$-space. All other variation is uniform noise. Lines are approximate cluster borders (Voronoi borders), and the small crosses are the prototype vectors. Associative clustering neglects the irrelevant variation in both margin spaces and models the relevant, dependent variation. In contrast, K-means, as all purely unsupervised clusterings, models all the variation including noise.

work, the baseline method will be independent K-means clusterings in both data spaces, since K-means is also prototype-based clustering for continuous data-like AC. For more detailed description and references of K-means, see, for example, [7].

We compare AC to the information bottleneck (IB) methods [17], [44]. The main problem with IB in our setting is the continuous nature of our data: IB works on nominal-valued data. We discretize the data first by K-means, resulting in a new algorithm called K-IB here. For discrete data, the closest alternative to AC among information bottleneck methods would be symmetric two-way IB [17]. Our sequential implementation is based on [40].

We first quantize the vectorial margins $\mathbf{x}$ and $\mathbf{y}$ separately by K-means without paying attention to possible dependencies between the two margins. This results in two sets of margin partitions which span a large, sparse contingency table that can be filled with frequencies of training data pairs $(\mathbf{x}_k, \mathbf{y}_k)$. The number of elementary Voronoi regions is chosen by using a validation set. In the second phase, the large table is compressed by standard IB to the desired size by aggregating the atomic margin clusters. In this stage, joins at the margins are made with the symmetric sequential algorithm [40] to explicitly maximize the dependency of margins in the resulting smaller contingency table.

The final partitions obtained by the combination of K-means and IB are of a very flexible form and, therefore, the method is expected to model the dependencies of the margin variables well. As a drawback, the final margin clusters will consist of many atomic Voronoi regions, and they are therefore not guaranteed to be particularly homogeneous with respect to the original continuous variables ($\mathbf{x}$ or $\mathbf{y}$). Interpretation of the clusters may then be difficult. Our empirical results support both the good performance of K-IB and the nonlocalness of the resulting clusters.

4 Validation of Associative Clustering

4.1 Demonstration with Artificial Data

Figs. 2 and 3 demonstrate two key properties of AC with artificial data sets that are as simple as possible.

The clusters focus on modeling those regions of the margin data spaces, that is, those subsets of data, where the cooccurring pairs $\mathbf{x}$ and $\mathbf{y}$ are dependent. This is clearly visible as the high-density area of cross clusters in Fig. 2.

AC neglects variation that is irrelevant to the dependencies between $\mathbf{x}$ and $\mathbf{y}$. In Fig. 3, the AC clusters have effectively become defined by only the relevant one of the two dimensions. By contrast, standard clustering methods, such as K-means, model variation in both dimensions.

4.2 Validation of Bootstrapped AC Analysis with Real Data

Especially in bioinformatics, it is often challenging to test new methods since there rarely exists any ground truth, that is, known correct answers. We validated the (bootstrapped) AC approach by searching for dependencies between data sets containing known, real-world duplicate measurements that should be more dependent than random pairs.

Expression profiles of orthologous man-mouse gene pairs with unique LocusIDs were derived from a public source [43] (http://expression.gnf.org/data_public_U95.gz, http://expression.gnf.org/data_public_U74.gz) using the HomoloGene [46] database and Affymetrix annotation files. The expression measurements include 46 human and 45 mouse arrays covering a wide range of tissues and cell-lines. For 21 of the tissues, expression values were available for both species.

We have derived two different data sets from the original data: 1) a larger one for this validation study, with known ground truth in the form of naturally multiplicated genes and 2) a smaller one for the actual analysis without any multiplicated genes (presented in Section 5).

Due to technicalities related to the Affymetrix oligonucleotide array platform, in the original data sets [43], one gene (LocusID) may have multiple expression profiles. In the verification data set, these profiles were considered as independent samples, resulting in a total of 4,500 gene expression profile pairs. These "duplicate orthologous genes," representing the same sequence-level similarity between the species, should cooccur in the significantly dependent cross clusters (see Section 2.3) more often than randomly chosen orthologous genes, and, since AC should model dependencies more effectively than K-means, also more often than in the cross clusters produced by K-means.

The validation study was carried out by exactly the same procedures as we will use in the rest of the experiments of the paper, to validate the setting.

The number of clusters was chosen to be such that each cross cluster would, on average, contain roughly 10 data points. For the verification set, this translates to 19 clusters in both margin spaces. We sampled 100 bootstrap data sets, computed AC for each, got 100 different contingency tables, and, from these, we computed a similarity matrix for the genes as described in Section 2.3.

The optimization parameter σ was chosen by leaving half of the data for validation.

We then tested with a rank sum test whether the similarity distribution of the known duplicates is different from the similarity distribution of all the other genes. In AC, the known duplicates turned out to cooccur unexpectedly frequently in dependent cross clusters (rank sum test; $p < 2.2 \times 10^{-16}$).

Compared to K-means, AC detected connections of the multiple ortholog profiles statistically significantly more often (sign test, $p < 0.001$). These two results support the validity of AC in finding dependent subsets of data better than standard unsupervised clustering.

5 Experimental Results: Dependencies between Man and Mouse

Functions of human genes are often studied indirectly, by studying model organisms such as the mouse. An underlying assumption is that so-called orthologous genes, that is, genes with a common evolutionary origin, have similar functional roles in both species. Exploration of dependencies (regularities and irregularities) in functioning of orthologous genes helps in assessing to which extent this assumption holds. In practice, gene pairs are defined as putative orthologs based on sequence similarity, and we seek for regularities and irregularities in their expression by associative clustering.

An exceptional level of functional conservation of an orthologous gene group may indicate important physiological similarities, whereas differentiation of function may be due to significant evolutionary changes. Large-scale studies on orthologous genes may ultimately lead to a deeper understanding of what makes each species unique. (For related approaches, see, e.g., [6], [9], [11], [14], [30]).

5.1 Data and Experiments

In the original data [43], multiple expression profiles may correspond to one gene. In Section 4.2, they were used for validating the methods, whereas, in this section, we use a single representative profile for each gene. The profiles corresponding to a same gene are averaged after discarding weakly correlating $(r < 0.65)$ profiles of the same gene, when multiple measurements from incomplete or potentially nonspecific probe sets are available. This results in a set of 2,818 orthologous gene pairs with unique LocusIDs.

5.2 Quantitative Comparisons of the Methods

A dependency-maximizing clustering method should 1) find dependencies and 2) represent the results as homogeneous clusters. We compared AC to a baseline method that does not search for dependencies at all, that is, separate K-means for both mouse and man, and to symmetric IB following a discretization with K-means (see Section 3). The both $\sigma : s$ of AC and the number of initial K-means clusters for IB were chosen using a validation set as in Section 4.2.

AC produced significantly more dependent clusters than standard K-means clustering (10-fold cross-validation, paired t-test with $\mathrm{d.f.} = 9$; $p < 0.001$). All methods were run in each fold from three different intializations, of which the best result according to each method's own cost function was selected. Averaged log-BF costs were -52.9 and -115.8 for AC and K-means, respectively. However, cluster homogeneity was not significantly reduced by focusing on dependency modeling (at the $p < 0.05$ significance level). Differences of the methods in cluster homogeneity have been visualized in Fig. 4.

K-IB produced significantly $(p < 0.001)$ more dependent clusterings (log-BF=10.24 on average over cross-validation folds) than AC and K-means. On the other hand, cross clusters from AC studies are significantly more homogeneous than those of K-IB and random clustering $(p < 0.002)$. The measure of homogeneity (actually dispersion) was the sum of intracluster variances.

In summary, as expected, AC extracts more dependencies than K-means and the clusters are more homogeneous (and hence easier to interpret) than those of K-IB. K-IB is a good method for searching for dependencies if homogeneity is not essential.

5.3 Biological Results: Findings of Mice and Men

Bootstrapped AC produces a similarity matrix for the genes, computed from the cooccurrence frequencies of genes in the AC cross clusters. The matrix is summarized with simple hierarchical clustering in this section, and a set of most homogeneous gene clusters is extracted by cutting the dendrogram at a specific cut-off level and discarding genes belonging to clusters smaller than three genes.

As the most reliable dependencies produced by a high cut-off are expected to be relatively trivial findings of similar behavior of orthologous genes in mouse and man, we set the threshold lower to include some unexpected findings as well. The (arbitrary) cut-off limit was set to include clusters with average cooccurrence frequency larger

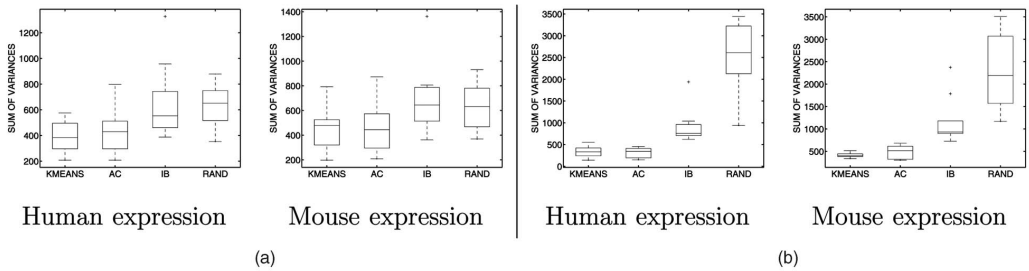


Fig. 4. Dispersion of (a) margin clusters and (b) cross clusters in mouse-man studies. AC produces clusters that are comparable to K-means, whereas the clusters of K-IB are more dispersed (significantly in (b)). RAND is a kind of an upper limit for cluster dispersion, obtained by randomly assigning samples to clusters.

than 80 percent (of the bootstrap samples). This resulted in 139 orthologous gene pairs in 31 clusters.

5.3.1 Overall Regularities in Ortholog Expression

Many orthologous genes are expected to be functionally similar, and similarity can, at its simplest, be measured by correlation. Weak correlation of expression of orthologous genes suggests differentiated gene function (or heavy noise), whereas strong correlation is an indication of functional conservation. To some extent, a global trend exists in our data: Median correlation of expression profiles of orthologous man-mouse gene pairs in the common 21 tissues is 0.33. It is expected that this trend dominates the AC analyses concerning unexpectedly common expression trends (large cross clusters) as well. Indeed, the more similar (highly correlating) the expression profiles of an orthologous gene pair are, the more often it tends to be located in an unexpectedly large cross cluster. This was measured by correlating the occurrence frequency with the correlation between the orthologs, and the resulting correlation coefficient $r = 0.41$ suggests that AC is indeed capable of detecting the simple tendency of the orthologs to depend linearly.

Weakly or negatively correlating orthologs are the other extreme; they are kinds of outliers and tend to be located in exceptionally small cross clusters. Expression similarity correlates negatively $(r = -0.38)$ with frequency of occurrence in small cross clusters.

5.3.2 General Functional Trends of Dependent Genes

Orthologous genes are often functionally similar, although some deviation may have occurred in the course of evolution. Orthologous gene groups with exceptional functional conservation could be expected to be of a specific importance for species survival.

Such a cross-species feature is likely to contribute to dependencies in the data and should be detected in AC analyses. A straightforward approach to study such functional trends is to check enrichment of Gene Ontology (GO) [1] categories among the most dependent genes.

The most enriched GO categories among the genes showing remarkable dependency (average cooccurrence level $\geq 80/100$, minimum cluster size 3) were ribosomal categories (all findings having EASE score with the conservative Bonferroni correction < 0.05 are listed; EASE [23] is a program that annotates the given gene list based on GO and calculates various statistics for it). The three most significantly enriched GOs, for both species, were cellular component categories "cytosolic ribosome (sensu Eukarya)" and "ribosome," and the molecular function category "structural constituent of ribosome." Also, the biological process "transmission of nerve impulse" was enriched for both species. For human, the "eukaryotic 48S initiation complex," "cytosolic small ribosomal subunit (sensu Eukarya)," "small ribosomal subunit," and "synaptic transmission" categories were also enriched.

The dependency structure of data is mostly explained by genes from these categories. A natural explanation for the enrichment of ribosomal functions in large cross clusters is that they often require coordinated effort of a large group of genes and function in cell maintenance tasks that are critical for species survival. High conservation of such genes has been suggested also in earlier studies (see, e.g., [26]). The current result is an additional indication of exceptional conservation of ribosomal genes and of their crucial role for the cellular functions of an organism.

By contrast, enrichment of the "transmission of nerve impulse" category is somewhat surprising and worth more careful studies. It is interesting to note that such genes seem to contribute more to commonalities in the data than genes with other conserved functions. No straightforward biological explanation for this phenomenon could be found so far.

5.3.3 Examples of Finer-Scale Regularities

Minor regularities are revealed by the individual clusters. In addition to conserved expression, AC can potentially reveal orthologs with functional deviation.

We used median correlation as a rough measure to order the clusters and picked two clusters: one with the highest (suggesting preservation of function) and one with the lowest (suggesting differentiation of function) median correlation as examples.

The cluster with the highest median ortholog correlation contained three genes with strongly testis-specific expression (LocusID pairs 8852-11643, 11055-53604, 1618-13164; Fig. 5). Literature studies confirmed that the function of these genes is related to reproduction. Disturbances in the function of the last gene are known to cause infertility although its functions are otherwise not well-known.

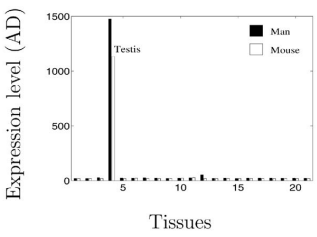


Fig. 5. Average expression profiles of the genes within the cluster showing the highest correlation between mouse and man. Only the 21 tissues which were measured for both species are shown for clarity. No genes were expressed $(AD < 200)$ in the remaining tissues. The tissue list is in the Appendix.

Although the presence of strongly correlated orthologs in the most dependent clusters of the two species is not surprising as such, the strong relationship of the three genes suggests a possibly unknown functional link.

The clusters having salient regularities suggest interactions: The gene products may have physical interaction, they may share a common pathway, or they may otherwise be responsible of similar biological functions. Even correlated expression within a single species is known to be a valuable cue for such interactions (see, e.g., [8], [13], [19]), and preservation of coexpression in evolution is an even stronger hint. Moreover, such "conserved correlations" have also been suggested to be useful in confirming orthologous relationships between genes [15].

Low between-species correlation in a cluster with five genes suggests differentiated gene function (Fig. 6). Three of the genes are known to be related to embryonic development and three are transcription factors. We were not able to find an interpretation for the cluster from the literature. It is reliable, however, and hence potentially interesting; the genes were clustered together in an exceptional cross cluster in over 80 out of 100 bootstrap samples. Our data is from adults, in which the embryonic genes may have unknown functions.

5.3.4 Functionally Exceptional Orthologs

Outliers, that is, genes having peculiarities in their function, can be sought by computing how often they end up in an unexpectedly small cross cluster in the bootstrap. Such genes are comparatively rare; only 1.5 percent of the orthologs end up in an exceptionally small cross cluster with a frequency of ≥ 50 percent. Such exceptional orthologs tend to correlate weakly or negatively, and potentially hint at differentiated gene function. Note that AC takes more than correlation into account as only three of the 43 found orthologs are among the 43 most weakly correlating orthologs. Hence, these exceptional genes could not have been found based on the correlation analysis alone.

Enrichment of certain GO categories among such exceptional orthologs would indicate functionalities that are more often differentiated between species. Interestingly, closest to significant enrichment were the "secretion" category with its subcategory "protein secretion" and the "signal transduction" category with subcategories of "cell communication," "signal transduction," and "cell surface receptor linked signal transduction" for human, and "cell communication" and "G-protein coupled receptor protein signaling pathway" for mouse. These categories have EASE score of < 0.05 without Bonferroni correction. With Bonferroni correction, the enrichment is not significant, however.

To some extent, the secretion categories above could be related to the overall signaling phenomena. The protein secretion category fits well into this picture since many of these signaling pathway initiators are, in fact, secreted molecules. For example, G protein pathways include a variety of extracellular agents like hormones, neurotransmitters, chemokines, and local mediators that are all systemically secreted molecules [33]. From the relative abundance of such orthologs among those with exceptional functionality, we may derive a hypothesis of their role in species divergence.

The most extreme gene (LocusIDs 998 and 12540 for human and mouse, respectively) occurs in an exceptionally small cluster in ≥ 80 of the 100 bootstrap iterations. The expressions in man and mouse correlate negatively (-0.47) in this case and the ortholog is exceptional already as such. The human gene is only expressed in neuronal tissues, whereas the mouse gene is more generally expressed (Fig. 7). Such outliers may be either real functional differences in the species or measurement errors. Whichever the reason, the detection of the outlier was useful.

Groups of orthologous genes with a similar but exceptional functional relationship would be more reliable findings than individual outliers. Unfortunately, cooccurrence of orthologous gene pairs in exceptionally small cross

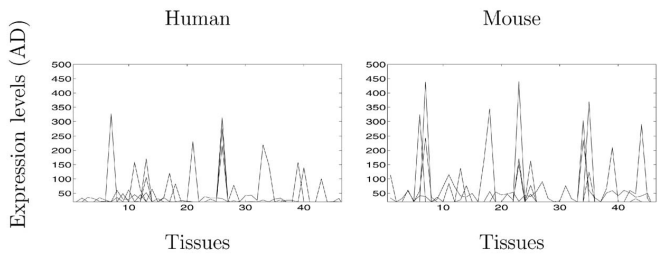


Fig. 6. Expression profile plots of the genes in the cluster with weakest median correlation between the orthologs. Since the correlation is low, no immediate relationships are visible. The cluster is very reliable, however, and hence the orthologs probably share some unexpected higher-order dependency.

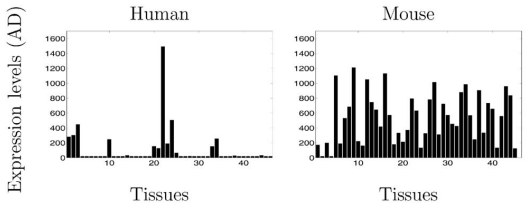


Fig. 7. The most strongly exceptional outlier gene, detected based on its most frequent occurrence in an unexpectedly small cross cluster. LocusIDs 998 and 12540 for human and mouse, respectively.

clusters is rare. The two cases with the most frequent cooccurrence in small cross clusters have a frequency of 45 out of 100 bootstrap iterations. It is interesting to note that, in both cases (Fig. 8), mouse genes are only weakly or not at all expressed in the 21 tissues common to the organisms. In the first case, the mouse and human genes are known to be related to translational regulation. Differences in the expression levels might hint at differentiation in the translational mechanisms. In the second case, the human genes (Protein tyrosine kinase 2 and Glia maturation factor, LocusID-pairs 5747-14083 and 2764-63985) are expressed specifically in neuronal tissues and are known to participate in the regulation of growth and differentiation of neurons.

5.4 Summary

In summary, AC reproduced known findings and performed as expected in comparison with alternative methods. Although this case study is technically interesting and completely new, its biological implications are not yet as convincing as in the second one (Section 6).

From the man-mouse orthologs, we found clusters of highly conserved orthologs, possibly unknown functional relationships between genes, and examples of exceptional relationships between orthologs suggesting differentiation in gene function between species. Some of the findings remain unexplained but could be used as starting points for more detailed studies.

6 Experimental Results: Dependencies between Gene Expression and Transcription Factor Binding

The baker's yeast, *Saccharomyces cerevisiae*, is a popular eukaryotic model organism due to the representativeness of its genetic regulation and because of its easy experimental handling.

Gene expression regulation operates on several levels, of which perhaps the most crucial is transcriptional control. This is handled by a set of regulatory proteins called *transcription factors (TFs)* that bind to DNA in the gene regulatory (promoter) region and can either enhance or suppress the gene's expression. In most cases, TFs interact *inter se* to make up macromolecular complexes before binding to the regulatory regions of DNA. Since TFs are manufactured by expressing the relevant genes, they are the key components of gene interaction networks. In this work, we focus on the dependencies between the TFs and gene expression, that is, on the gene regulatory network.

Regulatory interactions have been studied by measuring genome-wide expression with microarrays in knock-out mutation experiments and in time series experiments. In the knock-out experiments, a mutation is targeted to a single gene in the yeast genome to modify (usually knock out) the normal function of that gene. It is then hoped that, by measuring the gene expression changes with microarrays after the mutation, the role of the mutated gene in cellular processes is revealed. Genes belonging to the same regulatory pathway as the mutated gene could be unveiled, for example. In time series experiments, the goal is often to infer causality in the gene regulatory network based on the sequential changes in expression levels. However, since the interaction network between the genes is complicated, discerning the direct effects of the knock-out or the change of expression in a time series from noise and the mass of second-order effects can be very difficult, if not impossible. At least a comprehensive, very expensive high resolution time-series experiment with numerous replications would be required. The same holds for knock-out experiments. Thus, alternative approaches are worth exploring.

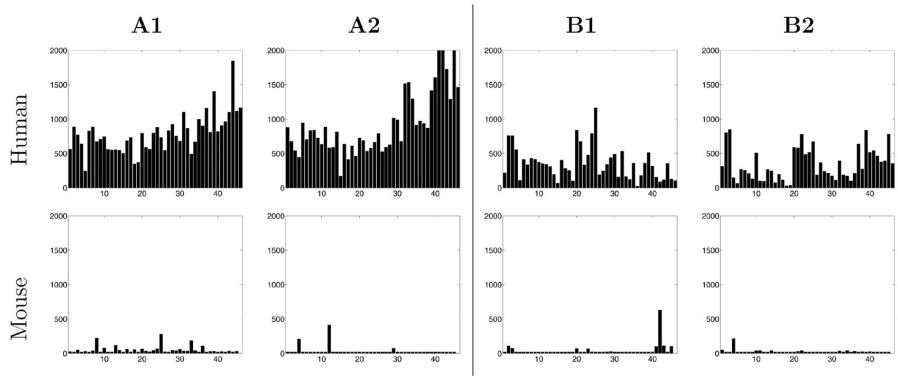


Fig. 8. Two examples (**A** and **B**) of frequently cooccurring and exceptional "clusters" of gene pairs. (They cooccurred frequently in exceptionally small cross clusters). Gene expression profiles belong to human-mouse LocusID pairs **A1** 10438-57316, **A2** 7458-22384 and **B1** 5747-14083, **B2** 2764-63985.

Gene expression is not the only source of information about gene regulation. For instance, microarray-based chromatin immunoprecipitation (ChIP) allows measuring the binding strength of the transcription factor proteins on any gene's promoter region [31]. This reveals which TFs are able to bind the specific gene's promoter and are thus potential regulators. But, many TFs bind numerous gene promoter regions and are still not operational regulators. The number of false positives can be very high and, thus, inferring the regulatory relationships based on the binding information alone is not in general possible.

Combining data from the several sources is a promising option, and exploratory models are perfectly suited for the first studies. We combine the functional information (gene expression) and the potential regulator information (TF binding). We make the following assumptions: First, it is assumed that the genes are coexpressed in groups that are unknown, cf., [16], [36]. Second, it is sensible to assume that a common set of transcription factors binds to the coexpressed genes. Otherwise, groupwise expression would be very unlikely. This is of course an oversimplification, but it has some biological justification. To be more realistic, we do not assume that all the genes are regulated in such a manner; we relax the simplification by assuming that only *subsets* of genes behave this way, only *a subset* of transcription factors need to be the same, and coexpression needs to take place only in *a subset* of knock-out experiments or time points.

Associative clustering, when applied to expression and TF binding data, makes precisely these assumptions, and we now aim to find subsets of genes whose expression is maximally dependent on their transcription factor binding profiles. These sets then act as hypotheses for expression coregulation.

6.1 Knock-Out Expression and TF Binding

The yeast expression used in this analysis has been measured from 300 different mutation strains with cDNA microarrays [25] (http://www.rii.com/publications/2000/cell_hughes.html). Transcription factor binding data on genes for 113 transcription factors was obtained from [31] (http://web.wi.mit.edu/young/regulator_network). After taking the logarithm of the expression ratios, imputing missing values with genewise averages, standardizing the treatmentwise variances to unity, and including only the genes appearing in both data sets, we had two full data matrices, each with 6,185 genes. The number of clusters in the margin spaces was chosen to produce roughly 10 data points in each cross cluster, resulting in 30 clusters in the expression space and 20 clusters in the TF-binding space.

6.1.1 Quantitative Evaluation

We first used this data to validate the performance of AC in the two tasks it addresses: maximizing the dependency and keeping the clusters homogeneous. These were measured in 10-fold cross-validation runs with prevalidated σ for AC and prevalidated number of K-means clusters for K-IB. Prevalidation was analogous for both methods: The data was divided into two equally sized parts and several parameter values were tried from three different random initializations. Of these, the parameter value giving the best AC cost was chosen. The final cross-validation runs were also started from three different random initializations.

AC discovered dependencies in the data significantly better than the reference methods (10-fold cross-validation, paired t-test; $d.f. = 9;\ p < 0.001$). The dependency was measured with (natural) logarithmic Bayes factor (log-BF), the average value of which was 8.84 for AC, -46.37 for IB, and -262.29 for K-means. The value of log-BF is traditionally interpreted to signify strong evidence against the null hypothesis if it is at least 6-10 [28].

The homogeneity, or actually dispersion, of the clusters was measured simply by the sum of the componentwise variances in cross-validation. The comparison was made for both margin clusters as well as for cross clusters. Margin clusters produced by AC were statistically significantly less dispersed than those produced by IB, but for cross clusters the difference was not significant.

6.1.2 Biological Results

We sought for biologically interesting findings by bootstrapping the AC (100 bootstrap data sets) and by otherwise using the same parameters as in the above cross-validation tests. A similarity matrix was generated for the genes from the bootstrap results (see Section 4.2) and summarized by the average-distance variant of hierarchical clustering. Clusters with average cooccurrence higher than 20 out of 100 and with the minimum size of 3 genes were chosen for the final analysis, resulting in 20 clusters.

The clusters were first screened with EASE, which found enriched gene ontology classes in 12 of the 20 clusters (Fisher's exact test, Bonferroni corrected; $p < 0.05$). It is of course likely that clusters without significant GO enrichments are also biologically meaningful, but their interpretation is more cumbersome and is therefore left for future work. In the following, we present a sample of four representative AC cluster types.

The first, most notable cluster is a large set of about one hundred genes that all code for ribosomal proteins. These genes are known to be expressed often very homogeneously, and they can also often be found in conventional cluster analyses, cf. [5], [34].

The next two clusters are examples of how AC identifies and highlights modules where a subset of the genes and their main regulator(s) have been previously identified in wet lab experiments. However, the modules also contain novel components not previously associated to the corresponding biological function.

The second cluster is an example of a cluster type rarely found in conventional analyses. It contains only four genes, of which three are known to code for proteins involved in lipid metabolism and one to code for a growth factor transporter. The most reliable and strongest transcription factor bindings in this cluster are by proteins INO2/YDR123Cp and INO4/YOL108Cp that are known to form a protein complex and then regulate lipid metabolism. The fact that AC detects two interacting TFs shows that the method can be used, to a certain extent, to predict TF interactions as well. Moreover, it also unveils which potential target genes are responsible for the lipid metabolism regulation observed in wet lab experiments. In other words, the reliability of gene function annotations is enhanced through the use of AC.

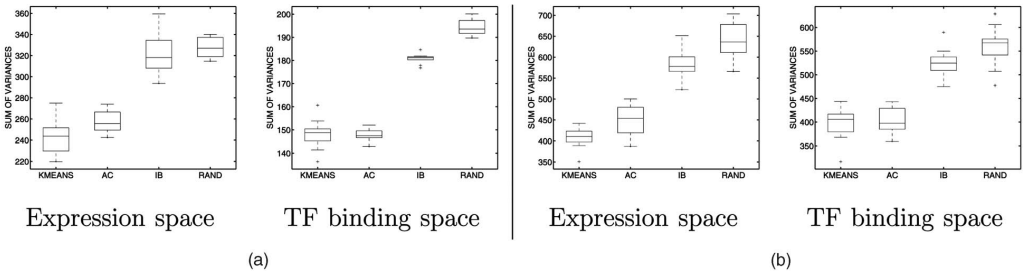


Fig. 9. (a) Margin cluster and (b) cross cluster dispersion for all methods in cell-cycle experiments, demonstrating that AC produces clusters that are almost as compact as K-means clusters, whereas the IB-clusters are significantly more dispersed. RAND is a kind of an upper limit for cluster dispersion, obtained by randomly assigning samples to clusters.

The third cluster of 31 genes contains 20 genes involved in amino acid and derivative metabolism. The best identified regulator for this cluster is GCN4/YEL009Cp, a transcriptional activator of amino acid biosynthetic genes known to respond to amino acid starvation. Here again, it is shown that the AC creates a partially new cluster and identifies a good candidate regulator.

About two thirds (28) of the genes in the fourth cluster, the most interesting so far, are of unknown molecular function. Even the biological process they contribute to may be unknown. The known genes map to such GO categories as "nuclear organization and biogenesis" and the most reliable transcription factor associated to genes in this cluster was YAP5p/YIR018Wp. This transcription factor is known to be activated by the main regulators (SBF and MBF [22]) of the START of the cell cycle, a time just before DNA replication. This clearly refers to cell-cycle regulation and to organization of the nucleus prior to replication.

6.2 Time Series Gene Expression and TF Binding

The expression data for this case study was measured during yeast cell cycle and was originally published in two different papers [10], [42] (http://genome-www.stanford.edu/cellcycle/links.html). The data consisted of 77 timepoints in total. The transcription factor binding data used here is the updated (2003) version of [31] for 106 transcription factors. In this case study, the missing values were imputed with the k-nearest neighbor method $(k = 10)$ [45] and logarithms were taken from both of the data sets. Including only the genes present in both data sets resulted in a total of 5,618 genes. The chosen cluster numbers were 30 in the expression space and 20 in the TF-binding space.

6.2.1 Numerical Results

The tests were run as described in Section 6.1. The differences in dependency modeling between all the methods were statistically significant also for this data pair (10-fold cross-validation, paired t-test; $\mathrm{d.f.} = 9$; $p < 0.001$). Natural logarithmic Bayes factor for AC was 32.27, for IB -13.17, and for K-means -92.30, implying that AC found a very strong dependency between the data sets.

The measure of cluster homogeneity, or actually dispersion, was the same as in the previous cases: the sum of the componentwise variances. For this data pair, AC produced significantly (10-fold cross-validation, paired t-test; $\mathrm{d.f.} = 9$; $p < 0.001$) less dispersed cross clusters and margin clusters than IB. Fig. 9 visualizes the margin cluster and cross cluster dispersion for all methods.

6.2.2 Biological Results

In a similar manner as in the previous case, we sought for biological findings from the bootstrapped AC clusters. The clusters with average distance smaller than 60 (times in the same dependent cross cluster out of 100) and with more than two genes were chosen. This resulted in a total of 16 clusters.

Gene ontology classes were enriched statistically significantly in 13 of the 16 clusters (EASE; Fisher's exact test, Bonferroni corrected; $p < 0.05$). In the similar spirit as in the knock-out mutation case, we give a representative sample of four clusters.

Two clusters are essentially the same as in the in knock-out case study, the ribosomal proteins being the first of them.

The second cluster is the same as the most interesting (fourth) cluster in the knock-out case. This provides more evidence that the cluster represents a biologically robust motif, having a homogeneous profile in both TF-binding and expression.

The third cluster (Fig. 10) contains a significantly high number of genes involved in cell cycle regulation and, more specifically, at the stage of entry into the mitotic cell cycle (nine genes out of 33). The main regulator identified in this module is SIP4p/YJL089Wp which is possibly involved in SNF1p/YDR477Wp-regulated transcriptional activation. This latter signaling factor is required for transcription in response to glucose limitation. Interestingly, SIP4p/YJL089Wp has a DNA-binding domain similar to the GAL4p/YPL248Cp transcription factor, involved in galactose response, another route in energy metabolism. Taken together, this cluster contains some clear references to cell cycle regulation on one hand and energy metabolism on the other and proposes a set of genes that can bridge and connect these two biological processes. Thereby, AC offers the hypothesis for a relation between biological functions, in addition to some clues on what genes could be involved.

The fourth cluster contains nine genes of unknown molecular function or associated biological process. The associated transcription factor ACE2p/YLR131Cp is known to activate expression of early G1-specific genes, localizes to daughter cell nuclei after cytokinesis, and there delays G1 progression in the daughters. Based on this data, the nine genes can be predicted to act during the G1 phase of the cell-cycle, thus specifying what kind of targeted experiments are needed to establish their function.

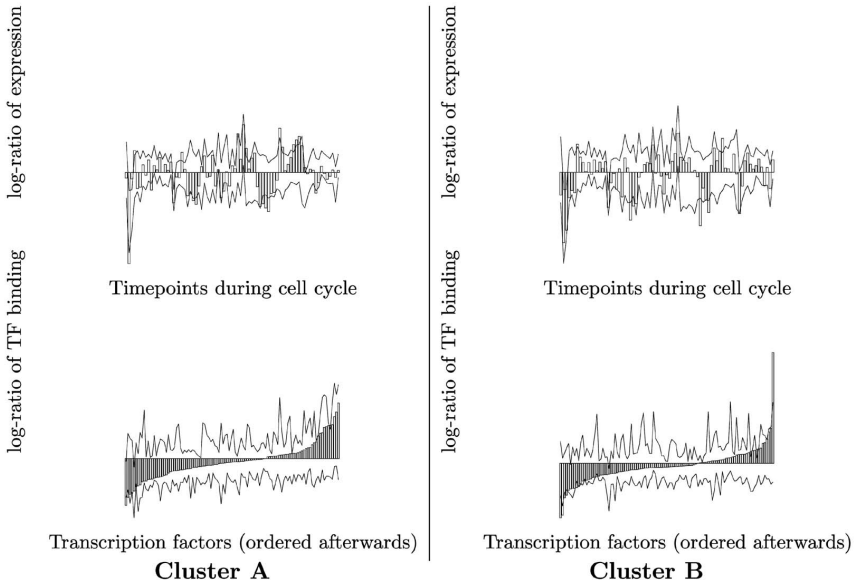


Fig. 10. Two examples of bootstrapped cross clusters, associated to cell cycle, that reveal both known and novel dependencies between gene expression and TF binding. The upper figures show the average expression profiles (bars) of the clusters and confidence intervals (curves). The periodicity of the cell cycle in the expression is clearly visible. The lower figures show the average TF-binding profile of the clusters with confidence intervals. The average TF-bindings rising above the confidence interval are considered reliable. Note that the confidence intervals are very conservative; they have been estimated based on random clusters. In **Cluster A**, there was only one reliable TF binding, SIP4. It could be verified from the literature (see text for details). SIP4 binds also the genes in **Cluster B**, but, additionally, there is one extremely strongly binding TF, SFL1 (the rightmost bar). Its putative regulatory interaction with the gene cluster during cell cycle is a new finding.

7 Conclusion and Future Work

We have introduced a new approach for a relatively little-studied machine learning or data mining problem: From data sets of cooccurring samples, find what is in common. We have formulated the problem probabilistically, extending earlier mutual information-based approaches. The new solution is better-justified for finite (relatively small) data sets.

The introduced method, coined associative clustering (AC), summarizes dependencies between data sets as clusters of similar samples having similar dependencies. Such a method is particularly needed for mining functional genomics data where measurements are available about different aspects of the same set of functioning genes. Then, a key challenge is to find commonalities between the measurements. The answer should reveal characteristics of the genes, not only characteristics of the measurement setups.

The work is pure machine learning in the sense that the model is a general-purpose semiparametric model which learns to fit a new data set instead of being manually tailored. As a result, it is probably not as accurate as more specific models, but it can be expected to be faster and easier to apply to new problems. Its main intended application area is in exploratory data analysis, "looking at the dependencies in the data" in the first stages of a research project.

The method was validated and applied in two functional genomics studies. The first found regularities and differences between the functioning of orthologous genes in different organisms, suggesting evolutionary conservation and divergence. The second explored regulatory interactions between gene expression and transcription factor binding. Both trivial and unexpected findings were made: known regularities, outliers, and hints about unexpected regularities.

While the proposed method was shown to be viable already as such, it can be further improved. We did not address the problem of choosing an optimal number of clusters. If clustering is interpreted as a partitioning or quantization of data to compress its presentation, then the exact number of clusters is not a crucial parameter, but nevertheless, the results could be improved by optimizing it. Since the task is formulated in Bayesian terms, Bayesian complexity control methods are applicable in principle. The setting is not standard, however, because of the nonstandard (new) use of the Bayes factors and because of discontinuities in the objective function.

Another direction of improvement is regularization of the solution. Dependency-searching methods may potentially overfit the data, which is well-known from canonical correlation analysis and can be avoided by regularization. We have developed two regularization methods for AC with one fixed margin. "Entropy regularization" was used here because it is easier in practice and has not been shown to be worse than the alternative [27]. In the present case, bootstrap also helped. Another related question is which kinds of priors to use for the distributional parameters. The simple constant Dirichlet priors used in this work may be too informative. Hierarchical modeling should be more appropriate but it is computationally more complex.

A third area worth investigating is the parameterization of the clusters. It should be investigated whether the hard Voronoi regions, used up to now because they are easily

interpretable and make the theory manageable, could be replaced by smooth and more regular-sized clusters. Alternatively, the degrees of freedom of the clusterings could be directly reduced to regularize the solution.

Finally, a comprehensive comparison of the relative merits of dependency maximization and more traditional Bayes networks and graphical models of the whole joint distribution should be carried out. It is clear that the two approaches focus on different properties of data and that our semiparametric models need less prior knowledge than specialized models of gene regulation, for instance, and are hence more general-purpose. We expect that exploratory models of the type introduced here are viable as complementary methods for gathering the necessary prior knowledge for the more specific models.

APPENDIX

TISSUES IN MOUSE-HUMAN DATA

The first 21 tissues are considered to be common for both species. (Listed in the following order: tissue number: human tissue: mouse tissue. Tissues are separated with commas.)

Common tissues: 1: cerebellum: cerebellum, 2: cortex: cortex, 3: amygdala: amygdala, 4: testis: testis, 5: placenta: placenta, 6: thyroid: thyroid, 7: prostate: prostate, 8: ovary: ovary, 9: uterus: uterus, 10: 0DRG: 0DRG, 11: salivary gland: salivary gland, 12: trachea: trachea, 13: lung: lung, 14: thymus: thymus, 15: spleen: spleen, 16: adrenal gland: adrenal gland, 17: kidney: kidney, 18: liver: liver, 19: heart: heart, 20: caudate nucleus: striatum, 21: spinal cord: spinal cord lower.

Noncommon tissues: 22: fetal brain: digits, 23: whole brain: gall bladder, 24: thalamus: hippocampus, 25: corpus callosum: large intestine, 26: pancreas: adipose tissue, 27: pituitary gland: lymph node, 28: prostate cancer: eye, 29: OVR278E: skeletal muscle, 30: OVR278S: snout epidermis, 31: fetal liver: tongue, 32: HUVEC: trigeminal, 33: THY+: bladder, 34: THY-: small intestine, 35: myelogenous k-562: stomach, 36: lymphoblastic molt-4: hypothalamus, 37: burkitts Daudi: epidermis, 38: bukitts Raji: spinal cord upper, 39: hep3b: bone, 40: A2058: brown fat, 41: DOHH2: olfactory bulb, 42: GA10: mammary gland, 43: HL60: umbilical cord, 44: K422: bone marrow, 45: ramos: frontal cortex, 46: WSU: -.

ACKNOWLEDGMENTS

The authors would like to thank Jaakko Peltonen for the code for IB. This work has been supported by the Academy of Finland, decisions #79017 and #207467. During part of the work, S. Kaski and J. Nikkilä were with Helsinki University of Technology.

REFERENCES

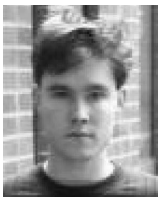
[1] M. Ashburner et al., "Gene Ontology: Tool for the Unification of Biology," *Nature Genetics,* vol. 25, pp. 25-29, 2000.
[2] M.S. Bazaraa, H.D. Sherali, and C.M. Shetty, *Nonlinear Programming: Theory and Algorithms.* New York: Wiley, 1993.
[3] S. Becker, "Mutual Information Maximization: Models of Cortical Self-Organization," *Network: Computation in Neural Systems,* vol. 7, pp. 7-31, 1996.
[4] S. Becker and G.E. Hinton, "Self-Organizing Neural Network that Discovers Surfaces in Random-Dot Stereograms," *Nature,* vol. 355, pp. 161-163, 1992.
[5] M. Beer and S. Tavazoie, "Predicting Gene Expression from Sequence," *Cell,* vol. 117, pp. 185-198, 2004.
[6] S. Bergmann, J. Ihmels, and N. Barkai, "Similarities and Differences in Genome-Wide Expression Data of Six Organisms," *PLoS Biology,* vol. 2, pp. 85-93, 2004.
[7] C.M. Bishop, *Neural Networks for Pattern Recognition.* New York: Oxford Univ. Press, 1995.
[8] H. Bono and Y. Okazaki, "Functional Transcriptomes: Comparative Analysis of Biological Pathways and Processes in Eukaryotes to Infer Genetic Networks among Transcripts," *Current Opinion in Structural Biology,* vol. 12, pp. 355-361, 2002.
[9] S.B. Carroll, "Genetics and the Making of Homo Sapiens," *Nature,* vol. 422, pp. 849-857, 2003.
[10] R.J. Cho, M.J. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L. Wodickaa, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman, D.J. Lockhart, and R.W. Davis, "A Genome-Wide Transcriptional Analysis of the Mitotic Cell Cycle," *Molecular Cell,* vol. 2, pp. 65-73, 1998.
[11] A.G. Clark et al., "Inferring Nonneutral Evolution from Human-Chimp-Mouse Orthologous Gene Trios," *Science,* vol. 302, pp. 1960-1963, 2003.
[12] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap.* New York: Chapman & Hall, 1993.
[13] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein, "Cluster Analysis and Display of Genome-Wide Expression Patterns," *Proc. Nat'l Academy of Sciences, USA,* vol. 95, pp. 14863-14868, 1998.
[14] W. Enard et al. "Intra- and Inter-Specific Variation of Primate Gene Expression Patterns," *Science,* vol. 296, pp. 340-343, 2002.
[15] R.M. Ewing and J.-M. Claverie, "EST Databases as Multi-Conditional Gene Expression Datasets," *Proc. Pacific Symp. Biocomputing,* vol. 5, pp. 427-439, 2000.
[16] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian Networks to Analyze Expression Data," *J. Computational Biology,* vol. 7, pp. 559-584, 2000.
[17] N. Friedman, O. Mosenzon, N. Slonim, and N. Tishby, "Multivariate Information Bottleneck," *Proc. 17th Conf. Uncertainty in Artificial Intelligence (UAI),* pp. 152-161, San Francisco: Morgan Kaufmann, 2001.
[18] V. Ganti, J. Gehrke, R. Ramakrishnan, and W.-Y. Loh, "A Framework for Measuring Changes in Data Characteristics," *Proc. 18th ACM Symp. Principles of Database Systems,* pp. 126-137, 1999.
[19] H. Ge, Z. Liu, G.M. Church, and M. Vidal, "Correlation between Transcriptome and Interactome Mapping Data from Saccharomyces Cerevisiae," *Nature Genetics,* vol. 29, pp. 482-486, 2001.
[20] I.J. Good, "On the Application of Symmetric Dirichlet Distributions and Their Mixtures to Contingency Tables," *Annals of Statistics,* vol. 4, pp. 1159-1189, 1976.
[21] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning.* New York: Springer, 2001.
[22] C.E. Horak, N.M. Luscombe, J. Qian, P. Bertone, S. Piccirrillo, M. Gerstein, and M. Snyder, "Complex Transcriptional Circuitry at the G1/S Transition in Saccharomyces Cerevisiae," *Genes and Development,* vol. 16, pp. 3017-3033, 2002.
[23] D. Hosack, G. Dennis Jr., B. Sherman, H. Lane, and R. Lempicki, "Identifying Biological Themes within Lists of Genes with EASE," *Genome Biology,* vol. 4, p. R70, 2003.
[24] H. Hotelling, "Relations between Two Sets of Variates," *Biometrika,* vol. 28, pp. 321-377, 1936.
[25] T.R. Hughes et al. "Functional Discovery via a Compendium of Expression Profiles," *Cell,* vol. 102, pp. 109-126, 2000.
[26] J.L. Jiménez, M.P. Mitchell, and J.G. Sgouros, "Microarray Analysis of Orthologous Genes: Conservation of the Translational Machinery across Species at the Sequence and Expression Level," *Genome Biology,* vol. 4, p. R4, 2002.
[27] S. Kaski, J. Sinkkonen, and A. Klami, "Discriminative Clustering," *Neurocomputing,* to appear.
[28] R.E. Kass and A.E. Raftery, "Bayes Factors," *J. Am. Statistical Assoc.,* vol. 90, pp. 773-795, 1995.
[29] M.K. Kerr and G.A. Churchill, "Bootstrapping Cluster Analysis: Assessing the Reliability of Conclusions from Microarray Experiments," *Proc. Nat'l Academy of Sciences,* vol. 98, pp. 8961-8965, 2001.

[30] P. Khaitovich, G. Weiss, M. Lachmann, I. Hellmann, W. Enard, B. Muetzel, U. Wirkner, W. Ansorge, and S. Pääbo, "A Neutral Model of Transcriptome Evolution," *PLoS Biology*, vol. 2, pp. 0682-0689, 2004.
[31] T.I. Lee et al., "Transcriptional Regulatory Networks in Saccharomyces Cerevisiae," *Science*, vol. 298, pp. 799-804, 2002.
[32] G.J. McLachlan, K.-A. Do, and C. Ambroise, *Analyzing Microarray Gene Expression Data*. New York: Wiley, 2004.
[33] S.R. Neves, P.T. Ram, and R. Iyengar, "G Protein Pathways," *Science*, vol. 296, pp. 1636-1639, 2002.
[34] J. Nikkilä, P. Törönen, S. Kaski, J. Venna, E. Castrén, and G. Wong, "Analysis and Visualization of Gene Expression Data Using Self-Organizing Maps," *Neural Networks*, special issue on new developments on self-organizing maps, vol. 15, pp. 953-966, 2002.
[35] J. Peltonen, J. Sinkkonen, and S. Kaski, "Sequential Information Bottleneck for Finite Data," *Proc. 21st Int'l Conf. Machine Learning*, pp. 647-654, 2004.
[36] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman, "Module Networks: Identifying Regulatory Modules and their Condition-Specific Regulators from Gene Expression Data," *Nature Genetics*, vol. 34, pp. 166-176, 2003.
[37] J. Sinkkonen and S. Kaski, "Clustering Based on Conditional Distributions in an Auxiliary Space," *Neural Computation*, vol. 14, pp. 217-239, 2002.
[38] J. Sinkkonen, S. Kaski, J. Nikkilä, and L. Lahti, "Associative Clustering (AC): Technical Details," Technical Report A84, Publications in Computer and Information Science, Laboratory of Computer and Information Science, Helsinki Univ. of Technology, 2005.
[39] J. Sinkkonen, J. Nikkilä, L. Lahti, and S. Kaski, "Associative Clustering," *Proc. 15th European Conf. Machine Learning*, pp. 396-406, 2004.
[40] N. Slonim, "The Information Bottleneck: Theory and Applications," PhD thesis, Hebrew Univ., 2002.
[41] N. Slonim, N. Friedman, and N. Tishby, "Unsupervised Document Classification Using Sequential Information Maximization," *Proc. 25th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval*, pp. 129-136, ACM Press, 2002.
[42] P. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher, "Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast Saccharomyces Cerevisiae by Microarray Hybridization," *Molecular Biology of the Cell*, vol. 9, pp. 3273-3297, 1998.
[43] A.I. Su et al., "Large-Scale Analysis of the Human and Mouse Transcriptomes," *Proc. Nat'l Academy of Sciences, USA*, vol. 99, pp. 4465-4470, 2002.
[44] N. Tishby, F.C. Pereira, and W. Bialek, "The Information Bottleneck Method," *Proc. 37th Ann. Allerton Conf. Comm., Control, and Computing*, pp. 368-377, 1999.
[45] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R.B. Altman, "Missing Value Estimation Methods for DNA Microarrays," *Bioinformatics*, vol. 17, pp. 520-525, 2001.
[46] D.L. Wheeler et al., "Database Resources of the National Center for Biotechnology," *Nucleic Acids Research*, vol. 31, pp. 28-33, 2003.

Samuel Kaski (M'96–SM'02) received the DSc (PhD) degree in computer science from the Helsinki University of Technology, Espoo, Finland, in 1997. He is currently professor of computer science at the University of Helsinki, Finland. His main research areas are statistical machine learning and data mining, bioinformatics, and information retrieval. He is a senior member of the IEEE.

Janne Nikkilä received the MSc (Tech) degree from the Helsinki University of Technology, Espoo, Finland, in 1999. He is currently finalizing his PhD thesis about exploratory clustering analysis methods applied to genomic data sets at the Laboratory of Computer and Information Science (Neural Networks Research Centre), Helsinki University of Technology, and is also partly affiliated with the University of Helsinki.

Janne Sinkkonen received the MA degree in psychology from the University of Helsinki in 1996 and the PhD degree in machine learning from the Helsinki University of Technology (HUT) in 2004. He has worked as a researcher in a Helsinki University brain research group during the 1990s and as a researcher at the HUT Neural Networks Research Centre during the 2000s. He is currently at Xtract Ltd.



Leo Lahti received the MSc (Tech) degree from the Helsinki University of Technology, Espoo, Finland, in 2003. He is currently a postgraduate researcher at the Laboratory of Computer and Information Science (Neural Networks Research Centre), Helsinki University of Technology, focusing on research and development of data analysis methods for bioinformatics.

Juha E.A. Knuuttila is a graduate student at the University of Jyväskylä, Finland, where he is pursuing the masters degree in molecular biology. For the doctoral thesis, he has started studying the plasticity of the adult brain in different research paradigms by measuring the alterations of gene expression and the activity status of some plasticity related proteins at the Neuroscience Center, University of Helsinki, Finland.

Christophe Roos graduated in genetics and mathematics (1982) at the University of Helsinki, Finland, whereafter he performed a PhD thesis in molecular biology (1986) at the University of Strasbourg, France. After having directed a Drosophila developmental biology research group at the University of Helsinki, he joined Medicel Ltd. (2000), a company developing a systems biology software platform. His principal scientific interests proceed from the use of bioinformatics in developmental biology.