KUOPION YLIOPISTON JULKAISUJA G. - A.I. VIRTANEN -INSTITUUTTI 55
KUOPIO UNIVERSITY PUBLICATIONS G.
A.I. VIRTANEN INSTITUTE FOR MOLECULAR SCIENCES 55

PETRI PEHKONEN

Methods for Mining Data from Genome Wide High-Throughput Technologies

Doctoral dissertation

To be presented by permission of the Faculty of Business and Information Technology of the University of Kuopio for public examination in Auditorium, Tietoteknia building, University of Kuopio, on Saturday 10th November 2007, at 12 noon

Department of Neurobiology
A.I. Virtanen Institute for Molecular Sciences
Department of Computer Science
Faculty of Business and Information Technology
University of Kuopio

KUOPION YLIOPISTO

KUOPIO 2007

Distributor: Kuopio University Library
P.O. Box 1627
FI-70211 KUOPIO
FINLAND
Tel. +358 17 163 430
Fax +358 17 163 410
http://www.uku.fi/kirjasto/julkaisutoiminta/julkmyyn.html

Series Editors: Research Director Olli Gröhn, Ph.D.
Department of Neurobiology
A.I. Virtanen Institute for Molecular Sciences

Research Director Michael Courtney, Ph.D.
Department of Neurobiology
A.I. Virtanen Institute for Molecular Sciences

Author's address: Department of Biosciences
University of Kuopio
P.O. Box 1627
FI-70211 KUOPIO
FINLAND
Tel. +358 44 366 8027
Fax +358 17 281 1510
E-mail: petri.pehkonen@uku.fi

Supervisors: Research Director Erkki Pesonen, Ph.D.
Department of Computer Science
University of Kuopio

Professor Gary Wong, Ph.D.
Department of Biosciences
University of Kuopio

Dr. Petri Törönen
Institute of Biotechnology
University of Kuopio

Reviewers: Docent Sampsa Hautaniemi, DTech
Computational Systems Biology Laboratory
Institute of Biomedicine and Genome-Scale Biology Research Program
Faculty of Medicine
University of Helsinki

Academy Professor Heikki Mannila, Ph.D.
Helsinki Institute for Information Technology
Helsinki University of Technology and University of Helsinki

Opponent: Professor Tapio Salakoski, Ph.D.
Department of Information Technology
University of Turku

ISBN 978-951-27-0614-3
ISBN 978-951-27-0436-1 (PDF)
ISSN 1458-7335

Kopijyvä
Kuopio 2007
Finland

Pehkonen, Petri, J. Methods for mining data from genome wide high throughput technologies. Kuopio University publications G. A.I. Virtanen Institute for Molecular Sciences 55. 2007. 91 p.
ISBN 978-951-27-0614-3
ISBN 978-951-27-0436-1 (PDF)
ISSN 1458-7335

Abstract

Current high-throughput technologies, like DNA-microarrays, produce measurement data concerning the structure and function of cellular molecules in a genome wide manner. Analysis of such data requires using efficient and robust computational tools. The standard output from microarray analysis is a set of genes which are co- or differentially expressed. Biological interpretation of such outcome aims at finding the mechanisms that cause such expression. This step often involves searching the biological databases and literature for biological attributes that are over-represented in the gene set. Recent methods and software programs use statistical approaches for finding such information. Still, there remain many questions which they do not address.

This work presents novel bioinformatic methods and software tools for biological interpretation of data obtained from high throughput technologies. The presented methods 1) discover expected relations of genes and experimental conditions by literature mining, 2) discover biological processes which can explain the co- or differential expression by using cluster analysis of functional information on genes, 3) discover putative regulatory elements which can explain the genes' co-expression, and 4) find the chromosomal locations with enrichment of co-expressed genes by using a segmentation procedure.

Methods presented in this work analyze categorical data representing the associations between genes and biological attributes. The methods include clustering and segmentation, and statistical evaluation of such results. For clustering of high dimensional binary data, we present a method based on *Non-negative Matrix factorization* (NMF). This recent matrix factorization method has shown good performance in the analysis of binary data. In segmentation, we apply heuristics in order to obtain results in reasonable time. As clustering and segmentation produce several solutions with different numbers of clusters, we show novel methods for results evaluation.

The developed methods outperform the alternatives in comparisons performed by using real and simulated data. The methods are applied to interpretation of several different datasets. These include gene expression data obtained from salmon fish under the treatment of environmental toxins, baker's yeast during cell cycle and under the influence of antifungal drug, and nematode including human Parkinson’s Disease related transgene.

Universal Decimal Classification: 575.111, 575.112
National Library of Medicine Classification: QU 26.5, QU 58.5, QU 450, QU 470
Medical Subject Headings: Computational Biology; Information Storage and Retrieval; Genes; Genome; Genomics; Gene Expression Profiling; Databases, Genetic; Microarray Analysis; Cluster Analysis; Transcription Factors; Bayes Theorem; Factor Analysis, Statistical; Models, Statistical

For Anni

Acknowledgements

This work has been carried out in the Institute of Applied Biotechnology (publication I), Department of Neurobiology of A. I. Virtanen Institute for Molecular Sciences (publications II-V), and the Bioscience Department at the University of Kuopio (publication V, thesis overview), during years 2004-2007. Sources of funding include projects supported by TEKES, the Academy of Finland, Pohjois-Savo Regional Fund of the Finnish Cultural Foundation, A.I. Virtanen Graduate School for Molecular Medicine, Emil Aaltonen Foundation, and the Department of Computer Science at the University of Kuopio.

I am very grateful to my supervisors, research director Erkki Pesonen, Ph.D., professor Garry Wong, Ph.D., and Ph.D. Petri Törönen for their guidance. Thanks for Garry and Petri for being also close workmates, advisors, and contributors in the included publications. Thanks also for Martti Penttonen for giving helpful instructions.

Thanks to co-authors: Ph.D. Merja Lakso, Ph.D. Suvi Vartiainen, Ph.D. Liisa Holm, M.Sc. Päivi Rösenstöm, M.Sc. Matti Kankainen, M.Sc. Heikki Koskinen, Ph.Lic. Hannu Mölsä, M.Sc. Eeva Vehniainen, Ph.D. Richard Nass, Ph.D. Caird Rexroad, Ph.D. Sergei Afanasyev, and Ph.D. Aimo Oikari. Special thanks to my workmate and advisor, Docent Aleksei Krasnov, for introducing me the world of salmon fish and microarrays. Thanks to my workmates Jussi Paananen, Mitja Kurki, Suvi Asikainen, Vuokko Aarnio, Markus Storvik, James Callaway, Kaja Reisner and Jani Kekäläinen. Thanks to my friends and everybody I have forgot to thank here.

Special thanks to the pre-reviewers of this thesis, Docent Sampsa Hautaniemi (DTech) and Professor Heikki Mannila (Ph.D.), whose comments helped in improving this work.

Very special thanks to my lovely wife Anni for her love, and my children Aatu, Iina and Unna, for tolerating me during preparing the thesis. Thanks to my mother and father for giving birth to me and raising me. Thanks to my three sisters Tarja, Päivi and Kirsi for big support.

List of abbreviations and symbols

AIC	Akaike information criterion
ANOVA	Analysis of variance
BF	Bayes factor
BIC	Bayesian information criterion
CBS	Circular binary segmentation
cDNA	Complementary DNA
CGH	Comparative genome hybridization
CREB	cAMP response element binding protein
CSP	Class specific prior
DEM	Data explaining model
DGM	Data generating model
DNA	Deoxyribonucleic acid
EBP	Empirical Bayes prior
EST	Expressed sequence tag
GO	Gene Ontology
GSEA	Gene set enrichment analysis
HD	High dimensional
HMM	Hidden Markov model
ICA	Independent component analysis
KEGG	Kyoto encyclopedia of genes and genomes
LR	Likelihood ratio
MAP	Maximum a posteriori
MDL	Minimum description length
MDS	Multidimensional scaling
MGED	Microarray gene expression group
MEBP	Modified empirical Bayes prior
MI	Mutual information
MIAME	Minimum information about microarray experiment
ML	Maximum likelihood
MLE	Maximum likelihood estimator
MLR	Maximum likelihood ratio
MM	Mismatch
mRNA	Messenger RNA
NHST	Null hypothesis significance testing
NMF	Non-negative matrix factorization
PCA	Principal component analysis
PM	Perfect match
RNA	Ribonucleic acid
SNP	Single nucleotide polymorphism
SOM	Self organizing map

SVD	Singular value decomposition
TF	Transcription factor
TFBS	Transcription factor binding site
QSAR	Quantitative structure-activity relationship

D	Data
D_{JS}	Jensen-Shannon divergence
D_{KL}	Kullback-Leibler divergence
H	Entropy
I_v	Number of classes of a multinomial variable
K	Number of parameters
k	Number of clusters or segments
m	Number of change-points
N	Number of data points
n	Cluster or segment size
P	Probability
Θ	Group of parameters
θ	Parameter
V	Number of dimensions
X	Data

List of original publications

This thesis is based on the following publications referred to by their Roman numerals (I-V):

I Koskinen, H., Pehkonen, P., Vehniainen, E., Krasnov, A., Rexroad, C., Afanasyev, S., Mölsä, H., and Oikari, A. Response of rainbow trout transcriptome to model chemical contaminants. *Biochemical and Biophysical Research Communications* 2004 320(3):745-53.

II Pehkonen, P., Wong, G., and Törönen, P. Theme discovery from gene lists for identification and viewing of multiple functional groups. *BMC Bioinformatics* 2005 6:162.

III Vartiainen, S., Pehkonen, P., Lakso, M., Nass, R., and Wong, G. Identification of gene expression changes in transgenic C. elegans overexpressing human α-synuclein. *Neurobiology of Disease* 2006 22(3):477-86.

IV *Kankainen, M., *Pehkonen, P., Rosenstöm, P., Törönen, P., Wong, G., and Holm, L. POXO: tool series to discover transcription factor binding sites from gene sets. *Nucleic Acids Research* 2006 34(Web Server issue):W534-40.

V Pehkonen, P., Wong, G., and Törönen, P. Heuristic Bayesian segmentation for locating co-expressed genome regions. Submitted manuscript.

* = first authorship

Table of contents

1 Introduction

Molecular biology is a study of living processes at the level of molecules. It is concerned with the reactions they undergo in the cells of living organisms, how these reactions are controlled, and how the molecules are assembled into larger structures such as membranes and fibres (Boudreault-Lapointe, 1988). Its main focus is in the macromolecules DNA, RNA, and proteins, and various smaller inter- and intracellular molecules. The major challenges in molecular biology are related to the complexity of the mechanisms comprised of the interactions and reactions of these molecules. Science must take enormous steps forward in developing technologies and methods that can address this problem.

After huge efforts, technologies have been established, which can obtain information about the structure and function of cellular molecules. The discovery of the molecular structure of DNA (Watson and Crick, 1953), by using data from *x-ray crystallography*, can be considered as a historical breakthrough. Soon afterwards, the structures of complete proteins were discovered using crystallography analysis (Kendrew, 1958). The next big milestone was the invention of the *DNA-sequencing* procedure (Maxam and Gilbert, 1977) which provided researchers with a huge amount of new information. These inventions were followed by the development of *high throughput laboratory technologies* in the 90's. These technologies utilize the common physical, chemical, and biological characteristics of molecules together with the very accurate measuring equipment in simultaneous measurement of structural and functional aspects of macromolecules. As an example, the DNA-microarray technology (Schena, 1995) utilizes simultaneous hybridization reactions for measuring gene expression levels of thousands of genes. The novelty of these technologies is that they provide an insight to the whole cell content at the given time point rather than only considering one or a few molecules at time.

The availability of high throughput technologies magnified rapidly the amount of information on the biomolecules. Subsequently, new challenges emerged related to the storing, analysis, interpretation, and integration of such data. Due to active development of methods for these purposes in mathematics and computer science, a new field of science referred to as bioinformatics came into existence during the last few decades. Due to the large amount of produced information, the focus of whole biosciences has moved from laboratory work to the informatics.

It is possible to distinguish multiple more or less overlapping subfields in bioinformatics. The earliest applications concern *sequence analysis*, such as alignment of two sequences for a pair wise comparison (Smith and Waterman, 1981), and more recently, determining the complete sequence from arbitrary sub-sequences obtained from genome shotgun sequencing (Myers et al., 2000; Celniker et al., 2002). Other sequence analysis approaches include prediction of protein structure from primary sequence (Bowie et al., 1991), and parallel alignment of multiple sequences (Lipman et al., 1989). A separate branch of bioinformatics involves quantitative techniques for modelling interactions between molecules, such as chemical compounds and proteins. This is called quantitative structure-activity relationship modelling (QSAR). Another major sub-field involves analysis of data obtained from high throughput screening of biopolymers, such as DNA-microarray array analysis. This thesis represents original contributions in this last field.

1.1 General background for the thesis

Work in this thesis involves the development and application of bioinformatic methods for analysis of data obtained from high throughput technologies, mainly the gene expression microarrays (Schena et al. 1995). The microarray technology utilizes binding of complementary nucleic acid molecules, a process that is called *hybridization*, to measure transcriptional activity of genes. Such activity is often referred to as the *gene expression level*. Microarrays facilitate performing such measurements simultaneously for thousands of genes. This enables genome wide (all genes) studies about the response of genes to different effects like chemical treatments, environmental conditions, or diseases. Several other high throughput technologies are also in use, such as protein arrays, chromatin immunoprecipitation (chIP) screening (Ren et al., 2000), and single nucleotide polymorphism (SNP) genotyping platforms (Lockhart et al., 1996), producing data on different aspects of macromolecules. The methods presented in this thesis are also compatible with some of these technologies (see chapter 2.2 for more detailed description).

DNA microarrays produce a large amount of quantitative data on gene expression levels under the tested conditions. The amount of measurements requires the use of computational analysis methods. These methods often summarize the data by grouping genes based on some similarity measure. The usual result of such analysis is a set of genes that show a similar expression level under each tested condition, co-expressed genes, or differential expression between two tested conditions. The principal step in the analysis is then taken by asking the question: what underlying biological mechanism explains the co-expression? This is also the principal research problem in this thesis.

In order to obtain a biologically meaningful interpretation for the expression change of a set of genes, the integration of various kinds of gene related information from the public biological databases is required. This thesis provides novel data mining methods and associated software for such interpretation. The presented methods identify expected relationships of genes and experimental conditions by literature mining, discover biological processes which can explain the co- or differential expression by applying a clustering method into functional information on genes, discover putative regulatory sequences which can explain co-expression by using an analysis pipeline, and find the chromosomal locations with over-representation of genes within a particular co-expression group by using a segmentation procedure.

Data analyzed in this work is comprised of associations between genes and biological attributes, such as biological conditions, chromosomal locations, textual expressions from literature, database records etc. Such associations are represented as discrete categorical data. There are some points that are emphasized in the analysis of such data. First, common discrete data sets are often of very high dimensionality (Patrikainen and Mannila, 2004). The traditional data analysis procedures are often insufficient to handle such data because of the problems related to algorithmics, such as increased time complexity (Bellman, 1961a), or statistics, such as inability to recognize patterns in sub-spaces of all data dimensions. Regarding the latter point, the performance of standard similarity measures has been found poor with high dimensional (Aggarwal, 2001) and discrete data (Kontkanen et al., 2001). Third, as noted before in Kontkanen et al. (2003), the common demand for such analysis methods is increasing rapidly. This is due to increased quantity of such data like market basket, questionnaire, and web-log data sets, as well as different biology related data sets.

Work reported in this thesis includes development and application of unsupervised machine learning methods, such as data clustering, segmentation, and related statistical evaluation. For clustering high dimensional binary data sets, we apply *Non-negative Matrix Factorization* (NMF) (Paatero and Tapper, 1994; Lee and Seung, 1999). NMF is a novel matrix factorization approach which has shown good performance with analysis of high dimensional binary data (Lee and Seung, 1999; Seppänen et al., 2003). For segmentation, we apply a heuristic approach in order to obtain results in a reasonable time.

A common challenge with clustering and segmentation is the selection of an appropriate number of clusters or segments. We have ended up with two different solutions. With the clustering procedure, we observe the stability of clusters with visualization in order to discover coherent clusters. We also perform evaluation by analyzing the over-representation of data attributes in clusters. With the segmentation, we developed a model selection method for evaluating different segmentation solutions obtained from a heuristic algorithm. The model selection can be considered as a problem of finding a good trade-off between the complexity and the goodness of fit. Such a problem is also known as *Occam's Razor* in the literature (Myung and Pitt, 1996). In our model selection, we use a Bayesian approach which addresses this problem directly by facilitating analysis of uncertainties related to the model parameters.

1.2 Original publications and personal contribution

The publications I-V present original research related to development and use of DNA-microarrays and analysis of generated quantitative data. As a main contribution for this thesis, the publications present novel data mining methods which aid in biological interpretation of such data. The data mining is based on the annotations (or descriptions) of genes obtained from biological databases and literature collections.

Publication I introduces a cDNA-microarray study of rainbow trout gene expression under treatments of environmental toxins. First, it presents a novel cDNA-microarray for salmon fish and its application to rainbow trout samples. Secondly, it presents the analysis of produced quantitative data and data mining for biological interpretation of results. A simple approach for mining MEDLINE literature database is introduced, which indicates the expected and novel findings in obtained microarray data. This publication is included as an introduction to high throughput technologies and general analysis of such data. It also serves as a starting point for the author's later developments in data mining. The author of this thesis developed and used bioinformatic tools to support different stages of the study: design of microarray, annotation of fish genes, microarray data analysis and data mining. The author also conceived and implemented the method for literature mining. The author also contributed to writing of the manuscript.

Publication II presents a novel approach for discovering functional themes from a set of genes, such as co-expressed genes. The introduced method creates a non-nested clustering scheme which facilitates discovery of coherent clusters and hierarchical relationships of data using visualization. As a clustering method, we use the Non-negative Matrix Factorization (NMF) based approach and study its performance in finding biological topics. We also show comparison of our method against the other existing tools in the same application. The original idea of clustering by using biological topics was given by the last author of the publication. The author of this thesis designed the method together with the last author, implemented the software for using method,

tested method with various datasets including artificial data and biological datasets, compared method against other methods, and drafted the manuscript.

Publication III presents a study concerning the *C. Elegans* nematodes, which were genetically modified by transferring human Parkinson's Disease related α-synuclein gene into their genome. DNA-microarrays were used to study the genome wide gene expression differences between transgenic and wildtype nematode strains. The method for clustering presented in publication II was applied in analysis of differentially regulated genes, in order to reveal separate biological functions and processes related with neurodegeneration. The author of the thesis participated in gene expression data analysis, mostly by performing analysis using the method presented originally in II. The author also made a biological interpretation for the outcome of the method and had a significant contribution in writing the manuscript.

Publication IV presents a large scale software pipeline for the discovery of transcriptional factor binding sites from promoter sequences of co-expressed gene sets. The method presented in II was further refined and integrated as the first part of the pipeline in order to find functionally similar gene groups, and analyze their sequences separately. Most of the other parts of the pipeline, such as pattern discovery and clustering methods, were developed and published separately before elsewhere. The author of this thesis contributed by developing the analysis series with the other first author (shared first authorship). This included mainly the refinement, implementation, and integration of the clustering method to the previously created tools. The author had also a contribution in writing the manuscript.

Publication V presents a segmentation method, which is based on the Bayesian modelling of segmentation solutions obtained from heuristic heuristic algorithm. The method was applied to study of chromatin remodelling in baker's yeast by using genes grouped into co-expression clusters according to their gene expression profiles during the cell cycle. The paper introduces a Bayesian model for evaluating heuristic segmentation solutions, including a simple prior for segmentation model, and a new empirical prior for multinomial or binomial data. Also, a benchmark for comparison of different segmentation methods is presented. The author of this thesis gave the original idea of segmentation and its application in the analysis of location specific gene expression. The development of methodology was mostly a joint work of the first and the last author. The idea for the empirical Bayes prior of segmentation model was originally given by the last author. Application of method and biological interpretation was done by the author of this thesis and the second author. The author of this thesis made the software implementation of the method, the application, and the method comparisons, and drafted the manuscript.

1.3 Structure of the thesis

In chapter 2, the background information about the roles and mechanisms of different macromolecules in cell are discussed. The biological problems of the thesis are formulated at the end of the chapter. In chapter 3, the methodological background for the thesis is presented. This includes the review of methods for clustering and segmentation, and the overview of related concepts in statistical modelling. In chapter 4, the methods developed in publications I-V of this thesis are summarized and discussed. Chapter 5 contains the conclusions and discussion concerning the future work.

Reading the biological background is recommended for understanding the basic biological entities and the motivation behind the methods. However, chapter 4 describes briefly the biological problem addressed by each presented method.

2 Biological background

2.1 Information about life

The main principle of molecular biology is called the Central Dogma. It was originally established by Francis H. Crick, based on his research on the structure and function of macromolecules of the cell (Crick, 1958; Crick, 1970). The Central Dogma defines a deterministic flow of "information about life" starting from a gene (figure 1). According to Central Dogma, gene is a segment of nucleic acid located in the organism's DNA. It stores heritable information needed for producing a functional product, usually a protein. Each protein (and any gene product) addresses some functions (e.g. maintaining and building membranes and fibres or response to environmental factors). In the Dogma, the information stored in a gene is copied into RNA in a process called transcription. This information is further copied (during translation) into an amino acid sequence, which is an early form of protein. The next step includes the protein folding into a multiform structure. Also, DNA mediates information to another DNA in the duplication during cell division.

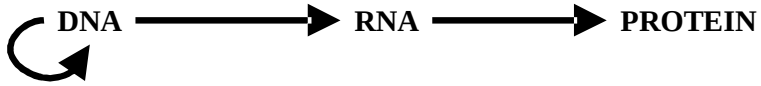

Figure 1. Central Dogma. Information is copied from DNA to RNA in transcription, and from RNA to protein in translation. DNA replication copies information from one DNA to another.

The Central Dogma presents the role of DNA as a passive store of heritable information. RNA and early forms of proteins are mediation stages distributing that information further. The finished protein is presented as an only element which has an active function in the cell. Later studies found points that extend, or are partially in conflict with, the Dogma (see for example Fire et al. 1998, Gerstein et al., 2007 and ENCODE Project Consortium, 2007). For example, an abundance of short genomic regions have been found, that are transcribed but not translated into protein. These micro RNA molecules could silence transcription by destroying the mRNA of target genes, in a process called the RNA interference (Fire et al. 1998). Also, a recent study on genome DNA elements reveals that there is much more functionality related with the genomic DNA that was thought before (ENCODE Project Consortium, 2007). This has also required renewing the definition of a gene (Gerstein et al., 2007).

Pathways

Further studies based on Central Dogma discovered more complex mechanisms in cells called pathways. A pathway can be considered as a molecular program comprised of consecutive steps of molecular interactions. Each pathway executes a particular biological function. The interacting molecules can be nucleic acids, proteins, or metabolites (Nelson and Cox, 2004). Each interaction implements some minor objective in a pathway such as receptor molecule activation through the impact of a signalling metabolite, binding of several proteins into a protein complex for executing some further task, or binding of protein to gene promoter DNA. Multiple chains of steps can take place in parallel and the pathway can overlap with, or include, another pathway (see for example Dohrmann et al., 1992).

According to the type of interactions, the pathways are often classified into signalling, metabolic, and gene regulatory pathways (Nelson and Cox, 2004). In signalling pathways, the intra- or extracellular molecules mediate signals that trigger particular events inside the cells, often through receptor proteins attached in the cell membrane. Metabolic pathways contain chemical modification of target molecules, such as proteins, with the help of nutrients and enzymes, in order to maintain vital functions of a cell, such as energy intake. Regulatory pathways address regulation of a gene, or often a set of genes that are regulated in concert. Larger mechanisms of a cell are often comprised of a combination of different types of pathways. For example, the regulation of genes via CREB transcription factor includes both signalling and gene regulation elements (see for example Zhang et al, 2005).

In publication II, III and IV, we present and apply a method that clusters genes according to gene functions. This method can reveal gene groups that associate to separate pathways or large parts of a single pathway.

Gene transcription and its regulation

As this thesis concentrates mainly on the analysis of gene expression data, the transcription and its regulation are the main focus. It is known, that gene regulation is often mediated by special proteins called transcription factors (TFs; Wray et al., 2003; Nelson and Cox, 2004). In such a process, TFs bind to TF binding sites (TFBSs) in the promoter region of a target gene. The promoter is the region locating upstream from the transcription starting site of a gene, which includes TFBSs. TF can start, enhance, repress or silence the transcription of the gene into mRNA (Nelson and Cox, 2004). TF proteins are also products of genes, and therefore regulation chains of multiple genes exist. As each such a protein can have several optional functions in a cell autonomously, or in complexes with other proteins, these chains often constitute complex networks. Figure 2 shows a simple hypothetical gene regulatory mechanism.

Recently, several new aspects on transcription regulation have been discovered. Some of these aspects are important for this thesis, and thus discussed here. The first point concerns the diversity of transcription. That is, the transcription is tissue, individual, population, and species specific (Hsieh et al., 2003; Whitehead and Crawford, 2005). This explains the large polymorphism between different organs such as brain and liver, as well as partially the differences between populations. We have detected such differences between the fish kidney and brain before, in Krasnov et al. (2005).

The second transcription regulation related aspect concerns the impact of chromosome structure. The chromosome is built from the complexes of DNA and protein molecules such as histones. This structure includes loose parts, called euchromatin, and parts called hetero-chromatin, that are tightly packed and bound over the histone proteins. It is well known, that the structure is mostly loose during the cell cycle interphase to allow DNA duplication, and packed prior to mitosis to allow cell division. However, recent studies show that the chromatin structure is also dynamically modified during the interphase to allow and block transcription of several genes in particular chromosome regions (see for example Gilbert and Ramsahoye, 2005). In publication V (see also chapters 2.3 and 4.4), we particularly search such active and passive regions of genome from yeast.

In addition there are several other factors that may affect transcription regulation. These include DNA-methylation, positions and moving of histones, type of histone molecules, operon like

structures, and duplicated promoter regions. These are important aspects, but not in the scope of this thesis. For further information on this topic, see for example Gilbert and Ramsahoye (2005).

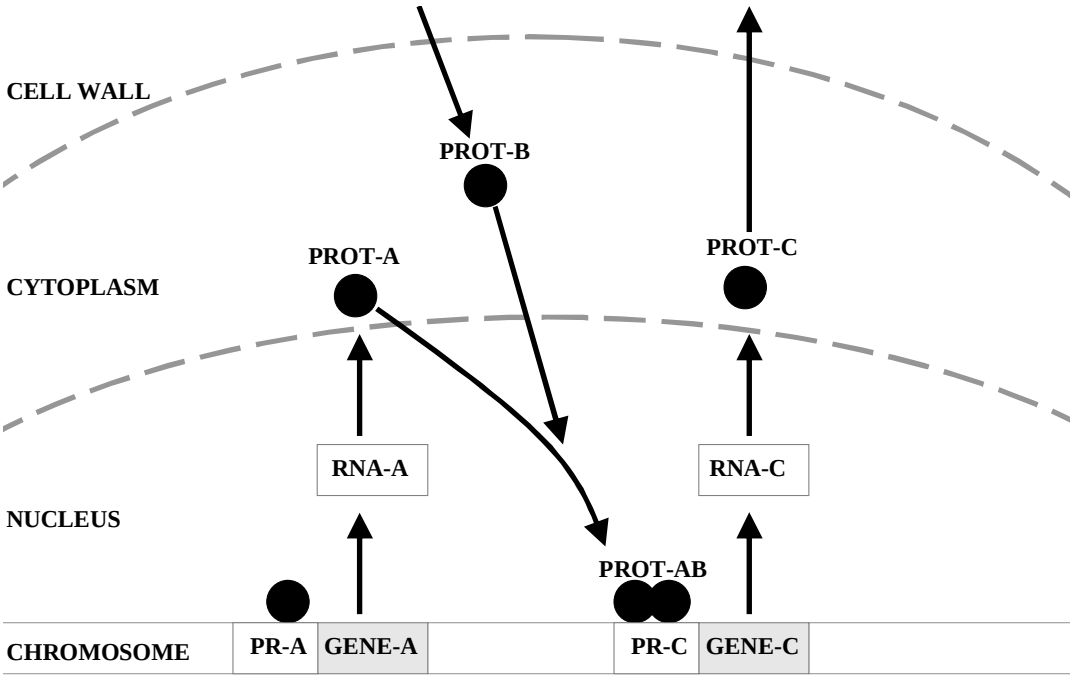

Figure 2. Hypothetical example of a standard macromolecule mechanism within cell. The regulatory proteins have opened the chromatin structure. Transcription factor has bound in the promoter area PR-A of gene GENE-A and started its transcription into mRNA. Gene GENE-A is also another transcription related gene which codes transcription factor protein PROT-A. From elsewhere, possibly from outside of the cell, protein PROT-B comes and creates a protein complex PROT-AB with PROT-A. The complex regulates the transcription of gene GENE-C. This gene finally codes RNA-C and PROT-C protein which has some further function.

2.2 Genome wide high-throughput technologies

Due to findings on general structure and function of macromolecules of the cell, a large effort was put in obtaining information on their specific mechanisms. The rapid genome wide DNA-sequencing technique developed during the last decade (Roach et al., 1995) contributed by providing researchers with the abundance of information on full genomic sequences of many organisms. This further facilitated development of various technologies that utilize the availability of genomic sequence in studies of biopolymers. A common principle with these technologies is that they all perform simultaneous screening of a very large gene set, possibly an entire genome. Such screening produces qualitative or quantitative information related to macromolecules in cell. Thus these methods are often referred to as *high throughput technologies*. They are currently in standard use in biology and medical research.

Several high throughput methods exist that utilize *a hybridization reaction*, the pairing of two complementary strands of nucleic acids. Such technologies for measuring gene expression include mainly two types of platforms: cDNA-microarrays (Schena, 1995) and oligo nucleotide arrays (Lockhart et al., 1996). Another application of the oligo nucleotide arrays is detection of

single nucleotide polymorphisms (SNP) from sample DNA. Also the high throughput comparative genome hybridization (CGH) is another application of microarray technology (Snijders et al., 2001). CGH measures gene copy numbers from the sample genome DNA. This is often used to detect abnormal duplication of genes in cancer studies. Chromatin immunoprepicitation (ChIP) method (Ren et al., 2000) measures the binding of regulatory proteins into promoters in a genome wide manner.

In addition to hybridization based methods, there are other technologies that obtain the data from cellular molecules in a high throughput manner. *Mass spectrometry* (MS) produces molecular level information about the sample contents, such as type of metabolic compounds or so-called fingerprints of proteins. Other applications of MS are SNP screening (Little et al., 1997) and high-throughput gene expression profiling (Ding and Cantor, 2003). Another kind of array based technology, protein arrays, measures the protein contents of cell. This technology is promising as it can address directly the problem about the existence of particular proteins, which is often addressed indirectly with gene expression arrays. Still, it is largely at the stage of development (Lee and Mrksich, 2002). In addition, there are tissue arrays which facilitate the screening of hundreds of tissues simultaneously for example to detect the existence of a specific protein.

Data obtained from different high throughput technologies provide information on different aspects of cellular molecules. Such aspects are integrated in the approach called *systems biology* (Ideker et al., 2001). For example Berger et al. (2004) combine gene copy number and expression data for cancer studies.

This thesis concentrates mainly in the gene expression microarray technologies and the produced data, while auxiliary data is integrated from public databases to interpret the results (see chapter 2.2.3). It should be noted, that despite this focus, the developed methods are applicable with other types of technologies mentioned in this chapter that involve similar king of profiling, such as protein arrays, or produce gene or protein lists as a result, such as protein arrays or MS gene expression profiling.

2.2.1 Gene expression microarrays

Gene expression microarrays (Schena et al. 1995) measure simultaneously the expression levels of a large amount of genes, such as an entire genome. This can be used directly to study transcription, or indirectly to predict the presence of proteins. Microarray technology is based on biopolymer sequences, referred to as *probes*, which are attached in the platform. The probes are organized in distinct spots or cells in the platform, each of which contains a large amount of probe material, representative for a specific gene of an organism. The probes are complementary to the expressed mRNA of cell, contained in the sample material. Thus, they pair with the mRNA when coming into contact in a reaction called *hybridization*. Before the hybridization, the sample mRNA is labelled with fluorescent dye, such as Cy-3 or Cy-5. Simultaneous hybridizations take place when the microarray platform and sample material come into contact. From the intensities of dyes, it is possible to infer the relative amounts of hybridizations. This is possible when the fluorescent light intensity emitted from the dye is digitized with scanner, when excited with a laser at a particular dye specific wavelength. The obtained values correspond to the relative amounts of mRNA molecules in the sample material, i.e. the *gene expression levels*.

Mainly two types of DNA-microarray platforms are currently in use: spotted complementary DNA (cDNA; Schena et al. 1995) and *in situ synthesized* oligo nucleotide arrays (Lockhart et al., 1996). For comparison, figure 3 shows parts from the scanned pictures of cDNA and Affymetrix chips. The cDNA arrays contain probes comprised of 300-500 bp long single strand cDNA sequences that have been spotted or printed on the platform. The cDNA material is obtained by synthesizing from expressed sequence tags (EST) obtained from the expressed genes' mRNA sequences. Often two samples are screened with each array. One of the samples is obtained from the tested condition or time point, whereas the other is often a control sample representing the cell in a normal or untreated state. As a result, the ratios of gene expression levels between the tested and control sample are often reported for each gene. This works also as a normalization for spot-to-spot variation (see the *Normalization* section in chapter 2.2.2).

In situ synthesized oligo arrays use probes that are short nucleotide sequences called oligo nucleotides. These are complementary with the sub-sequences of expressed mRNA of genes. *In situ synthesized* platforms contain often cells that include abundance of copies of oligos representing the tested genes. These cells are built in the array rather than printed as with cDNA arrays. They are often used to perform screening with one sample per array rather than using the control sample.

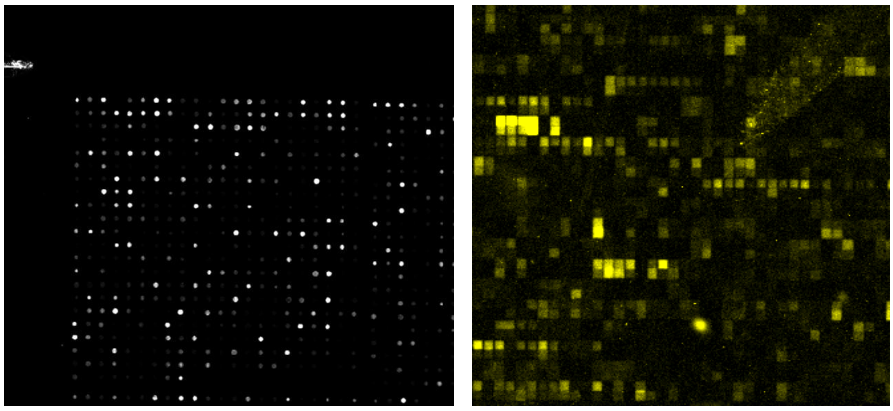

Figure 3. Parts of scanned images of cDNA (on left) and Affymetrix oligo chips. In cDNA chip the spots are often separated with larger empty spaces whereas Affymetrix probe wells are deposited firmly next to each other. Each spots in the cDNA chip is often a representative of single gene in the whole chip whereas in Affymetrix platform a single gene is represented by several neighbouring or distinctly located probe cells.

An example of *in situ synthesized* oligo method is the Affymetrix oligo chip technology (Lockhart et al., 1996). Each gene (or probe set) in this platform is represented by several probe cells in order to get more reliable measures. Each cell contains a large amount of copies of an oligo sequence representing perfect match (PM) for a particular gene transcript (mRNA). In addition there is a mismatch (MM) probe for each PM probe for detecting the amount of cross-hybridization to PM probes. MM probes contain a single nucleotide difference in the middle (Tuimala, 2003).

It has been shown that the Affymetrix chips are more reproducible and specific in detection of gene expression levels than cDNA chips (Woo et al., 2004; Bammler et al., 2005). The disadvantage is that a limited repertoire of these arrays exists and they are relatively expensive.

Therefore, the cDNA arrays are still very popular in studies of unfamiliar organisms or when aiming at low expenses. Another shortcoming of Affymetrix chips was reported by Dai et al. (2005). They showed that the definition of Affymetrix probe sets in the probe level was flawed due to use of an earlier version of genome and transcriptome annotation. By re-mapping the probes according to an up-to-date sequence, they showed that there was approximately 30%-50% difference in the sets of differentially expressed genes to the results with original probe sets. They propose re-analysis of earlier datasets by using the new mapping of probe sets they published (see Dai et al. 2005 for further information).

In this thesis we have used our own custom designed cDNA-microarrays for salmon fish in publication I. Creating custom made cDNA chips for salmon fish was necessary as a product did not exist in the market at that time. In publication III we used Affymetrix oligo chips for *C. elegans* nematode (Lockhart et al., 1996).

2.2.2 Analysis of microarray data

The principal steps of a regular microarray study are shown in figure 4. See the figure legend for brief description of different steps. This thesis focuses on analysis and biological interpretation of produced data. The analysis contains pre-processing, normalization, and grouping steps. The usual aim of analysis is to find the genes that behave similarly under treatments, co-expressed genes, or genes that show different expression level between the studied treatment and control condition. The focus is on analysis of data obtained from cDNA and Affymterix chips as they are used in this thesis.

Preprocessing

Hybridized and washed microarrays are first scanned in a microarray scanner which digitizes the array image under the excitation of a laser at a particular dye specific wavelength (Tuimala, 2003). As a result, an image is produced for each fluorescent dye. It shows the quantity of hybridizations of target mRNA molecules within each probe spot or cell as a colour intensity.

The aim of further preprocessing is to obtain intensity of each probe cell or spot from the scanned image. Therefore, the first step is to determine the location of each spot or cell in the array. The usual procedure to perform this with both cDNA and oligo chips is called *gridding*. It involves overlaying a grid over the array image so that each grid square delineates a spot or probe cell. Several gridding methods have been presented including manual, semi-automatic, or completely automatic methods. Manual or semi-automatic methods are very time consuming and non-reproducible (Tuimala, 2003). Thus mainly automatic methods are currently used such as Bowman et al. (2002) and Deng and Duan (2004) with cDNA microarrays and Global Gridding algorithm (Affymetrix Inc., 2005) with Affymetrix oligo chips.

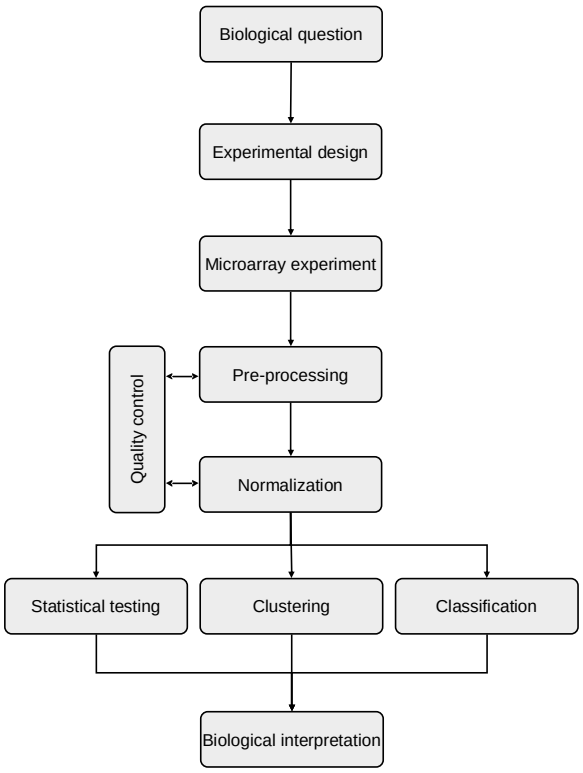

Figure 4. Simplified flow diagram of a microarray experiment and data analysis lifecycle. The experiment is motivated by a biological question. This can be often created by a new hypothesis of some phenomenon such as response of gene expression during a particular treatment. This is followed by the experimental design which has a large impact on how reliable results can be obtained in the further steps. An example of this is the number of chip replicates per condition, as the low number of replicates reduces the statistical power in selection of differentially expressed genes. The raw data produced by experiment is pre-processed and normalized using procedures documented in chapter 2.2.2. Quality control should precede and follow every procedure performed at this stage. The next stages depend largely on the type of study. If the aim is to search genes that are differentially expressed between two conditions or in one among several conditions, statistical testing is performed. If the aim is to discover co-expressed genes in an unsupervised way, the clustering is usually performed for genes. Studies that aim at grouping samples can use clustering or classification. The final stage is biological interpretation which also contains the data mining methods developed in this thesis, documented in chapter 4.

Often the scanning light intensities between different array platforms or between different parts of a single array vary, adding bias to the spot or cell intensities. Thus, the next step is to determine the degree of intensity that is obtained due to this effect. This is referred to as background intensity. With cDNA chips the background intensity is measured from the background of each spot within the aligned grid square (see figure 3). Separation of background from the spot area is usually performed using a technique called segmentation. Abundance of different segmentation methods exist for different purposes, including those described in chapter 3.2.2. With cDNA chips, very simple methods such as simple histogram segmentation in TIGR SpotFinder software (Saeed et al., 2003) are used. After detecting the spot boundaries, it is possible to read the background intensity from the pixels outside the boundaries.

Due to different architecture, the background intensity is measured differently with Affymetrix oligo chips than with cDNA-microarrays. Each Affymetrix probe cell, representing either PM or MM probe of some probe set, is firmly located next to the neighbouring cells (see figure 3). For background detection, an algorithm called Zoning algorithm (Affymetrix Inc., 2002) is used. It divides an array into zones (16 by default), within each of which the cells are ranked according to the average cell intensities. The most dim (2%) of cells in the each zone are chosen to represent the background of that zone. In the subsequent steps, each array cell is assigned with the estimate of background by summing the weighted background values of each array zone. Here, the weight is the distance between the cell and a particular array zone centre (Affymetrix Inc., 2005).

Intensity value for each spot or cell in the array is often calculated by subtracting the estimate for the background intensity in the grid square from the complete intensity. For example with cDNA chips the background adjusted intensity of each spot is calculated in TIGR SpotFinder (Saeed et al., 2003) with formula $I - BKG * A$, where BKG is the median average intensity, I is the sum of all pixel intensities in the spot area, and A is the number of spot area pixels. Calculation of PM and MM probe intensities in Affymetrix oligo chips is conducted simply by subtracting the background estimate for each cell from the detected complete intensity (Affymetrix Inc., 2005). The calculation of cell intensity from different pixel intensities is not reported comprehensively by Affymetrix.

All stages of microarray data preprocessing, also normalization, should include quality control. This is often performed by comparing intensities between different replicate chips and also between different conditions. High linear correlation is expected between the replicate chips whereas, due to assumption on distribution of genome expression levels mentioned above, linearity is expected also between any two biological conditions (Tuimala, 2003). One often performed step is hierarchical clustering (agglomerative hierarchical clustering: Johnson, 1967; see also chapter 3.2.1 for more detailed description) of all chips using correlation as a similarity measure. This is a handy approach to reveal correlations between different chips.

Removing spots that are artefacts is also a general procedure at the preprocessing stage. Two measures are commonly used to detect the validity of a spot: signal-to-background ratio and signal-to-noise ratio (Tuimala, 2003). The first is often simply the ratio of average spot pixel intensity and average background pixel intensity. This is used as a basis of filtering in publication I. Signal-to-noise ratio is often interpreted as a ratio of mean pixel intensity and standard deviation of pixel intensities in the spot and background areas. There are also other approaches such as discarding spots based on their shape or detecting spot saturation, both available in TIGR SpotFinder (Saeed et al., 2003).

Normalization

Normalization is a procedure to remove non-biological variation from data, and can be also considered as a part of pre-processing pipeline (Tuimala, 2003). Several sources of such variation exist, including random effects and systematic bias. Normalization aims rather at removing the systematic effects, such as dissimilar dye effect, scanner malfunction, uneven hybridization, variation in chip printing, variation in plates etc. Specific description of common sources of bias is given in (Tuimala, 2003).

A general assumption behind the normalization is often that the logarithmic gene expression levels in the whole genome have normal distribution with nearly the same mean and standard deviation despite the sample treatment or condition (Tuimala, 2003). Therefore in chips with a large number of genes the normalization is often based on transforming the chip log-intensity distributions so that they become similar. With low number of genes, control genes which are assumed to be expressed with similar levels under any condition, are commonly used instead of this assumption (Tuimala, 2003).

Commonly, the normalization means transforming the data to more normal like. With microarray data, this is usually performed by logarithmic transform of intensity values or intensity ratios. In the context of microarrays, the normalization also means centralization and

standardization of chip intensity distributions (Tuimala, 2003). Centralization contains techniques for scaling the distributions to the same mean or median. This is simply obtained by subtracting logarithm of mean or median from the log-intensity values. Local methods like Lowess can make centralization based on the local intensities on chip. In turn the standardization, in common terms, means standardizing the shape of the distributions. The statistical technique called standardization divides the difference of a value and the mean of all values by their standard deviation (Tuimala, 2003). The mean of the resulting distribution is zero and the standard deviation is 1. With normal distribution, this means transformation to the standard normal distribution. Z-score standardization of microarray data is discussed more in (Cheadle et al., 2003).

A gene expression matrix, comprised of genes on rows and samples as columns, is often normalized in two directions: per-chip and per-gene. Per-chip normalization normalizes log-intensity distributions for each array whereas per-gene normalization normalizes values of each gene across all chips. The aim of per-chip normalization is to remove array specific bias whereas per-gene procedures remove spot-to-spot variation (Tuimala, 2003). Mean or median centering are common approaches in both cases. Median centering is rather performed with per-gene normalization as there are often rather small amount of observations (Tuimala, 2003).

Local normalization addresses the systematic bias locally, unlike the global methods discussed above. Lowess normalization (Cleveland, 1979) is a popular local method. It uses a sliding window algorithm to estimate the regression curve through the (hypothetical) scatter plot of test and control chip. Each intensity value is then corrected by subtracting the curve from the original values. Lowess is used in publication I for analysis of salmon fish microarray data.

One approach to remove the array specific bias in two colour cDNA chip experiments is the dye-swap experimental design and normalization (Kerr et al., 2001; Yang et al., 2001). In addition to hybridization with test and control samples dyed with Cy5 and Cy3, dye-swap involves replicating the microarray hybridization where the dye assignment is reversed between the two samples. This can effectively remove the bias that is affected by different dye or mRNA amounts in the different sample materials (Dabney and Storey, 2005). We use dye-swap in publication I to get more trustworthy estimates of differential expression.

Affymetrix normalization

With cDNA chips, each array contains often a single spot or a few replicated spots in the subarrays (see for example publication I) for a single gene. Such measures are often normalized using some approach described above, and used as estimates of gene expression levels in further analysis such as statistical testing or clustering (see the two sections below). The Affymetrix chips, in turn, use several PM oligo probe cells to represent a single gene in a single array. In addition, the MM probes (see explanation in chapter 2.2.1) are used to estimate the level of cross-hybridization with the PM probes. Therefore for each gene, an estimate of gene expression level must be obtained from the intensities of these probe cells.

There are currently three main procedures to obtain an expression level for Affymetrix probe set: Affymetrix MicroArray Suite (MAS) 5.0 algorithm (Affymetrix Inc., 2005), dChip method (Li and Wong, 2001a; Li and Wong, 2001b) and Robust Multi-array Average (RMA) method (Irizarry et al., 2003a). The more advanced methods: dChip and RMA perform three pre-processing steps: background correction, normalization and summarization. MAS 5.0 does not

involve the normalization step. It should be noted, that the correction of background has been already made once by Affymetrix's own procedure described in the *Preprocessing* section above. However in this context, in addition to light intensity correction, it also implicates the correction of PM cross-hybridization using the intensity of MM cell.

Affymetrix MAS 5.0 method corrects the cross-hybridization on PM cells using the *ideal mismatch* procedure (Affymetrix Inc., 2005). This approach subtracts the MM cell intensity from PM cell intensity when PM>MM. Otherwise the PM intensity is set equal to MM intensity (plus a small constant). In MAS 5.0, the probe intensities are not normalized but rather the user is expected to normalize summarized probe set intensities afterwards using some methods documented in the paragraphs above. The summarization is made using Tukey's biweight method (Saviozzi and Calogero, 2003).

In dChip, the correction for cross-hybridization is made simply by subtracting MM from PM intensities (Li and Wong, 2001a). It is also possible to discard MM information completely and use only the PM intensities in further procedures. The normalization is conducted by a method that uses an invariant set of probes in different chips to scale chips to equal level. This is discussed more in (Li and Wong, 2001b). Summarization in dChip is based on the idea that the variation of individual probes across different arrays can be considerably higher than the variation between probes of the corresponding probe set (Saviozzi and Calogero, 2003). Thus, in addition to information obtained from a single array, it uses probe intensities from all chips. The dChip model for summarization assumes that the observed probe intensity is the multiplication of the true gene expression level and probe effect, plus an error term (Li and Wong, 2001a; Li and Wong, 2001b). An iterative procedure is used to fit this model into data, and to estimate the true expression level referred to as Model Based Expression Index (MBEI).

In RMA, the background is corrected using a model which assumes that the observed intensity is the sum of true intensity and background noise (Irizarry et al., 2003a). This is aimed to remove both the background intensity variation and bias from cross-hybridization. The computation of this model is difficult (McGee and Chen, 2006; Bolstad, 2004), and thus the current implementations, such as Bioconductor (Gentleman, 2004), use an *ad hoc* solutions instead of standard approaches like the EM algorithm. RMA conducts normalization for the background corrected probe intensities by using quantile normalization (Irizarry et al., 2003a; Saviozzi and Calogero, 2003). It equalizes the distributions between the normalized chips. A summarization of probe intensities into a single gene expression level is performed using median polishing method (Irizarry et al., 2003a). The model used by RMA assumes that the observed probe intensity is the sum of the true gene expression level, probe affinity across all arrays, and an error term.

dChip was the first method that takes into account the probe variation across the different arrays of an experiment, and thus it served as a significant refinement from the MAS 5.0 like methods. Many comparisons like (Irizarry et al., 2003b), show that dChip clearly over-performs MAS 5.0. We use dChip for pre-processing in publication III. In comparison, RMA has been found to outperform MAS 5.0 and also slightly the dChip method (Irizarry et al., 2003b).

Finding differentially expressed genes

After pre-processing and normalization, the aim is often to find the genes with differing expression between two tested conditions, or genes that differ in expression in some proportion of conditions.

With two colour chips, the logarithms of ratios between test and control intensities are often calculated, and thus one sample Student's t-test is often used to test the difference of measurements from zero (Tuimala, 2003). When samples in the test groups have dependent pairs in the control group, the paired Student's t-test can be used. With independent groups the two sample Student t-test should be used. In publication I, we use one sample Student's t-test with two colour cDNA micorarrays. In publication IV, two sample Student's t-test with one colour Affymetrix array is used.

Student's tests assume normally distributed population, and with two measurement groups equal variances for both distributions. When expression levels are assumed to arise from the normal distributions with unequal variances, Welch t-test is suitable alternative for Student's t-test for two independent groups (Tuimala, 2003). There also exist several non parametric tests used when the assumption of normality cannot be made. However, no clear consensus exists when to assume variances equal. Some sources recommend assuming variances always unequal between test and control conditions (Knudsen, 2002).

Detecting genes that are differentially expressed in some of the several conditions is usually performed using analysis of variance (ANOVA). This is a generalization of t-test for multiple tested groups. Several post-hoc tests exist that can be used to detect in which condition pairs the differences exist (Tuimala, 2003).

Due to multiple testing, the probability for obtaining smaller p-values by chance is increased from the originally chosen *α-level*, such as 0.05. Often, this level is still used for filtering with the ratio or fold change (see Tuimala, 2003) filtering in parallel. This removes also the genes that show low p-value but do not have a large change of expression. This approach is also used in publications I and IV. The other approach is to apply p-value correction. Due to a high number of tests, the traditional Bonferroni correction is not applicable (Tuimala, 2003). Currently, the Benjamini-Hochberg False Discovery Rate (FDR; Benjamini and Hochberg, 1995) has become a standard in gene expression data analysis. Statistical tests are further discussed in chapter 3.3.

Finding co-expressed genes

For obtaining co-expressed genes from data containing multiple time points or conditions, the classification and clustering approaches are often used. The most popular methods are k-means (MacQueen, 1967), hierarchical clustering, or self organizing map (SOM) (Kohonen, 1988, 1995). Clustering is used for finding gene groups where genes have similar expression levels within a set, and different sets have divergent expression levels. Such groups are often referred to as co-expressed gene groups. A more complete description of clustering procedures is given in the chapter 3.2.1. Classification contains supervised techniques that are not in the scope of this thesis.

The next step is to find explanations for co-expression or differential expression. This can either confirm or create new hypotheses, such as the relationship of a particular pathway. This step is often referred to as biological interpretation.

2.2.3 Biological interpretation

Biological interpretation of a gene set usually involves browsing gene annotations such as functional descriptions of genes. Often the aim is to find out if there exists some specific annotation that correlates with the co- or differential expression. The primary source of such information is

scientific biomedical literature. Therefore the manual interpretation of gene sets involves mainly browsing and reading the literature for descriptions of genes.

As the genome wide studies often produce gene sets containing even hundreds of genes, which can be altogether associated with hundreds or even thousands of different biological themes, the manual interpretation is not very convenient. Thus, lately such information has been collected and structured into the databases which associate genes with different types of annotations covering often whole genomes of several different organisms. Such annotations include a description of gene functions, related metabolic, signalling and transcriptional pathways, interactions with other genes, cellular locations of produced proteins, chromosomal locations etc. These databases do not only enable easy manual browsing of annotations but also computer aided statistical inference and data mining approaches based on the annotations.

Publicly available information about genes

Currently there exists an abundance of annotation databases. In addition there are public high throughput measurement data sets available, such as gene expression data obtained from microarray screening. Information from these databases is often used with the in-house experimental data in analysis and biological interpretation of data. Databases also facilitate the development of tools that utilize such information. The public databases can be classified as follows:

- **Microarray databases.** Several databases exist for storing and retrieving public microarray data. The major microarray databases are Gene Expression Omnibus (GEO) (Barrett et al. 2005), ArrayExpress (Brazma and Parkinson, 2006) and Stanford Microarray Database (SMD) (Sherlock et al., 2001). They store the information with required metadata defined in the Minimum Information About Microarray Experiment (MIAME) standard (Brazma et al., 2001). MIAME describes the required attributes for describing protocols, instruments and data-analysis methods used in a microarray project. It is one part of the standardization work of Microarray Gene Expression Group (MGED) to make microarray experiments reusable. Many scientific journals have started to require submission of microarray data into MIAME compliant public databanks before publishing a new manuscript.

- **Functional descriptions of genes.** Functional annotations are an example of attributes which are currently in frequent use in microarray (and other high throughput method) data analysis and interpretation. Assignment of such annotations has often been performed manually or semi-automatically from published literature by curators of the databases. Gene Ontology (GO) database (Ashburner et al., 2000) is a formal controlled vocabulary of functional description of genes. In addition it is a semantic definition of relationships between different biological processes, molecular functions, and cellular locations. Gene products in several different model organisms are associated in the database with these concepts. All studies presented in this thesis involve using GO database. Another similar database for functional information is the much simpler Münich Information centre for Protein Sequences (MIPS) database (Mewes et al., 2000).

- **Genome sequence information.** The information on genome sequences of model organisms is publicly available in databases that are specialized in a particular organism, such as WormBase (Stein et al., 2001) or FlyBase (Ashburner and Drysdale, 1994), or in the databases that cover multiple organisms and sophisticated tools for retrieving and integrating information, such as ENSEMBL (Hubbard et al., 2002). In addition to raw nucleotide sequence, they often contain information on the sub-elements such as locations of genes, exons, introns, regulatory elements etc.

- **Biological pathways.** Interactions of macromolecules and simpler chemical compounds create complex mechanisms that regulate and maintain the function of an organism. These mechanisms, called pathways, can be divided into categories such as metabolic, signalling or transcriptional pathways (Nelson and Cox, 2004). There exist databases that describe these mechanisms such as Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) and Reactome (Joshi-Tope et al., 2005). Similarly as functional descriptions, the pathway information can be used for interpreting high-throughput data for example for finding the pathways which are activated or repressed as a result of some treatment.

- **Scientific literature.** The literature databanks are sources of the most up-to-date information on genes. Thus also the curators of annotation databases obtain information there by manual reading or by using semi-automatic text mining tools. The most used source of literature in biosciences is the Pubmed databank which contains abstracts and metadata of hundreds of thousands of scientific articles published during the last decades including mostly MEDLINE articles. Pubmed includes sophisticated user interface with different kind of searching and retrieval tools and also an application programming interface (API) which enables queries using standard Common Gateway Interface (CGI) and Web-services protocols. In publication I we utilize Pubmed API to perform retrieval of abstracts for our text mining application.

Methods and tools for interpretation

Abundance of methods and tools exists that perform interpretation of high throughput screening results such as co-expressed genes. Most of the tools compare the frequency of gene annotations, obtained from some public database, in the user supplied gene list to the background set of genes. The background should include the rest of the genes, which did not fulfil the selection criteria, such as genes in chip which are not differentially expressed. The usual outcome from these methods is a sorted list of annotations that are considered important. These methods have been beneficial to data analysis by guiding the process towards the most important features in the gene list. In addition, the observation of multiple genes associated with the same annotation increases confidence in results obtained from high throughput methods.

Several software programs for finding over-represented GO annotations exist, such as SGDTermFinder (Boyle et al., 2004), GOToolBox (Martin et al., 2004), DAVID (Dennis et al., 2003; Huang et al., 2007) and EASE (Hosack et al., 2003). Often they either give a sorted list of significant GO terms or a graph showing the ontology associations between the most significant GO

terms as output. In publication II we have described method and associated GENERATOR software which represents an alternative approach for these methods by clustering the gene set into the subsets of genes which associate to similar GO-terms. It should be noted that a more recent feature of DAVID includes similar application of clustering large gene lists (Huang et al., 2007).

Also many tools exist for analyzing gene lists that use other kinds of data than GO-terms. For example DAVID includes tools for discovery of enriched attributes from KEGG pathway database, and also from several other databases (Dennis et al., 2003; Huang et al., 2007). Enriched chromosomal bands for a set of differentially or co-expressed genes can be found by using a tool available in dChip microarray analysis software (Li and Wong, 2003). In turn the over-represented sequence patterns are searched by the programs such as MEME (Bailey et al., 2006). This is similar to the POXO pipeline presented in publication IV. Simple text mining to see gene associated topics from Pubmed database can be done with PubMatrix (Becker et al., 2003) to interpret microarray results. This tool performs similar type of association searches as implemented in publication I.

More recent approach for finding over-represented attributes is the Gene Set Enrichment Analysis (GSEA) method (Subramanian et al., 2005). Instead of calculating over-representation of an attribute in the pre-selected set of genes, the method performs a walk down in the sorted list from the other extreme (e.g. over- or under-expressed end). Score for enrichment is then calculated for the gene subset created in the each position of walk. This has shown to detect many signals that were undetected or weak with the standard methods that use pre-selected gene list (Subramanian et al., 2005).

Another recent approach for analysis of gene sets is the Connectivity Map project (Lamb et al., 2006). Its principal aim is to facilitate finding associations between drugs, diseases and genes. Its main part is a reference database, including large number of results from gene expression screening performed for cultured human cells under the treatments of various treatments such as small molecules. The effects of such treatments are represented as signatures of over- and under-expressed lists of genes, referred to as reference signatures. Connectivity Map facilitates searching the reference signatures which correlate with the user given query signature. A query signature, i.e. user given sets of over-and under expressed genes, is used to search correlating reference signatures. The Connectivity Map approach has shown to recognize drugs with related actions, and to discover potential unknown mechanisms. However, many questions are still open, for example how many small molecules should be profiled to obtain suitable coverage (Lamb et al., 2006).

2.3 Biological problems in this thesis

Methods and tools developed in this thesis address the following problems in biological interpretation of genome wide data sets:

(i) Detecting the novel and expected findings in an experiment

In our DNA-microarray data analysis in publication I, we screened the expression levels of fish species rainbow trout (*Oncorhynchus mykiss*) under the treatments of different environmental toxins. As a result, we could obtain genes with abnormally high or low expression under treatment of each toxin. As a further step, we wanted to know, whether the same genes were affected by the same toxins in the previous studies. In order to answer this, we needed to search the scientific

literature for the effects of the toxins to expression of each fish gene we had on our microarray study.

Manual reading of thousands of toxin related articles, even only the abstracts, would have been too enormous an effort. Thus we decided to use a computational approach. We considered this as a problem for detecting strength of the associations between each gene-toxin pair in scientific texts. A gene and a toxin were considered to have an association when keywords describing the gene and the toxin appeared in the same abstract. An approach for determining the strengths of such associations is presented in publication I and chapter 4.1.

(ii) Theme discovery from gene sets

The usual outcome of genome wide screening is a set of genes with a particular characteristic, such as co-expression. The further step is to find the relating biological themes which could explain such characteristic. The previous tools for such analysis have reported the over-represented attributres in the user supplied gene list. This does not take into account, that a gene set often consists of separate gene groups, each responsible of a different biological process or a larger part of a single pathway.

In publication II, we aim to separate the gene groups that associate with separate biological themes, and analyze them separately. This is considered as a clustering problem, where genes are grouped according to associated biological attributes. The method is described in chapter 4.2.

(iii) Finding transcription factor binding sites from co-expressed genes

Transcription factors regulate the volume of gene transcription by enhancing and repressing the transcription. Therefore, the co-expression of a set of genes suggests that such set might be regulated by the same transcription factors. This can be studied in the pattern analysis of promoter DNA-sequences of co-expressed genes. The frequent patterns can be hypothesized to be putative biding sites of transcription factors. This problem is considered in publication IV, which describes a software tool series for searching TFBS regions from the co-expressed genes. This is briefly described in chapter 4.3.

(iv) Finding the chromosomal regions with co-expressed genes

In publication V, we have presented an approach for finding chromosomal regions containing co-expressed genes. Such analysis is important in many biological problems, such as study of chromatin remodelling, the dynamic changes in chromatin structure that open and close the chromatin structure, and thus regulate the transcription. Particularly, we focus on studying such phenomena in baker's yeast in publication V.

In our approach, we represent each gene of a chromosome as a single data point of sequential data, which is ordered according to genes' chromosomal locations. Each data point is a multinomial value, or vector, indicating gene's membership in a co-expression cluster. From this representation, we can find the regions with over-representation of some co-expression cluster(s). This is considered as a segmentation problem. The method is described in chapter 4.4.

3 Methodological background

In this chapter, the methodology related to this work is reviewed. This includes unsupervised machine learning techniques, clustering and segmentation, and statistical modelling and evaluation used with such methods.

The chapter is divided into three subsections. Section 3.1 describes the representation of data for biological problems described in chapter 2.3. Section 3.2 discusses algorithmic techniques that are used for clustering and segmentation. Section 3.3 reviews some related statistical modelling techniques that can be applied when the partitioning of data is considered as a statistical model.

3.1 Data representation

Methods developed in this thesis analyze categorical binary or multinomial data that indicates associations between genes and biological attributes. Such attributes represent GO terms (publications II, III and IV), text from scientific literature (publication I), or biological conditions of the experiment series (publication V). The data can be represented as matrix *X*:

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1M} \\ x_{21} & x_{22} & \dots & x_{2M} \\ \dots & \dots & \dots & \dots \\ x_{N1} & x_{N2} & \dots & x_{NM} \end{bmatrix}, \qquad (3.1)$$

where the *N* rows indicate genes (referred also to as data points, objects or observations) and *M* columns indicate the biological attributes (referred also to as variables, attributes or dimensions). In this thesis the data is often denoted just with symbol X or optionally with symbol *D*.

In a common clustering problem, the data points are exchangeable and thus also denoted with $X = \{x_1, x_2, ...x_N\}$ indicating a set of objects. In such a set, each data point $x_i = (x_{i1}, x_{i2}, ...x_{iM})$ is an *M*-dimensional data vector. In the segmentation problem the data is sequential and thus indicated with $X = (x_1, x_2, ...x_N)$.

3.2 Review of clustering and segmentation methods

This chapter reviews the methodology used for data clustering and segmentation. These unsupervised machine learning techniques are central to the analysis of gene expression data and associated data mining in this thesis (in all included publications). The algorithms for clustering and segmentation are reviewed in sections 3.2.1 and 3.2.2. Section 3.2.3 focuses particularly on techniques used with high dimensional data sets. Section 3.2.4 reviews methodology used for evaluation of clustering and segmentation methods.

3.2.1 Clustering

Clustering produces a partitioning of data where similar objects are deposited into the same group and dissimilar objects in different groups. Optimal clustering is *a NP-complete* procedure for which the fastest optimal solution runs in not less than exponential time (Garey, 1982). Therefore heuristic

methods are generally used. The heuristic methods are often sub-divided into hierarchical and partitioning methods (Jain and Dubes, 1988), the latter of which includes optimization based methods as a subgroup. In addition there are methods that do not fit fully or at all into this categorization. These include graph theoretical clustering (such as Shamir and Sharan, 2000), unsupervised methods that are based on neural networks (such as Kohonen, 1995), density based (such as Sander et al., 1998) and model based clustering (such as McLachlan, 1988).

Often the data clustering is based on measuring similarity between the data point pairs. The standard similarity measure is *Euclidean distance*. This and *Pearson correlation* are commonly used in gene expression data clustering (see for example analysis of gene expression data in publications I, III and V). With discrete data, such as binary data, measures such as czekanowski-dice distance can be used (see for example Martin et al., 2004). Moreover the probabilistic and information theoretic measures have been widely used with discrete and symbolic data objects (see chapter 3.3). A good review of similarity measures is given in Spertus et al. (2005).

Another very common problem in clustering is determining the appropriate number of clusters. This is often considered as a problem of statistical modelling, which is directly addressed by model based clustering algorithms. This particular problem is discussed further in chapter 3.3. Below, the clustering problem and the most common clustering methods are briefly described with the references to contributions in this thesis.

Problem description

In a general clustering problem, *M*-dimensional data $D = \{d_1, d_2, ...d_N\}$ is concerned with *N* data points. Clustering produces a grouping of such data into *k* different partitions. Thus the data points can be assigned with cluster membership indices $I = \{I_1, I_2, ...I_N \mid I_i \in 1..k\}$. Clustering solution with *k* clusters can be considered as a set of subsets $D_k = \{D_k^{(1)}, D_k^{(2)}, ...D_k^{(k)}\}$. Each subset $D_k^{(i)} = \{d_j \mid I_j = i\}$ represents cluster *i* which includes the data points with membership indices equal to *i*. There exist

$$S(N,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^i \binom{k}{i}(k-i)^N \qquad (3.2)$$

different non-empty clustering solutions for *k* clusters (Stirling number of the second kind) (Virmajoki, 2004). Thus, a search over the whole solution space has exponential time complexity.

Optimal methods

Algorithms for obtaining the optimal clustering solution (according to some objective function) have been rarely discussed in the literature (Fränti et al., 2002). The algorithm presented in (Fränti et al., 2002) creates a graph of all possible clustering results and applies the branch-and-bound technique (Fukunage and Narendra, 1975) to detect the optimal clustering solution, for a given goodness criterion. This algorithm has been still reported to have an exponential time complexity, which limits its usage to small data sets (such as <100 data points) and a small number of clusters (such as <6 clusters).

Partitional methods

Partitional clustering methods group data into explicitly distinct groups (Jain and Dubes, 1988). The simplest partitional methods consider each data point only once, in a greedy fashion. Often they require the user to give the minimum distance as a parameter, which they use to decide whether a data point is assigned into a considered cluster or not (Virmajoki, 2004). The results are often weak as the solution depends largely on the minimum distance parameter and the order in which the data points are deposited into each cluster. One example of this is for example the *nearest neighbour clustering* (presented for example in Jain and Dubes, 1988).

In optimization based methods, such as traditional *k-means clustering* (MacQueen, 1967) and its different variants, data points are grouped into pre-selected *k* number of groups by using iterative optimization of some objective function. The k-means algorithm starts by creating *k* centroid vectors, which are often randomly initialized. The following two optimization steps are then iterated: 1) assignment of data-points into clusters with the nearest cluster centroids, and 2) updating of cluster centroids according to assigned data points. The steps are iterated a given the number of times or until the algorithm converges. K-means converges into the local optimum which depends completely on the initialization of centroid vectors and the number of clusters. Because of the often fast convergence of k-means algorithm, it is often repeated several times to produce several locally optimal solutions of which the one is chosen using some evaluation measure (Steinley, 2006).

Another method with a strong relation to k-means is *Self Organizing Map* (SOM) (Kohonen, 1988; Kohonen, 1995). The difference between SOM and k-means is that SOM defines a neighbourhood function which results in the establishment of similar neighbouring clusters. The update of each cluster centroid depends on the neighbouring clusters according to a neighbourhood function. The update effect is weakened as a function of distance from the main updated cluster centroid.

We have used partitional clustering such as k-means in publications I, II and V for analysis of gene expression data in order to find the co-expressed genes. Partitional approach which uses a matrix factorization for clustering (see 3.2.3) is also used for creating individual partitions to the developed non-nested scheme in publications II, III and IV.

Hierarchical methods

Several different hierarchical clustering methods are available (see for example Kaufman and Rousseeuw, 1990). Traditional hierarchical methods start with either the whole data or each observation as a single partition. The first approach is called divisive hierarchical clustering as it divides the data into two nested sub-partitions iteratively until each observation forms a single partition. The latter is called merging or agglomerative hierarchical clustering (Lance and Williams, 1967; Johnson, 1967) because it merges the two nearest partitions until the whole data forms a single cluster.

In agglomerative methods, the distance (or dissimilarity) between two clusters is defined using some distance measure (or dissimilarity measure), such as *Euclidean*. The standard approaches to determine such distance between two clusters are *single linkage*, *complete linkage* and *average linkage*. Single and complete linkage methods define the distance between the two

clusters as a distance between their nearest and farthest data points, respectively. Average linkage uses the average of distances between all possible data point pairs. In addition, in *Ward's method* (Ward, 1963) the merged clusters are selected so that the increase in within-clusters squared sums, or within-clusters variances, is minimized. A broad description of these methods and their properties is given in Olson (1995).

As a result, the hierarchical methods produce a set of nested partitions which is often visualized with a binary tree called a dendogram, where each node represents a single partition. The edges of the dendogram represent the division or merging procedures and their lengths usually correspond to the distances between the divided or joint clusters. Such visualization is advantageous as it gives an overview of the otherwise complex output, highlights hierarchical relationships in data, and reveals dense clusters that stay similar in multiple levels (do not decrease size in divisive and do not increase size in agglomerative clustering). See for example Mikheev et al. (1988).

The disadvantage of hierarchical clustering is the greedy proceeding approach that optimizes the objective function only locally in each merge or split (Virmajoki, 2004). Moreover, the partitioning result obtained is not explicit as some further method must be used to find either coherent individual clusters or a clustering solution. In addition, the problem with standard hierarchical clustering using single-linkage distance has been the chaining effect, a problem where separate clusters are joined through a sequence of mediating data points (Wishart, 1968).

We have used the agglomerative hierarchical clustering with Ward's method in publication I to get an overview of the correlation of gene expression response in salmon fish between different chemical contaminants. Pearson correlation was used as a similarity measure for ranks of gene expression levels.

Graph theoretical clustering

Graph theoretical clustering considers data points as vertices of a graph and the distances or similarities between the data points as graph edges. The aim is to partition the graph based on the similarity matrix obtained from the distances i.e. the edges. An example of graph theoretical clustering is the CLICK algorithm (Shamir and Sharan, 2000) that forms sets of close vertices referred to as kernels, and in the iterative process extends them.

Model based clustering

Model based clustering (for example McLachlan, 1988) is a probabilistic approach that considers data as samples obtained from an unknown number of populations with unknown parameter values. Such populations can be represented by a mixture of the probability density distributions that are assumed to follow some pre-defined parametric family such as Gaussian. The unknown parameters of this mixture density, such as means and (co)variances, are estimated (see review of parameter estimation in chapter 3.3) in the clustering process.

In model based clustering, the appropriateness of a clustering solution is measured rather as a function the fit of data with the mixture model than as a function of the distances between the data vectors (Kontkanen et al., 2003). Such a model facilitates also the direct evaluation of solutions with different numbers of clusters (Oh and Raftery, 2003). Optimization of model parameters is often performed by using the Expectation-Maximization algorithm as in Pan (2006) or in Fraley and Raftery (2007).

3.2.2 Segmentation

Techniques that separate homogeneous segments in sequential data are known as segmentation, partitioning, or change-point analysis (Li et al., 2002). Applications of segmentation include analysis of observations obtained at different time points from financial stock data (Hsu, 1979), and dividing DNA sequences into homogeneous parts (Liu and Lawrence, 1999; Bernaola-Galvan, 2000). There exist mainly three types of segmentation algorithms based on dynamic programming, its approximation, and heuristics (see description and references below; see also review of segmentation methods used in the data mining literature in Keogh et al., 2003). In addition, online methods exist, that are suitable for real time data streams (e.g. Fearnhead and Liu, 2007). In this work, we use heuristic method in a non-online segmentation problem (publication V). Our segmentation aims at separating genome regions with co-expressed genes from each other, and from surrounding areas.

Problem description

In this work, sequential data D_t is considered, where the time parameter t varies from *1* to a fixed N. Our general model for D_t assumes that they are independent and originate from k different segments. Data in each segment follows a distribution of some parametric family (multivariate multinomial in publication V) with parameter values that differ from other segments. The general aim of the segmentation is to find the unknown change-points, i.e. the locations in the sequence where a segment changes to another. Segmentation into k non-overlapping segments is often referred to as *k-segmentation problem*. Typically, as in this work (publication V), the number of segments (k) is also unknown and should be estimated. These unknowns are typically considered as parameters of a statistical model representing the data segmentation.

Segmentation can be considered as a special case of clustering, where data points in the obtained clusters must overlap along the time dimension. The number of different segmentation solutions for k segments ($m = k - 1$ change-points) is (as stated in publication V):

$$S(N,k) = \binom{N-1}{k-1}. \quad (3.3)$$

A search over all necessary solutions can be reduced from the exponential time of a common clustering problem into quadratic time (see *Optimal methods* section below) with dynamic programming as shown in Bellman (1961b).

Optimal methods

A dynamic programming based method (Bellman, 1961b) enumerates over the whole segmentation solution space by taking into account only the set of non-overlapping sub-solutions of complete segmentation solutions. This is advantageous, as the algorithm finds the optimal (according to the used cost function) solution to the *k-segmentation* problem in time $O(kN^2)$, if the cost can be calculated in linear time. As an advantage, dynamic programming based methods facilitate the "fully" Bayesian approaches, where integration over probability distribution representing the

probabilities of different segmentation solutions is performed. Such method is presented for example in Liu and Lawrence (1999).

Some studies argue that dynamic programming algorithm is too slow with long sequenced data (Himberg et al., 2001) whereas others have argued that it is still within the limits of feasibility (Kehagias et al., 2005). However, it should be clear that in the applications that require fast processing, include large data sets, and/or large number of change-points, only the methods that use approximation or heuristics are practical. This is also the case in publication V (see also chapter 4.4), where we have designed a segmentation method, which we aim to implement in a Web based software for analysis of quite large datasets.

Approximate methods

Approximate algorithms aim typically at approximating the results of dynamic programming based segmentation (Guha et al., 2001; Salmenkivi et al., 2003; Terzi and Tsaparas, 2006; Fearnhead, 2006). In addition to change-point estimation for fixed segment number, the recent methods concern also the unknown segment number (e.g. Salmenkivi et al., 2003 and Fearnhead, 2006).

Divide-and-Segment (Terzi and Tsparas, 2006) is an approach which performs a dynamic programming based segmentation for the initially chosen borders. This can be used to reduce the running times drastically.

A recent method presented by Fearnhead (2006) performs a direct simulation from posterior distribution for change-points and their number in quadratic time. Importantly, its approximation can complete in roughly linear time. Also, an online version of this method has been presented (Fearnhead and Liu, 2007).

Heuristic methods

Heuristic segmentation algorithms aim typically at excluding the sub-optimal steps from the algorithmic search (publication V). They facilitate the analysis with large sized and large dimensional data, and with large amount of change-points (Himberg et al., 2001). Hierarchical methods, based on similar binary splitting or merging as in hierarchical clustering, have been widely used. Typically, these methods produce several alternative solutions with different number of change-points, and require using some evaluation criterion for choosing the most proper.

Hierarchical Top-Down segmentation, or binary segmentation, was originally developed independently by multiple authors (Ramer, 1972; Douglas and Peucker, 1973; Sen and Srivastava, 1975). The algorithm creates repeated 1-to-2 segmentations for a given data sequence, and for the resulting sub-sequences, until each data point is surrounded by two change-points, or some stopping criterion is fulfilled. The algorithm is fast, as it runs in time $O(kN)$. A recursive version of the algorithm has been popular in several fields including bioinformatics (Bernaola-Galvan, 2000). It performs the segmentation in the fixed order of sub sequences, for example the left sub-sequence first and then the right one. The non-recursive version proceeds to the segmentation of a sub-sequence that increases the global cost the least. This is advantageous, as each solution can be considered as a representative for that segment number, unlike in the recursive algorithm. We use this method in publication V.

Circular Binary Segmentation (CBS) (Olshen et al., 2004) is an improvement of the original recursive binary segmentation algorithm. It is based on the observation that the original algorithm

leaves large segments with one changed sub-segment in the middle intact. CBS adds an extra step to binary segmentation, where each obtained sub-segment is considered separately with both ends attached together forming a closed circle. Then a statistical test for single square wave change is applied to the circle in order to detect a significantly changed sub-segment. Algorithm uses significance test as stopping criterion for recursion. CBS has been applied to CGH array data (Olshen et al., 2004).

The Bottom-Up (e.g. Keogh and Smyth, 1997) segmentation starts with all data points surrounded by segment borders. In the next step, the two most similar segments are merged. Such a merging is repeated until the complete data is in single segment or a stopping criterion is fulfilled. The algorithm is not as fast as the Top-Down method, as it runs in time $O(kN \log N)$).

Hidden Markov Model (HMM) based segmentation presented in Fridlyand et al. (2004) considers the unknown means at each segment as hidden states and the change-points as transitions between such states. A search of change-point locations is performed with unsupervised training of HMM on data, and then choosing the most probable state sequence. The number of states is decided by using Akaike Information Criterion (AIC; see *Model selection* in chapter 3.3).

3.2.3 Methods for high dimensional data

The typical objective of unsupervised machine learning, such as clustering, is to discover dense regions from data. The difficulty of such analysis increases drastically with the number of dimensions of space where the data points are located in. If we consider and object in the space, its (hyper)volume is exponentially grown as a function of dimensionality. This problem is known as the *"curse of dimensionality"*, coined originally by Richard Bellman (1961a).

In order to illustrate the curse, we can consider an imaginary M-dimensional hypergrid that constitutes of neighbouring hypercubes in (restricted) high dimensional (HD) space. We can consider the search of dense regions as the task of estimating the density of data-points within each of the hypercubes, so that the entire space will be mapped with some fixed precision. The first implication of Bellman's principle is that the number of such hypercubes needed for mapping the space grows exponentially with the number of dimensions (Bellman, 1961a). This has a large impact on the running time of the search and the needed memory storage (see for example Eccles and Su, 2004). The second implication of the "curse" is that with a fixed number of data points, the accuracy of density estimation decreases when the dimensions and hypervolume increase (Bellman, 1961a). In other words, the required number of data points for observing an equally dense cluster increases exponentially with the dimensionality (as stated in Koeppen, 2000). Thus, with very high dimensional data, the number of training samples should be extremely high to produce accurate estimates (Koeppen, 2000). Unfortunately, a relatively small amount of data is only available for many high dimensional analysis problems. For example, the data sets representing the associations between genes and GO-terms in publication II include nearly the same amount of dimensions as there are data points.

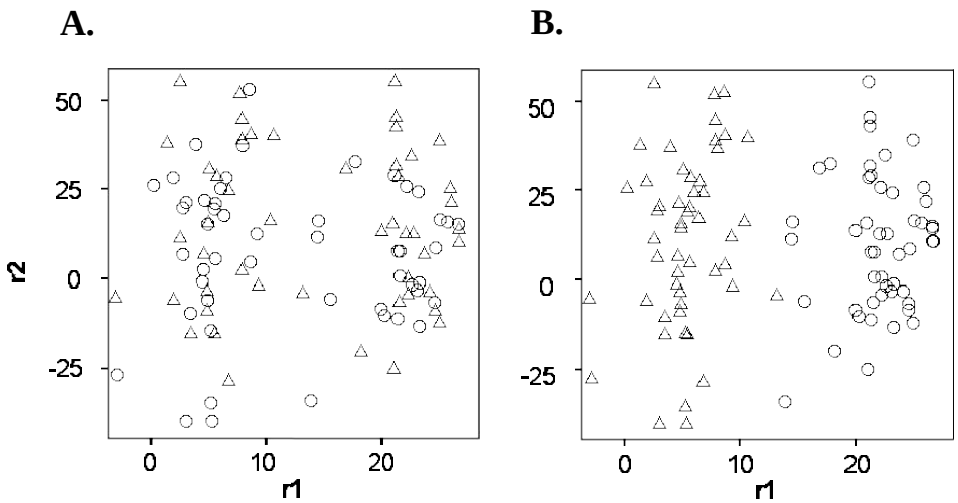

Figure 5. Clusters in subspace. Three dimensional artificial data where data points form two clusters only in the first dimension *r1* but not in the second and the third dimensions *r2* and *r3*. For clarity the dimensions *r1* and *r2* are only visualized. A: K-means clustering using Euclidian distance with 100 iterations was performed for data using data from all three dimensions. The resulting clusters are indicated with circles and triangles. Although the two clusters are clearly visible for the viewer, the clustering with all dimensions does not correlate with the groups. B: When k-means is performed only for the first dimension the two clusters are clearly separated by clustering.

Several methods exist for analysis of HD data that simplify the problem of dimensionality in order to facilitate analysis in lower dimensions. They work either by reducing dimensionality prior to analysis (see *Dimension reduction* section below) or optimizing the relevant set of dimensions during the learning algorithm (see *Subspace and biclustering* section below). One simplification is based on the observation that the clusters in HD data tend to exist only in the subspaces of all attributes. The irrelevant dimensions often overwhelm such patterns and prevent their discovery as shown in figure 5. This is emphasized in the similarity analysis of two binary vectors, where the values are either same or not, without intermediate forms (when we assume that the similarity of two zeros is equal to the similarity of two ones and the similarity between one and zero is same as between zero and one). Moreover, the change of one or a few bits may have drastic effects on the distance between such data points, depending on the used similarity measure (Kontkanen et al., 2003). Another simplification is based on observation that there often exist correlating dimensions, suggesting the existence of a lower dimensional manifold structure of data. By producing a combination of a set of variables, which is often referred to as *latent variable*, it is possible to reduce the dimensionality without losing much of the information content in the data.

Dimension reduction

Dimensionality reduction methods have been used to reduce the number of variables to facilitate the further analysis or visualization of data. Their principal objective is usually to reduce data dimensions with minimal information loss and maximal noise and redundancy loss (Fodor, 2002).

The practical examples where dimensionality reduction is utilized include topic analysis or text mining (Bingham et al., 2002; Seppänen et al., 2003), noise reduction for clustering analysis (Ding et al., 2002), application directly for clustering analysis (Xu et al. 2003), and image analysis such as image classification in numerous different sources.

Dimension reduction is often divided into *feature extraction (feature transformation)* and *feature selection* (Fodor, 2002). The first considers methods that map the data into lower dimensional space. Multidimensional scaling (MDS) is a set of such techniques, traditionally used for visualization of HD data in 2D or 3D. These methods use the dissimilarities between the pairs of objects to map them into lower dimensional space. The classical MDS method presented in Torgerson (1958), minimizes the loss function referred to as *strain*. Thereafter, several MDS methods have been introduced that optimize different loss functions (Cox and Cox, 2001).

Another example of feature extraction is *principal component analysis* (PCA). In PCA the selected number of the orthogonal components is fitted into data so that the variance of data point projections on each component is maximized (Fodor, 2002). Because of orthogonality constraint, PCA can be interpreted as the rotation of standard coordinate axes which explain the most variance in data and fitting them into data. Usually PCA is computed by generating a covariance matrix for the input dataset, and obtaining the *eigen vectors*, the *principal components*, of covariance matrix. The associated *eigen values* report the amount of variance explained by each component.

PCA and related *singular values decomposition* (SVD) produce the optimal decomposition of data with respect to MSE. A recent method, *Independent component analysis* (ICA) (Comon, 1994), enforces rather statistical independence than maximal variance. ICA has been often performed by using the Infomax algorithm (Amari and Cichocki, 1996) or FastICA (Hyvärinen, 1999) to search the Maximum Likelihood (ML) estimate for the maximal independence. Also Bayesian versions of ICA have been developed (Rizwan et al., 2001).

Moreover, the mixture models (discussed in the context of model based clustering in chapter 3.3.1) have been also applied in representation of data with latent variables (Verbeek, 2004). In that case, the latent variables represent the joint mixture densities of the probability densities of original variables. A popular approach to optimize the mixture model is the Expectation-Maximization algorithm (Dempster et al., 1977).

Non-negative Matrix Factorization (NMF) is a recent feature extraction method, developed in (Paatero and Tapper, 1994; Lee and Seung, 1999). NMF algorithms are linear factorization methods with positive constraint. Thus, unlike PCA, ICA, SVD or VQ they search the components for positive data so that as a result the projections of observations on components have also positive values when the components are considered as a new axis. Although NMF forces sparse rotation of components it does not require orthogonality of components. NMF algorithm presented in Lee and Seung (2001) is used in publication II for clustering binary data created from associations of genes and database attributes. See more discussion in chapter 4.2.1.

Feature selection methods include variable ranking methods and variable subset selection methods (Guyon and Elisseeff, 2003). The ranking methods, also referred to as *filters*, select variables by ranking them with correlation coefficients. Subset selection methods include *wrappers* that assess subset of variables according to their usefulness as predictors. An extensive review of feature selection methods is given in Guyon and Elisseeff (2003).

Subspace clustering and biclustering

With high dimensional data, only a certain subset of data attributes or their combination is often relevant for each cluster. Subspace clustering is an approach that aims to detect such subspaces during the optimization of cluster borders for data points. Subspace can be considered either as a subset of original dimensions, referred to as *axis-parallel space*, or as a combination of them, referred to as *general space* (Patrikainen and Mannila, 2004). It should be noted that the first has similarity with the feature selection problem and the latter with the feature extraction.

The first subspace clustering algorithm was CLIQUE (Agrawal et al., 1998). It addresses the axis-parallel space as the most of the existing methods for subspace clustering (Patrikainen and Mannila, 2004). Currently, there exist several other methods for continuous and discrete data.

Biclustering is closely related to subspace clustering. The aim is to create clusters of data points and attributes simultaneously, and as a result find "blocks" from the hypothetical matrix of data with data points on rows and attributes on columns. Several different biclustering algorithms have been used in bioinformatics previously (Getz et al., 2000; Sheng et al., 2003; Carmona-Saez et al., 2006).

In publication II, we have introduced a NMF based approach for clustering genes using associations to the database attributes, obtained from the GO-database, as data. This approach has similar aspects with subspace clustering as the NMF tends to capture the correlations lying in the subspaces of data.

3.2.4 Method performance evaluation

A critical part of any study developing computational methods is to evaluate the method performance in the light of other existing methods. Such an evaluation often comprises computational aspects such as time and space complexity and statistical aspects such as accuracy of the learned solution. In this thesis, the latter aspect is the main concern.

With supervised machine learning methods, the procedures such as cross-validation exist that are valid for evaluation of obtained solutions (Candillier et al., 2006). With unsupervised methods, such as clustering or segmentation, the evaluation is more problematic as it tends to be subjective by nature (Candillier et al., 2006). A standard approach for evaluation of an unsupervised method is to use artificial data. The parameters in the data generating model are known, which facilitates the evaluation of learning results. Another way is to use real data. In such a case, the evaluation can be based on auxiliary labels of data points if such information is available. The third possibility is to use a human expert to interpret the meaningfulness of different results in order to find the best one. Below, the usage of artificial data and labelled real data in the evaluation are briefly reviewed.

Using artificial data

Method evaluation using the artificial data approach can be considered as a multi-step procedure. First, a model is created for generating data. Such a model can define various parameters for created data, such as the number of clusters created, size of clusters and data, relationships of clusters, skew of data, difference between clusters, variance of data etc. Secondly, datasets are created from the

model, often using different parameter values in creation. Third, the compared methods are applied in the data. Fourth, an evaluation is carried out in order to find out how well the methods find the created clusters and often how well the method performs under different values of parameters, such as cluster number and cluster sizes.

A standard way for evaluating the partitioning result created for artificial data is to use some measure which is based on comparison of gene memberships between a partitioning solution and the "true" partitioning of data. Such measures study indices of cluster membership pairs like Rand Index (Rand, 1971) and adjusted Rand Index (Hubert and Arabie, 1985), frequencies of matching and mismatching objects like classification accuracy, and difference between probability distributions defined by gene memberships like Mutual Information (Shannon and Weaver, 1949). These methods have been often used for evaluating and comparing partitioning methods such as clustering or segmentation (Yeung and Ruzzo, 2001; Verma and Meila, 2003; Schenker, 2003; Kuncheva et al., 2006). While each of these approaches have shown to have its own inconsistency (Meila, 2003), the common shortcoming is that these methods do not take into account the uncertainty related to probabilities of data in the model that generates the data and in the model that represents the obtained partitioning solution.

A shortcoming of methods that compare gene memberships can appear in multiple cases. For example when two partitions are generated which are very homogeneous in the data, it would be natural to consider them as one partition in the model that generates the data and in the obtained partitioning solution. This follows also from the Occam's Razor principle (Myung and Pitt, 1996), according to which the model (phenomenon) should be presented with as little parameters (assumptions about the phenomenon) as possible. This is not taken into account when comparing gene memberships but rather all partitions are treated equally. Another case when similar merging should be adequate is when there are several very small partitions generated in the data. It would be often natural to represent such data with a smaller number of partitions than in the generating model.

The uncertainty mentioned above can be taken into account when the model generating the data and the obtained segmentation solution are considered as statistical models which treat probabilities of data classes as parameters. Such an approach is discussed more in publication V (see also chapter 4.4.1) which presents a benchmark for evaluation of different segmentation methods.

Evaluation using auxiliary data

When real data is used to evaluate a method's performance, one option is to obtain related auxiliary data, and use it for validation. In this case, it is possible to perform exactly similar calculation of classification error as with supervised methods. The shortcoming is that the labels that would be the basis of training the supervised method are not necessarily the classes that should be found by the ideal learning algorithm (Candillier et al., 2006). Still it can be interesting to see the correspondence between the two classifications. The comparison of clustering methods for microarray data has been often performed using GO database records as data point labels (see for example Datta and Datta, 2006).

We perform evaluation with labelled auxiliary data (known yeast gene relations with cell cycle phases) to evaluate the segmentation method presented in publication V (also discussed in

chapter 4.4). Still, this is not used for comparison of several methods, but rather only to study the biological significance of results obtained from the developed method.

3.3 Overview of related concepts in statistical modelling

Statistical inference, in general, involves the use of statistics to infer unknown aspects from samples taken from a population. In data mining, statistical inference is commonly used for evaluation of hypotheses in the course of algorithmic search and evaluation and understanding of obtained solutions (Glymour et al., 1996). Such evaluation is frequently used in clustering or segmentation, to address the uncertainties of obtained solutions. The term s*tatistical modelling* is used as a designation for applying statistical inference in the evaluation of statistical models and their parameters (Nikkilä, 2005). The use of statistical modelling with machine learning algorithms is also referred to as *statistical machine learning* (Nikkilä, 2005).

This chapter reviews some of the techniques used commonly in statistical modelling within machine learning: estimation of model parameters, choosing a suitable model, and testing significance of hypotheses or models. These techniques are reviewed in the light of classical and Bayesian schools. References are given to the contributions in this thesis.

Parameter estimation

Parameter estimation is the process of trying to estimate the correct value for a term in a physical model by studying the output of trial simulations and selecting values in such a way that the model duplicates experimental observations as closely as possible (Ewing et al., 1994). In statistical modelling, there is often a restricted family of parametrically specified distributions. Such family is usually regarded as *a statistical model* (McCullagh, 2002).

Parameter estimation concerns often the approximation of data generating distribution (or population distribution) based on the empirical data. A parameter is usually considered as an unknown value that describes some characteristics of adistribution, such as average or variance. Classically, the concept of a parameter is distinguished from the other unknown aspects, such as missing data or a statistical model itself, by assuming that parameter is fixed in number and describes the whole population rather than only a part of it (Liu and Lawrence, 1999).

A standard approach to address the relation between a model, its parameters and data is the *likelihood* principle (Ewing et al., 1994; Forster, 2000; McCullagh, 2002; Nikkilä, 2005). The likelihood addresses the probability of observed data being produced by a particular model with particular parameters. Thus, the likelihood is the function of parameters as follows:

$$L(\theta) = p(x_1, x_2, ..x_n \mid M, \theta), \quad (3.4)$$

i.e. the conditional probability of data given the parameters $\theta = \{\theta_1, \theta_1, ...\theta_m\}$ and the model *M*. The likelihood function for a parameter of exponential family distribution is generated from the parametrized probability densities or probability masses $f(x \mid \theta)$. For such density (or mass), the likelihood is the function of parameters θ with fixed *x* representing the observed data (e.g. observed successes with binomial data).

Procedures for making parameter estimation are called estimators (Glymour et al., 1996). A classical method for parameter estimation is the Maximum Likelihood Estimator (MLE). Its idea is to search the values for parameters that maximize the likelihood function as follows:

$$\theta^{mle} = \arg\max_{\theta}(\log p(x_1, x_2, ..x_n \mid \theta)) . \quad (3.5)$$

Often a logarithm of the likelihood is taken (as above) as it is a monotonically increasing function and eases up some calculations. This is referred to as log-likelihood.

An important aspect related to estimators is the *bias and variance principle* (see for example Neville and Jensen, 2007). Bias of an estimator measures how far the expected value of an estimator goes from the expected value of a true population parameter. A small bias is desirable since a small biased estimator can produce more accurate results (Neville and Jensen, 2007). Variance, in turn, measures how far each estimate is from the expected value of an estimate, i.e. how much different estimates vary. A small variance is also desirable as estimator with small variance can produce more stable results (Neville and Jensen, 2007).

There is a known trade-off between bias and variance so that the added bias decreases variance (see for example Yu et al., 2006 or Neville and Jensen, 2007). That is, an unbiased estimator may adjust "too well" into the sample data which can lead to weak results, especially with small sample sizes. The bias added to an estimator is utilized for example in Bayesian inference in a form of *prior probability* of an event (see discussion for example in Domingos and Pazzani, 1997). Similar pre-observed information can be added as pseudo-counts for ML estimate, as done for example in Lawrence (1993). The bias-variance trade-off is also closely related to model selection discussed later in this chapter.

In this thesis, MLE based methods are applied in optimization and evaluation of segmentation model in publication V. See further discussion on this topic also in chapter 4.4.

Model selection

Statistical modelling often involves selecting a suitable model (see definition of model above in this chapter) among a collection of viable candidates. The competing models contain usually varying numbers of parameters or other unknowns (Glymour et al., 1996). Such a situation is addressed by the methods included under the topic *statistical model selection* (Pizarro et al., 2000). A general aim of model selection methods is to obtain a proper balance between the goodness of model fitness in data and model simplicity (parsimony). This is the principle often referred to as *Occam's razor* in the literature (Myung and Pitt, 1996). This has also close relation with the *bias-and-variance* principle discussed in the context of parameter estimation above in this chapter. That is, the bias tends to decrease and variance increase with the increased model complexity, and vice versa, as shown in Geman et al. (1992).

In a clustering (or segmentation) framework, the model selection is often used to choose a suitable number of clusters (Li, 2002; Mannila and Patrikainen, 2004; publication V). This involves considering the partitioning solution as a statistical model. Parameters of such a model represent the cluster borders and the data within the partitions (see publication V). This implicates that with the number of clusters the number of model parameters, i.e. the model complexity, increases. By

increasing the cluster number the fitness refines but the generalization with new or future data becomes weaker. In turn, too low a cluster number leads to a poor degree of fitting to any real features. Model selection for partitioning methods is discussed further for example in Breiman et al. (1984).

Standard methods of model selection contain classical hypothesis testing, Bayesian inference, minimum description length principle (MDL), empirical cross-validation and different information criterions (listed in Glymour et al., 1996 and Forster, 2000). These approaches provide an implementation of Occam's razor and can be considered from the bias-variance-complexity perspective. This aspect is shown for several model selection methods in Yu et al. (2006). In unsupervised machine learning, the information criterions are in frequent use, mainly probably due to their simplicity.

Akaike Information Criterion (Akaike, 1974) is a simple measure which approximates the Kullback-Leibler distance (Kullback and Leibler, 1951) between a model and assumed true model. It penalizes the log-likelihood of the data and model with the amount of parameters in the model. For a general model AIC is defined as follows (Li, 2002; Zhang, 2005):

$$AIC = -2\log(\hat{L}) + 2K\,, \qquad (3.6)$$

where the first term of summation is the log-likelihood and the second is the number of model parameters *K*, both multiplied by two.

Bayesian model selection is based on the posterior probability of a model given the data (see more detailed discussion later in this chapter). Depending on the number of parameters in the model, the integration can become a computationally hard task. In 1978, Schwarz published a method called Bayesian Information Criterion (BIC; Schwarz, 1978) which computes very simply an asymptotic approximation of posterior probability, by using maximum likelihood and a penalty term:

$$BIC = -\log(\hat{L}) + \frac{K}{2}\log(N)\,. \qquad (3.7)$$

where the penalty term depends also from data size *N*. Modified versions of BIC have also been developed into different applications such as Bayesian networks (Rusakov and Geiger, 2002) and segmentation (Zhang, 2005).

A shortcoming of many approximation based model selection methods such as AIC and BIC is that they rely on some simplifications and are thus restricted to some particular applications (publication V). Fully Bayesian methods are reported consistent in many contexts but can suffer from computational complexity (Glymour et al., 1996). This problem is further discussed in chapter 4.4 and publication V, where we compare model selection methods in segmentation framework.

Hypothesis testing

Hypothesis testing can be considered as a one-sided estimation, in which a testing rule either conjectures that the hypothesis is false or makes no conjecture (Glymour et al., 1996). A single

hypothesis can be viewed as a statement that proposes a model for the distribution of observed data. The testing then interrogates whether the sampled data conjectures that the model is false. As such, the hypothesis testing can be also considered as a tool of model selection.

With classical binary hypothesis tests there are two hypotheses one of which often represents the absence of some element like signal. This is referred to as *null hypothesis*, denoted often as *H0*. The a*lternative hypothesis* states that the element, such as signal, exists and is often denoted with *H1*. *H0* proposes the *null hypothesis model*, also called *a zero model*, denoted with *M0*. Thus the classical approach is often referred to as *null hypothesis significance testing* (NHST) in literature (Nix and Barnette, 1998). Traditional NHST seeks to gain evidence on *H1* in order to reject *H0* (Nix and Barnette, 1998). Rejection is based on preset α-level threshold and resulting p-values from tests. A distinct approach from NHST is e.g. Bayesian inference which represents models for both null and alternative hypotheses (Törönen, 2004).

NHST and the interpretation of its results has been widely criticised for many reasons in the recent literature (Frick, 1996; Nix and Barnette, 1998). These relate mostly to the facts that NHST does not report the size of the difference and that it is practically often possible to reach statistically significant difference by only increasing the sample size (Nix and Barnette, 1998). This has affected the alternative and additional measures, like confidence interval and effect size, have been taken in use. Still, it is argued that significance testing is suitable in many applications, such as situations when effect size is irrelevant (Frick, 1996).

A frequently used test scheme is testing equality of means or rates of two or more sample groups. Standard tests for testing means of continuous data are different t-tests and *analysis of variance* (ANOVA). These were briefly reviewed in the chapter 2.2.2. In this thesis we use these tests for selecting differentially expressed genes (see chapter 2.2.2 for specific references). Testing the equality of two rates is known as a *rate problem* in literature (Lee and Pope, 2006). In such case, the null hypothesis *M0* involves "same rate model" that assumes similar rates for compared groups. Standard tests for this purpose are Binomial test and Fisher's Exact test (Fisher, 1922). The first uses binomial distribution as a null model whereas the latter involves hypergeometric distribution. The tests (one sided) are calculated by summing the tail of distribution for observed number and more successes (alternatively *x* or less successes) (Lancaster, 1961). The resulting p-values report the probability of obtaining *x* or more (alternatively *x* or less) successes by chance. Fisher's Exact test represents *sampling without replacement* i.e. a situation that the sampled data would be picked one by one without returning back to the population. In turn the binomial test represents *sampling with replacement*.

In this thesis, Fisher's Exact test is used in publication I to discover strengths of associations between gene keywords and text sets. Publication II presents the method that uses Fisher's Exact test to find associated database attributes for each created cluster.

Bayesian modelling

Bayesian modelling differs from classical approaches of statistical modelling mainly in three major points (listed for example in Nikkilä, 2005). First, whereas the classical probability equals to the number of successes from the total number of outcomes, in the Bayesian context, the probability should be rather interpreted as a subjective degree of belief. Secondly, the Bayesian modelling facilitates treating all unknowns, such as models, parameters, or missing data, as random variables.

Therefore all unknowns can be presented with probability distributions. This implicates that the approaches that can not be directly addressed with classical probability theory, such as model selection, are more straightforward in the Bayesian context. Third, Bayesian probability theory involves the concepts of *prior* and *posterior probability*. Prior probability represents the initial belief on modelled event whereas the posterior probability is computed after the new data is observed (Kass and Raftery, 1995). Use of prior is advantageous especially when lot of uncertainty relates with the modelled event, in a condition that convenient prior exists (Kass and Raftery, 1995).

Bayesian parameter estimation is used for combining both prior information, such as experience or knowledge, and current observation in an estimation process (Glymour et al., 1996). Each set of values for parameters or unknowns can be considered as distinct hypotheses involving different models. This is opposite to the null model hypothesis testing in classical statistics. The posterior probability for parameter of interest θ_1 given the data D is defined using Bayes rule as follows (Kass and Raftery, 1995 or Nikkilä, 2005):

$$\begin{aligned} P(\theta_1|D) &= \frac{P(D,\theta_1)}{P(D)} \\ &= \frac{P(D|\theta_1)P(\theta_1)}{P(D)} \\ &= \frac{\int P(D|\theta_1,\Theta)P(\theta_1,\Theta)d\Theta}{P(D)} \\ &= \frac{\int P(D|\theta_1,\Theta)P(\theta_1,\Theta_2)d\Theta}{\iint P(D|\theta_1,\Theta)P(\theta_1,\Theta)d\theta_1 d\theta_2} \\ &\propto \int P(D|\theta_1,\Theta)P(\theta_1,\Theta)d\Theta \end{aligned} \quad , \qquad (3.8)$$

where the numerator includes the joint probability of all variables. This is integrated over all unknowns that are not of interest (Kass and Raftery, 1995), defined as a set of parameters Θ in this example case. According to Bayes rule, the joint probability can be expressed as a multiplication of likelihood of parameters given the data and prior probability for parameters (Kass and Raftery, 1995). The denominator, i.e. the marginal probability of data, is equal to integrating the joint probability over all unknowns. This is also known as a normalizing constant (Nikkilä, 2005).

The key point of Bayesian estimation is taking into account the uncertainty related to the parameters with prior which is integrated out within the integrated likelihood term (Nikkilä, 2005). This is often a significant advantage in modelling very uncertain phenomena that are inaccessible for direct sampling. One good example of the use of Bayesian estimation is oil reservoir modelling in petroleum industry where expert knowledge is integrated to the analysis as a form of prior (Ewing, 1994).

Bayesian model selection extends the hypothesis space from classical parameter space to include all compared models under a single unified model (see for example Forster, 2000). Each of

the models can have their own parameter spaces. The posterior probability of model *M* is defined as (Kass and Raftery, 1995):

$$P(M|D) = \frac{P(M)P(D|M)}{P(D)} = \frac{P(M)\int P(D|M,\Theta)\cdot P(\Theta|M)d\Theta}{\sum_M P(M)P(D|M)}, \quad (3.9)$$

where the prior probability of the model is multiplied with the marginal likelihood of the model in the numerator. The denominator contains normalizing constant which is equal to the sum of multiplications of priors and likelihoods for all models.

A standard approach to compare two models M_1 and M_2, with the related parameter sets Θ_1 and Θ_2, is to use the posterior odds ratio $P(M_1|D)/P(M_2|D)$. If the priors of different models are assumed equal, the ratio reduces to comparison of marginal likelihoods. This is referred to as Bayes Factor (BF) (Kass and Raftery, 1995):

$$B_{12} = \frac{P(D|M_1)}{P(D|M_2)} = \frac{\int P(D|M_1,\Theta_1)\cdot P(\Theta_1|M_1)d\Theta_1}{\int P(D|M_2,\Theta_2)\cdot P(\Theta_2|M_2)d\Theta_2}. \quad (3.10)$$

BF is reminescent of the likelihood ratio (LR) hypothesis testing of classical statistics. Namely, BF can be considered as a Bayesian alternative for classical hypothesis testing (Kass and Raftery, 1995). However, rather than comparing the maximized likelihoods of models, BF takes into account the uncertainty related to data and parameters (as stated in parameter estimation above). As traditional hypothesis testing, also BF includes implementation of Occam's razor and is thus sufficient for model selection. BF does not require additional penalization such as penalization terms in ML based model selection AIC or BIC (Kass and Raftery, 1995). A broad description of Bayes Factors is given in Kass and Raftery (1995). We use BF for model selection in publication V.

The result of Bayesian modelling is often the posterior distribution which can be used to create a decision on the particular parameter value. A natural approach to summarize the posterior is to report a value where the posterior distribution reaches its maximum. This is called Maximum A Posteriori estimate (MAP). It is reminiscent of the MLE approach of classical statistics (Nikkilä, 2005). When the posterior distribution is highly peaked, the proportion of the MAP value contributes to the area of whole posterior distribution and thus the MAP gives reasonable results. Otherwise, the use of MAP may lead to poor results. We use MAP like solution in model selection to choose appropriate segmentation solution in publication V (see chapter 4.4 for details).

Related information theory based measures

Information theory is based on compact representation of information for storage and communication purposes. Thus, it provides methodology that is suitable especially with analysis of symbolic data (Grosse et al., 2002), such as biosequences. Information theory is also closely related to statistical modelling (Nikkilä, 2005), and often used in the same applications.

A basic measure of information theory is Shannon's Entropy (Shannon and Weaver, 1949), referred often as information entropy:

$$H(X) = -\sum_i p(x_i) \log_2 p(x_i). \qquad (3.11)$$

Entropy reports the amount of information in the analyzed signal or event, and can be considered as a measure of uncertainty for an event related with a particular probability distribution.

In this thesis (publication V), and in several other studies (see for example Li, 1990 and Grosse et al., 2002), the concepts of information theory are used for measuring similarity or dependence. One of such measures is Mutual Information (MI) (Shannon and Weaver, 1949), which measures the dependence between two variables, say *X* and *Y*:

$$MI(X;Y) = \sum_{x \in X} \sum_{y \in Y} P(X,Y) \log \frac{P(X,Y)}{P(X)P(Y)}. \qquad (3.12)$$

MI defines how much the uncertainty related to one variable is reduced when the other variable becomes known (Li, 1990). If *X* and *Y* are independent, MI is zero. When *X* and *Y* are strongly dependent, MI between them is very large. MI has a direct relation to correlation measures and can also be considered as an information theoretic alternative for correlation. Indeed, the evaluation of classical correlation and MI reports the latter more suitable for comparing symbolic sequences (Li, 1990).

Often, the distance or difference between two variables is of interest. In information theory, difference or variables *X* and *Y* can be measured using Kullback-Leibler divergence (Kullback and Leibler, 1951):

$$D_{KL}(X,Y) = E_X\left(\log(X/Y)\right) = \sum P(X=i) \log(P(X=i)/P(Y=i)), \quad (3.13)$$

which defines the distance from assumed "true" probability distribution *Y* to the target distribution *X*. It should be noted that although D_{KL} measures distance from Y to X, it does not satisfy conditions of a distance metric (for example the symmetry) and cannot be considered as such. Jensen-Shannon divergence D_{JS} is often regarded as a symmetric version of D_{KL} (Borovkov, 1984; Grosse et al., 2002):

$$D_{JS}(X \parallel Y) = D_{KL}(X \parallel (X+Y)/2) + D_{KL}(Y \parallel (X+Y)/2). \qquad (3.14)$$

D_{JS} defines the symmetric difference between probability distributions *X* and *Y* (Grosse et al., 2002). Symmetry, mathematical interpretability through D_{KL}, and the possibility to compare also more than two distributions have made D_{JS} popular in several applications involving symbolic data (Grosse et al., 2002). D_{JS} and its modifications are used in publication V for measuring difference between models generating artificial data and models that represent segmentation solutions for the data.

4 Novel methods for mining genome wide data

This chapter summarizes the data mining techniques developed and applied in publications I-V and discusses some points that were not discussed in the publications. It should be noted that the chapter focuses fully on analysis, which involves integration of auxiliary database data with gene sets such as co- or differentially expressed genes.

Chapter 4.1 describes the methodology developed in publication I for detecting novel and expected results obtained from DNA-microarray experiments. As this method is very shortly discussed in publication I, the chapter includes some additional points and discussion. Chapter 4.2 summarizes publications II and III which describe a novel clustering method for functional interpretation of microarray data and its application. Some additional points are also discussed which are not presented in publications II and III. Chapter 4.3 summarizes shortly publication IV describing the POXO software tool for detecting transcription factor binding sites from co-expressed genes. Chapter 4.4 summarizes the segmentation method presented in publication V. Some additional illustrations of results are also given.

4.1 Simple text mining for detecting expected gene expression

In this chapter, a simple text mining method is presented for interpreting data obtained from DNA-microarrays. The method discovers associations between gene groups and sample treatments by searching related keywords from scientific text sets. Such associations are used in comparison with results from fish cDNA-microarray screening for environmental toxins, to observe the novel and expected findings.

The biological background for the presented method is described in chapter 2.3 (problem i). In short, we conducted cDNA-microarray experiments for salmon fish treated with different environmental toxins: β-naphthoflavone (BNF), cadmium (Cd), carbon tetrachloride (CT), and pyrene (Pyr). As a result, we obtained genes that are expressed abnormally under treatment of each toxin. In addition to microarray results, we wanted to get an overview of similar findings from previous scientific studies. This was made by performing literature screening to discover associations between the genes in the microarray, grouped according to associated GO-terms, and the four tested toxins.

4.1.1 Overview of the developed text mining method

Data representation

For data, we used the literature abstracts downloaded from MEDLINE database of scientific literature. First, all genes in the microarray were divided into groups according to functional categories obtained from the Gene Ontology (GO) database. For further use, the gene expression levels within each gene group were averaged to get a representative level for each group under each toxin. For each such gene group, the associated GO-term was used as a keyword. It should be noted that these groups could overlap in genes.

For data, four sets of abstracts from scientific articles associated with each of the four toxins (BNF, Cd, CT, and Pyr) were used. Borland Delphi 7.0 was used to develop a program that

executes automated searches to the Pubmed databank using its E-Utilities application programming interface (API). The names of the toxins were used as search words. For the four toxins, totally 14,334 abstracts that existed in Pubmed were downloaded.

The rationale was to analyze the strength of associations between gene groups and toxins by studying the frequency of each gene group keyword (GO-term) in each of the toxin abstract sets. Such frequencies of associations can be indicated as contingency tables presented in table 1.

Table 1. Contingency table describing the frequency of associations between a particular gene group with a particular toxin in the literature abstract sets. Number of abstracts including the gene group's keyword (GO-term) was interpreted as the frequency of associations.

	Gene group related	Other	Totals
Toxin related	a	b	a + b
Other	c	d	c + d
Totals	a + c	b + d	a + b + c+ d

Measuring strengths of associations

The presented approach contains an application of statistical measures to analyze frequency of word and gene group association pairs. Similar measures have often been used for analyzing over-representation of database attributes within gene sets (see for example Törönen, 2004 and publications I, II, III and IV). The method was developed as a part of publication I and applied with some modifications also in Krasnov et al. (2005).

For defining the strength of association between each gene group and toxin pair, we apply the measures from a hypergeometric distribution (see chapter 3.3) to the contingency table presented in table 1. The hypergeometric distribution as a null model, in this context, represents sampling of abstracts from the whole abstract population set, when not returning each sampled abstract back to the population. This should be an appropriate model for the situation in question.

The aim of the presented text mining method was to assign scores for the strengths of associations in order to sort them and find the strongest. The plain hypergeometric probability would not be very informative for this. That is, as the sample size increases, the distribution is spread among a larger amount of observations. This decreases systematically the probability of each observation and makes the probabilities for different contingency tables incomparable (Törönen, 2004).

We tested two different methods to define the strengths of associations. First, we calculated Z-scores from the hypergeometric distribution for each observed frequency. The Z-score indicates how far the observation deviates from its distribution mean when expressed in the units of the standard deviation of the distribution (Cheadle et al., 2003). This normalizes differently deviated probability distributions allowing better comparison of different contingency tables. Z-scores are also computationally very simple to calculate.

In addition, we also performed right sided Fisher's Exact test for comparison which reports the probability to obtain "*a* or more" associated abstracts in a random sample. Although the Fisher's

Exact test is a computationally harder approach, due to its exactness it should give more valid results than Z-statistics or associated p-value. Moreover, several studies on statistical tests highlight that it is recommendable to use Fisher's Exact test especially when the analyzed contingency table contains small or unbalanced frequencies (Berry and Mielke, 1988; Bradley et al., 1979; Kraemer and Woolson, 1987; Lawal and Upton, 1990; Tate and Hyer, 1973). In our case, we expected unbalanced data as the frequencies of keywords were expected to be relatively small among the all abstracts. Still, we found that the sorted list of enriched gene groups for each toxin (described by a particular GO-term) was largely similar when sorted by either the Fisher's Exact test p-value, or Z-score.

Methods presented in this chapter and other parts of thesis use largely the Fisher's exact test for 2X2 contingency tables. It should be noted that with larger contingency tables, such as tables describing data frequencies in all partitions of a particular segmentation or clustering solution, it becomes computationally impractical and approximations or other methods should be used.

4.1.2 Biological results

The presented text mining of MEDLINE abstracts provided valuable information on the associations of gene groups and the toxins. The method was assumed to produce rather true negatives as exact keywords were searched rather than enumerating over all of their regular expressions. Thus, we did not expect a highly strong correlation between text mining and microarray results. Moreover there was not enough data for fish and therefore abstracts for all species were obtained. In addition, we did not use any correction for p-values although repetitive testing was performed.

The method could still predict preferential activation of heat-shock proteins, RAS pathway of signal transduction, protein biosynthesis, and proteasome protein degradation during exposure to BNF (see publication I). Many other themes were also expected such as induction of microsomal proteins by BNF and CT. See detailed interpretation of results in publication I.

Still, a more sophisticated approach would be beneficial in order to associate genes or gene groups with biological conditions such as chemical treatments or clinical parameters. The approach we used may only recognize only portions of real occurrences. That is, we only search exact occurrences of keywords or terms whereas various expressions of the same concept are often present in natural language.

4.2 Theme discovery from gene sets

This section presents a novel exploratory approach for detection of representative biological themes from a set of genes, such as co- or differentially expressed genes. The method creates a non-nested scheme from several partitional clustering solutions performed for the gene set. Clustering uses binary data indicating the associations between genes and GO-terms. Clustering is performed using a NMF based method. The developed method and associated GENERATOR software were originally presented in publication II, applied in the analysis of differentially expressed genes from transgenic nematodes in publication III, and integrated as a part of TFBS discovery software in publication IV (see chapter 4.3).

The biological background and rationale of the developed method was given in chapter 2.3. In short, the aim was to reveal gene groups that would correspond to separate biological mechanisms, such as separate biological processes, pathways, or larger parts of the pathways.

4.2.1 Overview of the developed clustering scheme

Data representation

Input for the developed method includes two gene sets: the sample genes, e.g. co-expressed genes, and the background set of genes, e.g. the other genes in the genome or microarray. As data, we use annotations obtained from the Gene Ontology (GO) database (Ashburner, 2000). The associations between genes and GO-terms are represented as binary data in the sample gene set. This data matrix is used for clustering described in the section. For further information on GO, the reader is recommended to consult Ashburner (2000).

Non-negative matrix factorization based clustering

Due to our observations, the binary data representing the associations of genes and GO-classes was often very high dimensional. The data also seemed to contain small proportions of non-zero entries (see publication II). We also observed manually, that the genes tended to have similarities in subsets of dimensions. Often there were genes that could have been grouped according to few (such as 5-15) GO-terms related to a particular biological pathway whereas the rest of the attributes (such as 100-200) did not correlate. This is natural as genes often have several alternative functionalities, as mentioned in chapter 2.1.

Many methods that have been traditionally used for data clustering weight equally all dimensions (Fränti et al., 2003; Cai et al., 2004). With such methods, the patterns located in the subset of dimensions tend to be overwhelmed with the dimensions that are irrelevant for the cluster structure of the whole data or individual cluster, as discussed in chapter 3.2.3. Because of these reasons, we chose to use a clustering procedure that is based on NMF (Paatero and Tapper, 1994; Lee and Seung, 1999). NMF has shown good performance with sparse binary data and with a large number of dimensions in the topic discovery literature and image analysis (Seppänen et al., 2003; Lee and Seung, 1999). According to Rajapakse and Wyse (2003), NMF can also learn parts-based representation that is an intermediate between the local and global features (note that this is not the same as global and local optima). This motivated us to use NMF rather than PCA or ICA. We use the NMF algorithm presented in (Lee and Seung, 2001) which is based on the update rules under which the Euclidian distance is non-increasing.

Clustering is performed by depositing each gene into a cluster that corresponds to the NMF produced feature vector with the highest value (see publication II for details). We use the relation between the highest value and the sum of all values as a measure of fitness for each gene in its cluster. In the visualization, the fitness is used to present genes in a sorted order for each cluster.

As data points may exist that do not fit well into any cluster, it would be possible to use some threshold in the cluster membership, for example see Kim and Tidor (2003). This is an important aspect but not taken into account in our approach. In addition, there may be data points that get high loading on multiple feature vectors, indicating alternative clustering solutions. This problem is considered by repeating the clustering several times for each number of clusters and observing the conservation of clusters.

Non-nested clustering scheme

Partitioning of genes into *r* subgroups is not itself a very informative result as the parameter *r* given by the user can take all values between *1* and data size *N*. Model selection, discussed in chapter 3.3, has been often used in evaluating different clustering solutions in order to find the most suitable number of clusters. In our approach, we tested stochastic complexity (Rissanen, 1987; Rissanen, 1996) based on Minimum Description Length (MDL) principle. We used the simple model of stochastic complexity for binary classification problems given by Gyllenberg et al. (1994, 1997, 2000) for evaluating clustering solutions with different *r* (results not shown). The method worked well with artificial data sets that we used to test the method. With real data, the MDL method evaluated the result with one cluster the best in most of the cases. Still, the manual interpretation of obtained clusters revealed biologically very reasonable groupings. We concluded that this was caused by the nature of the biological data, i.e. the tendency to include similarities in the subsets of all attributes.

One way to proceed would be to include feature selection within the model selection procedure to find relevant subspaces for cluster structure. In our approach, we have used a different kind of solution, and rather search individual clusters than complete clustering solutions. In publication II, we have proposed a non-nested scheme for combining information from several clustering solutions. The scheme combines clustering solutions each of which has a different number of clusters *r*, where *r* grows gradually from two into a user given number. Each clustering is executed from a random starting initialization using NMF, producing an independent division level to the scheme.

Correlations between each cluster *A* in level *r* and each cluster *B* of previous level *r - 1* are calculated by comparing cluster memberships of genes as binomial distributions with correlation. In the system described in publication III, we also calculate p-values of correlations. The strongest correlation for each cluster is denoted by a line between the corresponding clusters. The resulting visualizations can be seen in publication II (figures 1, 2 and 3) and III (figure 2).

The core model of the developed scheme has mainly two advantages. First, it utilizes the non-deterministic nature of NMF: if there exists a cluster that stays similar in different clustering solutions despite the changing initializations, then such a cluster can be interpreted to represent a non-random outcome. We utilize this feature by repeating clustering with the same and different numbers of *r*, and observe the correlations. Secondly, the scheme reveals hierarchical relationships of data: it is possible to see the division of larger parts of data with higher similarity into smaller partitions of data points that share lower similarity. This is apparent in the analyzed data sets, see for example figure 2 in publication III, where catalytic activity related genes are divided into subgroups all under catalytic activity topic. The division is not forced by any constraint, like 1-to-2 division or merging in hierarchical clustering. Rather it implies true existence of smaller parts that compose the larger entity.

Describing cluster contents with associated attributes

The main purpose of the developed method was to give a quick and simple overview of the over-represented biological themes in the analyzed gene list. Thus, rather than reporting gene groups, the

rationale was to discover and show the associated theme for each cluster generated from the most relevant attributes.

As mentioned previously in this chapter, the associations between genes and GO-terms create high dimensional binary data that tends to contain clusters only in the small subsets of all data attributes. It would be possible to use another decomposition matrix *H* to sort the attributes according to their relevance (weight) for each factor of *W* (representing a cluster) to obtain these subspaces. This has been performed for gene expression analysis study in Carmona-Saez et al. (2006). Such relevance would consider attributes with stronger association to the analyzed cluster than other attributes. Naturally, there could be multiple gene clusters that have association with the same subset of attributes.

In our analysis of cluster contents, we have used an approach that, to our knowledge, is novel. We consider as relevant the attributes that characterize only the analyzed cluster and not the other clustered genes. In other words, we obtain GO-classes which are unexpectedly frequent in the analyzed cluster, when the expectation is obtained from the population of all clustered genes. By this, we obtain a desired effect in our clustering scheme when a cluster with substructure of attributes is split into two "sub clusters". In the resulting sub clusters, the relevance of attributes enriched in the original cluster decreases, and the relevance of the attributes that define the structure for the new clusters increases. Due to this, the attributes reported for larger clusters give a coarse level description of the analyzed list, whereas the attributes reported for smaller clusters tend to give a more detailed description.

The developed method is targeted for the analysis of gene sets that fulfil some criterion such as co- or differential expression. Therefore, we are also interested in the attributes that show enrichment in such a set when comparing to the genes that have not fulfilled the criterion. This has been the standard way in the analysis of gene sets before with the so-called sorted list approach (Hosack et al., 2003; Boyle et al., 2004). The purpose of this analysis is to discover attributes that are so much more frequent in the sample gene list that they can be considered as biologically meaningful. Based on the presented points, our final description of cluster contents is obtained by filtering attributes by O.log(p) and S.log(p) measures, and sorting them by using C.log(p), the total enrichment in each cluster. See publication II for a description of these measures.

4.2.2 Biological results and evaluation

Analysis of yeast genes sensitive for H_2O_2 stress.

The developed method was applied in publication II to analysis of non-essential yeast genes that were found sensitive to oxidative stress in (Thorpe et al., 2004). The analysis of these genes in the original study reveals their association to mitochondrion. In addition, we analyzed a differentially expressed gene set obtained from microarray screening of itraconanzole, an antifungal drug (Hughes et al., 2000). This is not discussed here. For details, see publication II.

Results of the analysis steps (see publication II, figure 1) show that the oxidative stress sensitive genes are divided into coherent sub-groups each representing a separate biological functionality in the clustered data. Themes corresponding to each resulting cluster are: I, organellar and mitochondrial ribosome; II, tRNA ligase; III, mitochondrial membrane; IV, transcription regulation; and V, mitochondrial genome maintenance. A detailed discussion of result has been given in publication II. Briefly, the apparent existence of mitochondrial and organellar ribosome

themes (cluster I) seems reasonable, as these are known targets of oxidative stress (Thorpe et al., 2004). Cluster II shows tRNA ligase theme that was not reported in the original analysis of this data.

Analysis of up-regulated nematode genes over-expressing a human Parkinson's Disease related transgene

Inclusion of α-synuclein gene is a known hallmark of several neurodegenerative diseases in humans (such as Parkinson's Disease). In publication III, we performed a genome wide expression screening in *C. elegans* nematode over-expressing both wild-type and A53T human α-synuclein transgenes. In comparison between α-synuclein and wildtype strains, 433 genes were up- and 67 genes down-regulated by fold change > or <2 and significance $P < 0.05$ (see publication III for details). We performed analysis with GENERATOR presented in publication II for these gene lists. The clustering result for up-regulated genes is shown in publication III, in figure 2.

The whole gene list seems to enrich development and embryo development related attributes. This may be expedient or may suggest a bias in the timing in the worm growth process. Note that in the latter case, it would be especially useful to perform the proposed clustering analysis in order to exclude the genes that are implicated from such bias as a one cluster. Apparent themes in the lower clustering levels are catalytic activity/proteasome, mitochondrion, development, and reproduction related clusters. The first two themes were previously suggested to be related to the α-synuclein pathway before (McNaught et al., 2002, 2004; Zhou et al., 2004).

Comparison against methods used in the same application

Among the many tools for finding enriched biological topics, using often the GO-terms as data, two basic methods exist. The first method simply calculates statistical significance of over-representation and reports the sorted list of attributes for the user (see for example Hosack et al., 2003). The other method also calculates the significance but visualizes the significant GO-terms in the graph that is obtained from the relations of GO-classes indicated in the database (see for example Boyle et al., 2004). In publication II, we give a comparison of our method and these methods. Our comparison shows the developed method gives a much simpler summary which can still report the most important biological themes.

During the review of publication II, a tool making a similar type of analysis was published. GOProxy, contained in GOToolBox (Martin et al., 2004), creates a hierarchical clustering of gene set using Czekanowski-dice distance. We also performed a comparison of GOProxy against our method. The results from GOProxy with default settings included a large amount of clusters, several of which were minimal or near minimal sized. We proposed (publication II) that this could be caused by two aspects. First, the obtained clusters are allowed to overlap as they are selected in the different levels of hierarchical tree. Secondly, the use of a standard clustering approach can have a tendency to obtain very small clusters with the used data as it rather detects groups with highly similar genes.

4.3 Tool for finding transcription factor binding sites

It is known that gene regulation is often mediated by specific proteins called transcription factors (Wray et al., 2003). These proteins work by binding onto transcription factor binding sites (more generally *cis*-elements) that are small DNA regions in the promoter region of a gene. Such regulation often concerns several genes located apart rather than a single gene, as these genes can be regulated by the same transcription factors. A central question of molecular biology is to find the transcription factors and their binding sites.

This chapter discusses the POXO tool series which is intended to discover transcription factor binding sites from a set of co-expressed genes. The tool is reported in publication IV. POXO can perform functional gene set evaluation and grouping, sequence retrieval, pattern discovery and pattern verification. Some of these modules have been published previously elsewhere such as GENERATOR (publication II), POCO (Kankainen and Holm, 2005), and POBO (Kankainen and Holm, 2004). POXO has been implemented as Web software using Perl, C and C++. Biological results obtained by using the POXO tool, presented in publication IV, are not in the scope of this thesis.

The principal objective of POXO is to analyze the frequencies of different sequence patterns in the promoter DNA-sequences of user given gene sets, in order to reveal the putative binding sites. Figure 1 in publication IV shows the steps of the POXO tool with a typical workflow. The most important parts are briefly described below.

Finding functional sub-groups

Input for POXO is a set of genes, e.g. co-expressed genes. The first step of the pipeline is to examine what biological functions the genes are involved in (GENERATOR module). In this examination, the functional descriptions from GO-database (Ashburner et al., 2000) are associated with the genes in order to discover the over-represented themes. The functional annotations can also be used to group genes into functionally analogous subsets, if the experiment perturbed several biological processes. The grouping is performed using a new implementation of the GENERATOR clustering method presented in chapter 4.2 and publication II.

Some slight modifications to the original GENERATOR are introduced in the POXO implementation. The first modification is the calculation of statistical significances of correlations between clusters of adjacent clustering solutions, instead of plain correlations. Only significant correlations are visualized for the user rather that the best correlation for each cluster as in the original tool. Another modification is that NMF with r factors is repeated multiple times from random initializations and the solution with the smallest squared sum difference is chosen as a representative for the rth level in the visualization.

Grouping of genes allows POXO to be used not only to study the regulation of genes, but also the regulation of a specific function. This should also improve the discovery of regulatory elements as focus on a more specific sub-set of functionally similar genes is facilitated (publication IV).

Discovering representative patterns

The next step is to discover the putative *cis*-elements from the set of promoter sequences of genes. This is made by searching sequence motifs that are significantly over-represented, and thus suggest a regulatory function. POXO facilitates the search of over-represented sequence patterns in a sequence set versus the background set such as a genome, using the POCO module. Additionally, patterns that maximize discrimination between two sequence sets can be searched. This enables discovery of patterns that are over-represented in one and under-presented in the other sequence set. Additionally, the POCO 2nd iteration tool enables the user to search pattern combinations of patterns that were discovered by POCO.

Analysis of representative patterns is implemented by using a bootstrap approach to create pattern occurrence distributions for the inputted set of genes, and background distributions for the promoters in the genome of the analyzed organism. These are used as a basis for testing significance of each pattern found in the input set of promoters.

Clustering of patterns

As patterns reported by the pattern discovery tool can overlap, it is desirable to obtain a non-redundant set of these patterns. Such a set can be produced, for example, by using clustering. In the POXO tool, the found patterns are clustered using phi coefficient of correlation as a distance measure (used also in Burset and Guigo, 1996) and agglomerative hierarchical clustering. As a result of clustering, the consensus patterns are produced that can be used as input for the next steps.

Pattern evaluation, visualization and phylogenetic footprinting

POXO facilitates testing of a preliminary hypothesis about the regulatory factor and binding site. This is implemented by the POBO module which can be used to screen and evaluate a predetermined pattern from either one or two sequence sets (Kankainen and Holm, 2004). The pattern distributions are generated similarly to POCO using a bootstrap approach.

POXO has also tools for visualization of the locations of the discovered patterns. This can be used to verify patterns and to examine their locations in the sequences. The visualization can efficiently reveal details such as similar distance of patterns from the transcription start site and co-occurrence of separate patterns.

POXO can also perform verification of discovered patterns against databases of known *cis*-elements. Patterns can also be verified using phylogenetic footprinting, in which the conservation of the patterns is examined in homologous sequences across different model organisms.

4.4 Heuristic Bayesian genome segmentation

This chapter presents development of a heuristic Bayesian segmentation method which was applied to find the chromosomal locations of co-expressed genes. The work has been originally described in publication V. We consider the chromosome as sequential data where each data point represents a single gene. In order to find the regions with homogeneous gene expression from this data, we have developed a segmentation method using Top-Down heuristics and Bayesian model selection. For a description of the general segmentation problem, including the symbols used below, see chapter 3.2.2 above.

As mentioned in chapter 2.1, the location specific expression can be caused by different biological factors, such as impact of regulatory proteins in the opened chromatin region. The dynamic changes in chromosome structure that permit or prevent gene expression are referred to as *chromatin remodelling*. Our particular interest has been to find chromosome regions in baker's yeast (*Saccharomyces cerevisiae*) which include genes with similar gene expression levels during the cell cycle. Such analysis could reveal regions which become active and passive during the chromatin remodelling in the cell cycle.

4.4.1 Overview of the developed methodology

Data representation

In our application, the N genes of a chromosome are represented as a sequence of data $D = (d_1, d_2, ...d_N)$. Each data point d_i corresponds to the ith gene in the chromosome, when ordered according to their physical locations. As data we have used groupings of genes according to their gene expression levels in different time points during the cell cycle. These groupings were obtained by clustering genes using preprocessed gene expression data from (Spellman, 1998). This preprocessing included background correction, logarithmic ratio transformation, and removing bad quality spots, as reported in Spellman et al., (1988). As a quality measure, a pixel-to-pixel correlation between Cy3 and Cy5 intensities had been used (Spellman et al., 1998). It is clear that this preprocessing is not sufficient as such. Still we consider the dataset suitable for our purposes, as 1) our style to represent the data as gene expression clusters reduces the noise and 2) the segmentation method addresses rather large groups of genes, decreasing the resonance of random failures in data.

Preprocessed data was normalized using per-gene and per-chip median-centering (see chapter 2.2.2 for an explanation). As a clustering method, k-means clustering with Euclidian distance was used (see an explanation of clustering in chapter 3.2.1). See also chapter 2.2.2 for an explanation of the preprocessing methods.

Rather than one clustering result, we use several clustering solutions with different number of clusters. Each clustering solution creates one dimension in the data. Therefore each data point (a gene) $d_i = (d_{i1}, d_{i2}, ...d_{iM})$ is indicated as an M-dimensional vector, where each multinomial value $d_{ij} \in \{1, ..., K\}$ indicates the cluster number of the ith gene in the jth clustering solution with K clusters. In our application we created clustering solutions for 1/3 of genes which had the highest expression variance with *3, 4, 5* and *6* clusters creating data with *M=4* dimensions. The remaining genes were included as an additional group of non changing genes in the data creating *K=4, K=5, K=6* and *K=7* classes for each dimension, respectively. This way we could map the co-expression of genes from general to specific.

Algorithm

Our final aim has been to implement the presented method as Web server based software. Due to scalability, this requires using an algorithm with an adequate running time. Therefore we use a heuristic Top-Down method which performs series of 1-to-2 segmentations, reminiscent of the divisive hierarchical clustering. Similar algorithms have been used before in image analysis (Douglas and Peucker, 1975) and in bioinformatics (Bernaola-Galvan, 2000; Grosse et al., 2002; Li

et al., 2002). The time complexity of the method with k segments is $O(kN)$. As a result we can obtain a single segmentation solution for each number of change-points very rapidly.

For selecting each next change-point for 1-to-2 segmentation in our algorithm, we use a Maximum Likelihood Ratio (MLR) based approach as a score. This is despite we use Bayes Factor (BF) for evaluating the obtained segmentation solutions. The reason for using MLR here is that we found that BF with the Top-Down algorithm tended to lead solutions with many small segments when leaving the larger segments intact. In our simulations, we concluded that this could be related to the uncertainty in peripheral region in 1-to-2 segmentation. In our simulation with single segment random data, BF gives negative support, i.e. no support for segmentation. Still, the negative support of BF is slightly higher in peripheral regions. This is discussed more in publication V. The result of comparison between MLR and BF in 1-to-2 segmentation of random segment is visualized in the publication V in figure 1.

For each created sub-sequence, we search the location that has maximum increase for the MLR. The location with the largest increase among all existing sub-sequences is then chosen for the subsequent 1-to-2 segmentation by the algorithm.

It has been shown, for example in (Grosse et al., 2002), that the logarithm of this MLR is equal to Jensen-Shannon divergence D_{JS}, multiplied with the data size i.e. $\log(\hat{L}_k / \hat{L}_0) = N \times D_{JS}$. Therefore, our scoring approach is similar than with segmentation methods presented in Grosse et al. (2002) and Li et al. (2002).

Selecting a good number of segments

We use a Bayesian approach for evaluating the created segmentation solutions for each number of k segments. The Bayesian approach for segmentation, presented in Liu and Lawrence (1999), has been implemented using a dynamic programming based segmentation algorithm presented originally in (Bellman, 1961b). It enumerates over all possible change-point solutions which facilitates obtaining the posterior for any parameter of interest. Thus, it is referred often to as a full Bayesian approach. In our approach, we do not use a fully Bayesian method, as we evaluate only one solution for each number of change points k, provided by a heuristic algorithm.

When any model is not favoured *a priori*, maximizing the posterior odds ratio of models given the data is the same as maximizing the BF. In our approach, we do not favour any segmentation model *a priori* differently from (Liu and Lawrence, 1995) who use probability 0.5 for the null model (data without segmentation) and divide the rest by 0.5 among the remaining *k_max* models. Also, we use the model without change-points as a null model which is equal in any BF comparison for a particular data. Therefore, we seek for a segmentation model that maximizes the marginal likelihood:

$$\begin{aligned} P(D|M_k) &= \int\int P(D, \Theta, \Psi \mid M_k) d\Theta d\Psi \\ &= \int\int P(D \mid \Theta, \Psi, M_k) \cdot P(\Theta) \cdot P(\Psi \mid M_k) d\Theta d\Psi \\ &= \int\int P(D \mid \Theta, M_k) \cdot P(\Theta) \cdot P(\Psi \mid M_k) d\Theta d\Psi \quad , \\ &= \int\int P(D, \Theta \mid M_k) \cdot P(\Psi \mid M_k) d\Theta d\Psi \end{aligned} \tag{4.1}$$

where M_k is the segmentation model with *k* segments (*m=k-1* change-points), and the parameters $\Theta = \{\theta_{cvi} \mid c = 1..k, v = 1..V, i = 1..I_v\}$ describing the probabilities of data classes in the segments. Here *v* and *V* correspond to the different dimensions and the number of them. Symbols *i* and I_v correspond to the data classes and the number of them in each dimension. We assume that the parameters representing the locations of change-points in a segmentation solution Ψ do not have an impact on the likelihood and thus $P(D \mid \Psi, M_k, \Theta) = P(D \mid M_k, \Theta)$.

As indicated in the last line of equation 4.1, we need to define the joint probability of data *D,* and parameters Θ describing the probabilities of data classes in the segments. Thus we need to define the likelihood of data given the parameters and the prior distribution for the parameters. We assume two simplifications that ease up the calculations. First, we consider the segments independent of each other. This has mainly two advantages: 1) the joint probability of several segments can be calculated via multiplications and 2) by optimizing a model for a sub-sequence, the complete model is also optimized with an equal amount. Secondly, we assume that the different data dimensions are independent. Therefore, the joint probability of all dimensions equals to multiplication of probabilities of the dimensions.

As the likelihood and prior are selected so that both functions are from the same conjugate family, the joint probability becomes relatively simple. With multinomial likelihood and Dirichlet prior the joint probability takes the following form after integration:

$$\int P(x,\Theta)d\Theta = \int P(x \mid \Theta) \cdot P(\Theta) \cdot d\Theta = \prod_{v=1}^{V} \frac{\Gamma\left(\sum_{i=1}^{I_v} \alpha_{vi}\right)}{\Gamma\left(\sum_{i=1}^{I_v} x_{vi} + \sum_{i=1}^{I_v} \alpha_{vi}\right)} \prod_{i=1}^{I_v} \frac{\Gamma(x_{vi} + \alpha_{vi})}{\Gamma(\alpha_{vi})} . \qquad (4.2)$$

A similar form is known in literature and has been often used in applications with similar type of data analysis problems (Marttinen, 2006; Corander, 2004; Buntine, 2002).

An essential part of the Bayesian approach is to define the used prior for the parameters. As presented above, we chose the commonly used Dirichlet prior for describing prior belief on parameters Θ in the created segments. Still the Dirichlet prior weights α describing the pseudocounts of data classes must be defined.

The prior used with this type of unsupervised machine learning method should be objective, so that information from the user would not be required. The definition of such a prior contains two aspects. First, the proportion of each pseudo-count from their total sum must be defined. This represents our expectation on the balance of different class probabilities. Secondly, the prior sum itself i.e. the total count of all prior observations has to be set. This corresponds to the magnitude the prior knowledge is allowed to have impact on the analysis.

Table 2. Different Dirichlet prior weights. Symbols: I_v indicates number of classes in the dimension v, n_c marks size of segment, and the notation $P(D \in i)$ indicates the proportion of class i in the whole data.

Prior name	Sum per dimension v	Weight per class i	References
FLAT1	$\sum_i \alpha_i = 1$	$\alpha_i = 1/I_v$	Perks, 1947
FLAT	$\sum_i \alpha_i = I_v$	$\alpha_i = 1$	Liu and Lawrence; Ramensky et al. 2000
CSP1	$\sum_i \alpha_i = 1$	$\alpha_i = P(D \in i)$	-
CSP	$\sum_i \alpha_i = I_v$	$\alpha_i = I_v \times P(D \in i)$	Buntine, 2002
EBP	$\sum_i \alpha_i = \sqrt{n_c}$	$\alpha_i = \sqrt{n_c} \times P(D \in i)$	Lawrence et al., 1993; Carlin and Louis, 2000
MEBP	$\sum_i \alpha_i = \sum_i \sqrt{n_c \times P(D \in i)}$	$\alpha_i = \sqrt{n_c \times P(D \in i)}$	Publication V

Commonly used priors can be divided mainly to A) data independent priors which do not contain any information of the frequencies of data classes in the data set, and B) data dependent priors which are scaled according to proportion of classes in the data. Table 2 shows priors of type A (FLAT and FLAT1), and priors of type B (CSP1, CSP, EBP) with different prior sums. As explained in publication V, we observed an unwanted behaviour of very commonly used CSP and EBP priors with our data sets. The problem appeared with small segments and small probabilities of data classes. We noticed that the problem was related with the very small prior values in gamma function, in the denominator of the latter term of equation 4.2. Thus, we modified the EBP by taking the square root also from class proportion in order to increase the smallest prior weights. This is referred to as Modified Empirical Bayes Prior (MEBP):

$$MEBP = \sqrt{n_c \times P(D \in i)}\,. \quad (4.3)$$

The original aim was to create an *ad-hoc* modification which would mute the described behaviour of EBP. Still, we noticed the analogy with the Chi-square test. This aspect is omitted here and the details can be found from publication V.

For the prior representing individual segmentation solution with k segments, we use a very simple proper prior which is equal for all possible segmentation solutions with k segments (m change-points). The prior is defined by dividing the probability 1 with the number of all possible segmentation solutions that exist for m change-points in data sized N:

$$P(\psi) = \frac{1}{\binom{N-1}{m}}. \quad (4.4)$$

This is almost similar to the prior used in (Liu and Lawrence, 1995) but instead of *N - 1* they use *N*. This should not be exactly correct as there are *N - 1* positions for change-points in *N* sized data.

Finally, our criterion for evaluating the segmentation solutions takes the following form:

$$P(D|M) = P(\Psi)\int P(D \mid M,\Theta)P(\Theta)d\Theta$$

$$\propto \frac{1}{\binom{N-1}{m}} \times \prod_{c=1}^{C}\prod_{v=1}^{V} \frac{\Gamma\left(\sum_{i=1}^{I_v}\alpha_{cvi}\right)}{\Gamma\left(\sum_{i=1}^{I_v}x_{cvi}+\sum_{i=1}^{I_v}\alpha_{cvi}\right)} \prod_{i=1}^{I_v}\frac{\Gamma(x_{cvi}+\alpha_{cvi})}{\Gamma(\alpha_{cvi})}, \quad (4.5)$$

where as prior weights α we compare performance of CSP, CSP1, EBP, FLAT, FLAT1 to the developed MEBP prior. Further we take the logarithm of the equation 4.5. This eases up many calculations, for example by changing very low or high numbers to reasonable scale.

Benchmark system for method evaluation

In publication V, we also developed a benchmarking system for comparing the performance of our method against competing model selection methods, used with segmentation. The competing methods included similar Bayesian methods, with 5 different Dirichlet priors: CSP, CSP1, FLAT, FLAT1 and EBP, and the three ML based measures AIC, BIC and the modified BIC, referred to as BIC2 (Zhang, 2005). The comparison was based on generation of artificial data from a predefined model. In comparison, we emphasize the goodness of each method in prediction of future data. We have used three different types of artificial data: i) data with several large segments; ii) data with few large segments; and iii) data with several small segments. Each type of data was tested with a different number of data classes: 2, 10 and 30, all with three data dimensions. We create 100 simulated data sets for each type of data. The generation of data is explained further in publication V.

As mentioned in chapter 3.2.4, the standard way for evaluation of clustering created for artificial data is comparison of gene memberships in different clusters. An example of such methods is Mutual Information (MI), presented in equation 3.12. The shortcoming of these methods is that they take into account only the gene memberships and not the data itself. Therefore, the uncertainty related to probabilities of data classes within segments is omitted. Such an aspect is taken into account when the model generating the data and the obtained segmentation solution are considered as statistical models. In this thesis these two models are referred to as Data Generating Model (DGM) and Data Explaining Model (DEM), respectively.

In the evaluation of each method, we measured the closeness of each produced DEM with the DGM. We ended up using a symmetric measure called Jensen-Shannon Divergence, referred here to as D_{JS}. Prior information must be used as pseudo-counts with this measure, as Bayesian methods aim to optimize the biased model (see publication V). The difficulty here is how to choose the prior for evaluation. Thus, we use a consensus of several D_{JS} measures for evaluation, each with different type of prior, and also D_{JS} without a prior (see also details from publication V). We do not use the developed MEBP prior as a part of this consensus, as it could favour our method.

We create the consensus of different D_{JS} scores with different priors as follows. First, we calculate a subtraction between each D_{JS} score for our model selection method (Bayes with MEBP prior) and a competing method for each replicated data creation and its segmentation. Secondly, we calculate a Z-score for each such difference for each type of data (100 repetitions for each). As a result, we have a set of Z-scores, each obtained with different D_{JS} priors, for the performance of a particular method with particular type of data. This set is referred to as CONS-JSD score. We use the CONS-JSD score as a consensus measure for method performance by reporting its average (table 1 in publication V) and percentiles (figure 6). Positive score marks better performance of our Bayes method with MEBP prior, whereas negative score indicates that the competing method performs better.

4.4.2 Results and evaluation

Method comparison results

The result of method comparison with CONS-JSD measure is shown in publication V in table 2. The table values are averages for CONS-JSD sets. Figure 6 shows a graphical visualization of percentiles of CONS-JSD sets. The results are discussed in detail in publication V. Briefly, it is apparent that the performance of ML-based methods, when compared to Bayesian method, is dependent on the number of data classes. With any type of data with 2 classes, AIC seems to shows bad performance when compared to the developed method. The detailed analysis (data available not shown) shows that AIC tends to select solutions with much larger number of segments than reported by D_{JS}. When the number of classes is higher, the performance of AIC slightly approaches the Bayesian method, with large segmented data (i). BIC and BIC2 have slightly better performance with low class number but with higher class number it is much worse. The reason is that BIC and BIC2 tend to favour too general solutions i.e. the solutions with too small number of segments. Also the behaviour of all ML-based methods is not so good with smaller data sets (ii and iii).

The differences of tested Dirichlet priors to our MEBP are in turn much smaller. MEBP is slightly worse with 2 class data, mostly with large data sets (i). With most other types of data it seems to clearly outperform the other priors. EBP has the most similar performance. This is expected, as EBP is the most similar of the tested priors with MEBP.

Application in finding yeast cell cycle related chromosome regions

The developed segmentation method was applied to the analysis of location specific gene regulation and chromatin remodelling in baker's yeast (*S. Cereviciae*). This involved mapping the yeast cell cycle co-expression clusters as sequential data according to the chromosomal order of genes. The background information related to this problem was given in chapter 4.4.1. The developed Bayesian segmentation with MEBP was then applied to this data to discover regions including genes associated with the same co-expression clusters. For each chromosome, we also compared the obtained BF value to the BF obtained from segmentation of randomized data for that chromosome. Based on this, we selected the chromosomes with the highest differences to randomized data for further inspection. This is shown in detail in publication V.

Results of the segmentation for chromosomes II, IV, VI and XII are shown in figure 3 in publication V. The figure also shows the annotations of genes to different cell cycle phases obtained from Spellman (2002). As it was expected, chromosome IV which had the largest difference to

randomized data, included the largest amount of coherent segments, and segments that correlate with the cell cycle phase annotations. Interpretation of results is given in publication V.

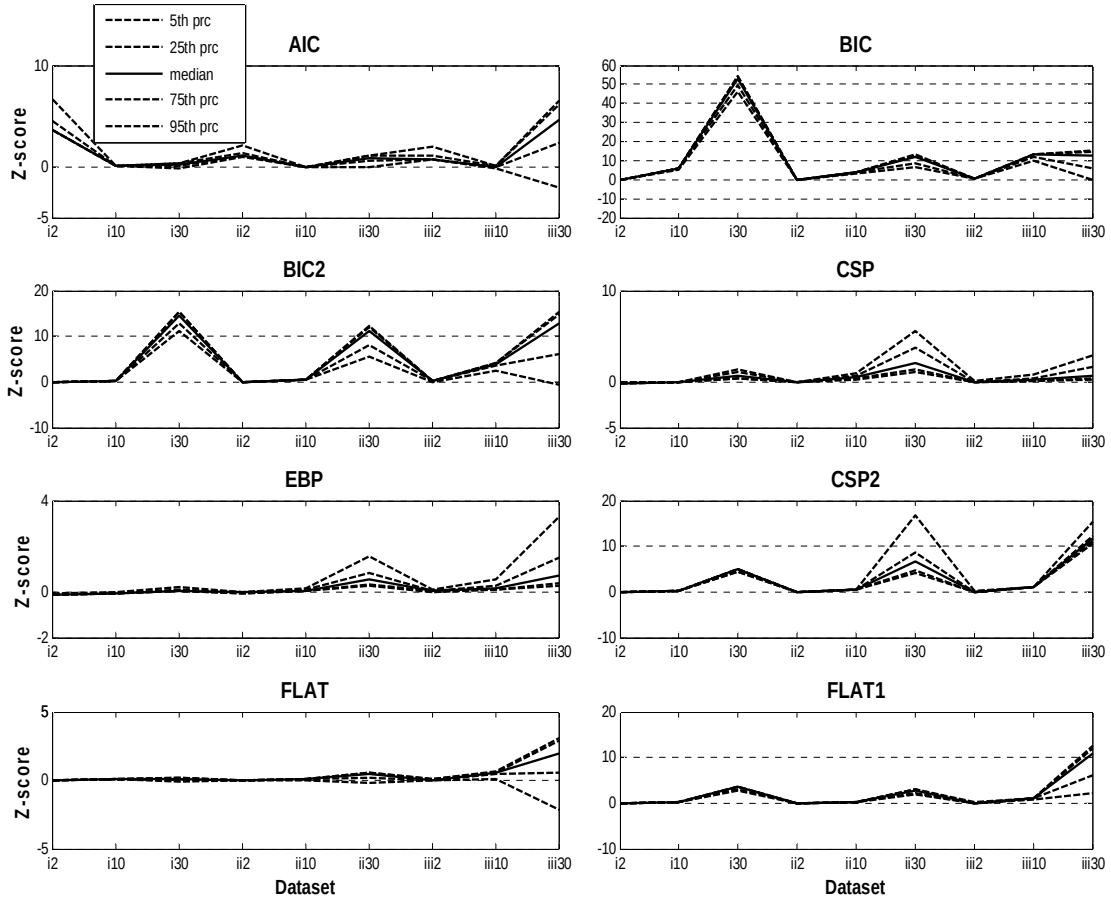

Figure 6. Graphs showing the comparison of our Bayesian model selection method with the developed MEBP prior against each competing model selection method. These include ML-based methods AIC, BIC and BIC2, and Bayes methods with different priors EBP, CSP1, CSP, FLAT and FLAT1. Types of data are indicated in the x-axis of graphs. These correspond to data of type i, ii and iii, described above, with the different number of classes: 2, 10 and 30. The values on Y-axis represent the average of CONS-JSD (set of Z-scores) of differences between our Bayes method with the developed MEBP prior and each competing method. A positive Z-score corresponds to better performance of MEBP than the compared method.

5 Conclusions and future work

Methods developed in this thesis facilitate the analysis of genome wide high throughput data sets in a novel fashion. The application of technologies such as microarrays often produces sets of genes chosen by some criterion, for example co-expression under some tested conditions. The key point in further analysis is to provide biological interpretation of this result. This often requires integration of auxiliary information, obtained from biological databases or other sources, into analysis. In this work, our goal has been to automate such interpretation by creating computational methods and associated software tools. Often, similar methods report the database attributes over-represented in the user given gene list. Our further goal was to develop methods that can also detect complex relationships contained in such data.

Contributions of this work

We have addressed multiple different aspects related to biological interpretation of high throughput data sets. First, we have developed a simple text mining approach that could observe expected and unexpected response of fish genes to environmental toxins (publication I). Secondly, we have developed a method for discovery of separate biological themes from gene lists. We used this method to discover themes from: genes showing divergent expression as a result of an antifungal drug (publication II); genes sensitive for oxidative stress (publication II); and genes showing differential expression in a nematode as a result of transferring a gene related to neurodegenerative diseases (publication III). Third, we have developed a Web enabled tool for the discovery of transcription factor binding sites from co-expressed genes (publication IV). Fourth, we have developed a segmentation method for locating chromosomal regions containing genes with homogeneous expression profiles in the yeast cell cycle (publication V). An evaluation shows that our results are reasonable and that the methods outperform the existing alternatives.

We have shown that the developed methods can help in the interpretation of data, created from the associations between genes and biological attributes. Such data is categorical, and often of very high dimensionality. This applies especially to the clustering method presented in publication II, which uses database records from the GO database as attributes.

The principal question in clustering concerns the appropriate number of clusters. We have presented two solutions in this work. First, in publication II we observe the stability of clusters in different clustering solutions, obtained non-deterministically from a partitional clustering procedure. We have produced an evaluation of this method and conclude that it is suitable for exploratory analysis of gene sets. Secondly, in publication V, we represent segmentation as Bayesian model in order to evaluating the appropriateness of different solutions. We have evaluated this approach using simulated data and found that it outperforms the existing methods in the field.

The developed software should be very useful for the molecular biologists performing microarray data analysis or other high throughput screening. The POXO tool, published in publication IV, is an example of a large scale software series, which fully implements one practical entity, the discovery of regulatory binding sites from the user obtained co-expressed genes. The use of such a system is very advantageous, as it is possible to define a pipeline of different tools according to user's needs. In addition, the user does not have to be concerned about the compatibility between different analysis modules.

Future work

The methods developed in this thesis aid in biological interpretation of the data obtained from DNA-microarrays and other high throughput technologies. A natural step further would be to increase automation and interpretability, so that the results would be more interpretable for a standard molecular biologists. Another step further would be to carry the systems from exploratory data analysis toward the knowledge mining. This would involve representing the obtained biological interpretations from the presented methods as biological signatures for treatments, conditions, and diseases. Such knowledge could be advantageous, for example, in drug target research.

In this work, we present methods that are designed for the analysis of categorical data representing the associations between genes and biological attributes. There exist plenty of other biological databases and literature databanks, which could be used in a similar manner. However, as such data is often of high dimensionality, this would require further development of some of the presented and used methods, such as the segmentation method presented in publication V.

This work presents bioinformatic methods that are, or are to be, implemented as software programs for the end user. Implementation aids in getting the method known in the scientific community and among the potential users in industry. However, the absence of a common source or portal for such software can decrease the number of potential users. Furthermore, the heterogeneity among the software implementation styles, architectures, and user interfaces makes them often incompatible for integration. For the future, it would be advantageous to pursue common standards for implementation and publication of bioinformatic methods, software, or databases. This could possibly help in matching programs with the potential end users, processing bioinformatic tasks in an integrative manner, and enhancing the cooperative projects in the bioinformatics community.

6 References

Affymetrix Inc. (2002) Statistical Algorithms Description Document. Available at http://www.affymetrix.com.

Affymetrix Inc. (2005) Affymetrix GeneChip Operating Software with Auto Loader, version 1.4. Available at http://www.affymetrix.com.

Aggarwal, C.C., Hinneburg, A., and Keim, D.A. (2001) On the Surprising Behavior of Distance Metrics in High Dimensional Space. Lecture Notes in Computer Science *1973,* 420.

Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P. (1998) Automatic subspace clustering of high dimensional data for data mining applications, in Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 94-105.

Akaike, H. (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control *19,* 716-723.

Amari, S.-I., Cichocki, A., and Yang, H.H. (1996) A new learning algorithm for blind source separation. In Advances in Neural Information Processing Systems *8,* pp. 757-763. MIT Press, Cambridge, MA.

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

Ashburner, M., Drysdale, R. (1994) FlyBase--the Drosophila genetic database. Development *120(7),* 2077-9.

Bailey, T.L., Williams, N., Misleh, C., and Li, W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res 34(Web Server issue), W369-73.

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., Cunningham, M.L., Deng, S., Dressman, H.K., Fannin, R.D., Farin, F.M., Freedman, J.H., Fry, R.C., Harper, A., Humble, M.C., Hurban, P., Kavanagh, T.J., Kaufmann, W.K., Kerr, K.F., Jing, L., Lapidus, J.A., Lasarev, M.R., Li, J., Li, Y.J., Lobenhofer, E.K., Lu, X., Malek, R.L., Milton, S., Nagalla, S.R., O'malley, J.P., Palmer, V.S., Pattee, P., Paules, R.S., Perou, C.M., Phillips, K., Qin, L.X., Qiu, Y., Quigley, S.D., Rodland, M., Rusyn, I., Samson, L.D., Schwartz, D.A., Shi, Y., Shin, J.L., Sieber, S.O., Slifer, S., Speer, M.C., Spencer, P.S., Sproles, D.I., Swenberg, J.A., Suk, W.A., Sullivan, R.C., Tian, R., Tennant, R.W., Todd, S.A., Tucker, C.J., Van Houten, B., Weis, B.K., Xuan, S., Zarbl, H. (2005) Members of the

Toxicogenomics Research Consortium. Standardizing global gene expression analysis between laboratories and across platforms. Nat Methods *2(5)*, 351-6.

Barrett, T., Suzek, T.O., Troup, D.B., Wilhite, S.E., Ngau, W.C., Ledoux, P., Rudnev, D., Lash, A.E., Fujibuchi, W., and Edgar, R. (2005) NCBI GEO: mining millions of expression profiles--database and tools. Nucleic Acids Res 33(Database issue), D562-6.

Becker, K.G., Hosack, D.A., Dennis, G. Jr, Lempicki, R.A., Bright, T.J., Cheadle, C., and Engel, J. (2003) PubMatrix: a tool for multiplex literature mining. BMC Bioinformatics *4*, 61.

Bellman, R. (1961b) On the approximation of curves by line segments using dynamic programming. Commun ACM New York, NY, USA *4*, 284.

Bellman, R.E. (1957) Dynamic Programming. Princeton University Press, Princeton, NJ.

Bellman, R.E. (1961a) Adaptive Control Processes. Princeton University Press, Princeton, NJ.

Benjamini, Y., and Hochberg, Y. (1995) Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing. J Roy Stat Soc B Met *57(1)*, 289-300.

Berger, J.A., Hautaniemi, S., and Mitra, S.K. (2004) Comparative Analysis of Gene Expression and DNA Copy Number Data for Pancreatic and Breast Cancers Using an Orthogonal Decomposition. IEEE Computer Society Bioinformatics Conference - CSB.

Bernaola-Galvan, P., Grosse, I., Carpena, P., Oliver, J.L., Roman-Roldan, R., and Stanley, H.E. (2000) Finding Borders between Coding and Noncoding DNA Regions by an Entropic Segmentation Method. Phys Rev Lett *85*, 1342-1345.

Berry, K.J., and Mielke, P.W. (1988) Monte Carlo comparisons of the asymptotic chi-square and likelihood-ratio tests with the nonasymptotic chi-square test for sparse r×c tables, Psychol Bullet. 103 (2), 256–264.

Bingham, E., Mannila, H., and Seppänen, J. (2002) Topics in 0-1 data. To appear in KDD 2002.

Blei, M., Ng, A.Y., and Jordan, M.I. (2003) Latent Dirichlet Allocation. Journal of Machine Learning Research *3*, 993-1022.

Bolstad, B.M. (2004) Low Level Analysis of High-density oligonucleotide array data: Background, normalization and summarization [dissertation]. University of California at Berkeley.

Borovkov, A.A. (1984) Mathematical Statistics. Mir, Moscow.

Boudreault-Lapointe, L. (1988) Plant Biotechnology Vocabulary. Department of the Secretary of State of Canada, Terminology Bulletin 180.

Bowie, J.U., Luthy, R., and Eisenberg, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science *253(5016)*, 164-170.

Bowman, C., Baumgartner, R., and Booth, S. (2002) Automated analysis of gene-microarray images. IEEE CCECE 2002. Canadian Conference on Electrical and Computer Engineering. Electrical and Computer Engineering *2*, 1140-1144.

Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J.M., and Sherlock, G. (2004) GO::TermFinder – open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics *20(18)*, 3710-5.

Bradley, D.R., Bradley, T.D., McGrath, S.G., and Cutcomb, S.D. (1979) Type I error rate of the chi-square test of independence in R×C tables that have small expected frequencies. Psychol Bullet *86(6)*, 1290–1297.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart. J., Taylor, R., Vilo, J., and Vingron, M. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet *29(4)*, 365-71.

Brazma, A., Parkinson, H. (2006) ArrayExpress team, EMBL-EBI. ArrayExpress service for reviewers/editors of DNA microarray papers. Nat Biotechnol *24(11)*, 1321-2.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984) Classification and Regression Trees. Wadsworth International, Belmont, Ca.

Brunet, J.P., Tamayo, P., Golub, T.R., and Mesirov, J.P. (2004) Metagenes and molecular pattern discovery using matrix factorization. Proc Natl Acad Sci U S A *101(12)*, 4164-9.

Buntine, W.L. (2002) Variational Extensions to EM and Multinomial PCA. London, UK, pp. 23-34, 2002.

Burnham, K.P., and Anderson, D.R. (1998) Model Selection and Inference (Springer).

Burset, M., and Guigo, R. (1996) Evaluation of gene structure prediction programs. Genomics *34*, 353–367.

Cai, L., Huang, H., Blackshaw, S., Liu, J.S., Cepko, C., and Wong, W.H. (2004) Clustering analysis of SAGE data using a Poisson approach. Genome Biol *5(7)*, R51.

Candillier, L., Tellier, I., Torre, F., and Bousquet, O. (2006) Cascade Evaluation of Clustering Algorithms. In Johannes Fürnkranz, Tobias Scheffer and Myra Spiliopoulou, editors 17th European Conference on Machine Learning ECML'2006, Berlin, Germany, 18-22 september 2006 Lecture Notes in Computer Science, LNAI *4212*, pp. 574-581.

Carlin, B.P., and Louis, T.A. (2000) Bayes and empirical Bayes methods for data analysis. Boca Raton Fla., pp. 419.

Carmona-Saez, P., Pascual-Marqui, R.D., Tirado, F., Carazo, J.M., and Pascual-Montano, A. (2006) Biclustering of gene expression data by Non-smooth Non-negative Matrix Factorization. BMC Bioinformatics *7*, 78.

Celniker, S.E., Wheeler, D.A., Kronmiller, B., Carlson, J.W., Halpern, A., Patel, S., Adams, M., Champe, M., Dugan, S.P., Frise, E., Hodgson, A., George, R.A., Hoskins, R.A., Laverty, T., Muzny, D.M., Nelson, C.R., Pacleb, J.M., Park, S., Pfeiffer, B.D., Richards, S., Sodergren, E.J., Svirskas, R., Tabor, P.E., Wan, K., Stapleton, M., Sutton, G.G., Venter, C., Weinstock, G., Scherer, S.E., Myers, E.W., Gibbs, R.A., and Rubin, G.M. (2002) Finishing a whole-genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence. Genome Biol *3(12)*, RESEARCH0079.

Cheadle, C., Vawter, M.P., Freed, W.J., and Becker, K.G. (2003) Analysis of Microarray Data Using Z Score Transformation. Journal of Molecular Diagnostics 5(2).

Choudrey, R.A., and Roberts, S.J. (2001) Flexible Bayesian independent component analysis for blind source separation. In 3rd International Conference on Independent Component Analysis and Blind Signal Separation, pp. 90-95, San Diego.

Cleveland, W.S. (1979) Robust Locally Weighted Regression and Smoothing Scatterplots. Journal of the American Statistical Association. *74*, 829-836.

Cohen, B.A., Mitra, R.D., Hughesand, J.D., and Church, G.M. (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nat Genet UNITED STATES *26*, 183-186.

Comon, P. (1994) Independent component analysis — a new concept? Signal Processing *36*, 287–314.

Corander, J., Waldmann, P., Marttinen, P., and Sillanpaa, M.J. (2004) BAPS 2: enhanced possibilities for the analysis of genetic population structure. Bioinformatics *20*, 2363-2369.

Cox, M.F., and Cox, M.A.A. (2001) Multidimensional Scaling, Chapman and Hall.

Crick, F. (1970) Central Dogma of Molecular Biology. Nature. *227*, 561-563.

Crick, F.H.C. (1958): On Protein Synthesis. Symp. Soc. Exp. Biol. XII, 139-163.

Dabney, A.R., and Storey, J.D. (2005) A New Approach to Intensity-Dependent Normalization of Two-Channel Microarrays. UW Biostatistics Working Paper Series. Working Paper 266.

Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H, Watson SJ, Meng F. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res *33(20)*, e175.

Datta S, Datta S. Evaluation of clustering algorithms for gene expression data. BMC Bioinformatics *7 Suppl 4*, S17.

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

Deng, N., and Duan, H. (2004) The Automatic Gridding Algorithm Based on Projection for Microarray Image. Proceedings of the 2004 International Conference on Intelligent Mechatronics and Automation.

Dennis, G. Jr, Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., and Lempicki, R.A. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol *4(5)*, P3.

Diehr, G. (1985) Evaluation of a Branch and Bound Algorithm for Clustering. SIAM Journal on Scientific and Statistical Computing *6*, 268-284.

Ding, C., He, X., Zha, H., Simon, H. (2002) Adaptive Dimension Reduction for Clustering High Dimensional Data. Proc. ICDM 2002.

Dohrmann, P.R., Butler, G., Tamai, K., Dorland, S., Greene, J.R., Thiele, D.J., and Stillman, D.J. (1992) Parallel pathways of gene regulation: homologous regulators SWI5 and ACE2 differentially control transcription of HO and chitinase. Genes Dev, 93-104.

Domingos, P., and Pazzani, M. (1997) On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. Mach. Learn. *29*, 103-130.

Douglas, D.H., and Peucker, T.K. (1975) Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. Canadian Cartographer *10*, 112-122.

Eccles, I., and Su, M. (2004) Illustrating the curse of dimensionality numerically through different data distribution models. In ISICT '04: Proceedings of the 2004 international symposium on Information and communication technologies. Pp. 232-237. Las Vegas, Nevada. Trinity College Dublin.

Ewing, R.E.; Pilant, M.S.; Wade, J.G.; Watson, A.T. (1994) Estimating parameters in scientific computation - A survey of experience from oil and groundwater modeling. Computational Science and Engineering, IEEE [see also Computing in Science & Engineering]. *1*, 19-.

Fearnhead, P. (2006) Exact and efficient Bayesian inference for multiple changepoint problems. Statist. Comput. *16*, 203–213.

Fearnhead, P., and Liu, Z. (2007) Online Inference for Multiple Changepoint Problems. Journal of the Royal Statistical Society, Series B *69*, 589-605.

Fire, A., Xu, S., Montgomery, M., Kostas, S., Driver, S., and Mello, C. (1998) Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature *391(6669),* 806-11.

Fisher, R.A. (1922) On the interpretation of c2 from contingency tables, and on the calculation of P. J R Stat Soc *85*, 81-94.

Fodor, I. K. A Survey of Dimension Reduction Techniques. LLNL technical report, June 2002. UCRL-ID-148494.

Forster, M.R. (2000) Key Concepts in Model Selection: Performance and Generalizability. Journal of Mathematical Psychology *44(1)*, 205-231.

Fraley, C., and Raftery, A.E. (2007) Model-based Methods of Classification: Using the mclust Software in Chemometrics. Journal of Statistical Software 18(6).

Fränti, P., Virmajoki, O., and Kaukoranta, T. (2002) Branch-and-bound technique for solving optimal clustering. Proceedings. 16th International Conference on Pattern Recognition *2,* pp. 232-235.

Fränti, P., Xu, M., and Kärkkäinen, I. (2003) Classification of binary vectors by using DeltaSC-distance to minimize stochastic complexity. Pattern Recogn Lett 24(1–3), 65-73.

Frick, R.W. (1996) The appropriate use of null hypothesis testing. Psychological Methods *1*, 379-390.

Fridlyand, J., Snijders, A.M., Pinkel, D., Albertson, D.G., and Jain, A.N. (2004) Hidden Markov models approach to the analysis of array CGH data. Journal of Multivariate Analysis *90,* 132-153.

Fukunage, K., and Narendra, P.M. (1975) A Branch and Bound Algorithm for Computing k-Nearest Neighbors. Computers, IEEE Transactions on Computers *C-24(7)*, 750- 753.

Garey, M., Johnson, D., Witsenhausen, H. (1982) The complexity of the generalized Lloyd Max problem (Corresp.), IEEE Transactions on Information Theory *28(2)*, 255- 256.

Geman, S., Bienenstock, E., and Doursat, R. (1992) Neural networks and the bias/variance dilemma. Neural Computation *4*, 1–58.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, AJ., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y., and Zhang, J. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biology *5*, R80.

Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M. What is a gene, post-ENCODE? History and updated definition. Genome Res *17(6)*, 669-81.

Getz, G., Levine, E., and Domany, E. (2000) Coupled two-way clustering analysis of gene microarray data. Proc Natl Acad Sci USA *97*, 12079-12084.

Gibbons, F.D., Roth, F.P. (2002) Judging the quality of gene expression-based clustering methods using gene annotation. Genome Res *12(10)*, 1574-81.

Gilbert, N., and Ramsahoye, B. (2005) The relationship between chromatin structure and transcriptional activity in mammalian genomes. Brief Funct Genomic Proteomic *4(2)*, 129-42.

Glymour, C., Madigan, D., Pregibon, D., and Smyth, P. 1996. Statistical inference and data mining. Commun. ACM *39(11)*, 35-41.

Grosse, I., Bernaola-Galvan, P., Carpena, P., Roman-Roldan, R., Oliver, J., and Stanley, H.E. (2002) Analysis of symbolic sequences using the Jensen-Shannon divergence. Phys.Rev.E *65*, 041905.

Guha, S., Koudas, N., and Shim, K. (2001) Data-streams and histograms. In STOC, 471–475.

Gyllenberg, M., and Koski, T. (2000) Probabilistic Models for Bacterial Taxonomy, TUCS Technical Report, No. *325*, Turku Center for Computer Science, Finland, February 2000.

Gyllenberg, M., Koski, M., and Verlaan, M. (1994) Clustering and quantization of binary vectors with stochastic complexity. Proc. IEEE Internat. Symposium on Information Theory, Trondheim, Germany, 1994.

Gyllenberg, M., Koski, T., and Verlaan, M. (1997) Classification of binary vectors by stochastic complexity. Journal of Multivariate Analysis *63*, 47–72.

Heger, A., and Holm, L. (2003) Sensitive pattern discovery with 'fuzzy' alignments of distantly related proteins. Bioinformatics 19(Suppl 1), i130-7.

Himberg, J., Korpiaho, K., Mannila, H., Tikanmäki, J., and Toivonen, H. (2001) Time Series Segmentation for Context Recognition in Mobile Devices. In ICDM '01: Proceedings of the 2001 IEEE International Conference on Data Mining. 2001. IEEE Computer Society.Washington, DC, USA.

Hosack, D.A., Dennis, G. Jr, Sherman, B.T., Lane, H.C., and Lempicki, R.A. (2003) Identifying biological themes within lists of genes with EASE. Genome Biol *4(10)*, R70.

Hosack, D.A., Dennis, G. Jr, Sherman, B.T., Lane, H.C., and Lempicki, R.A. (2003) Identifying biological themes within lists of genes with EASE. Genome Biol *4(10)*, R70.

Hsieh, W.P., Chu, T.M., Wolfinger, R.D., and Gibson, G. (2003) Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. Genetics *165*, 747-757.

Hsu, D.A. (1979) Detecting Shifts of Parameter in Gamma Sequences with Applications to Stock Price and Air Traffic Flow Analysis. Journal of American Statistical Association *74*, 31-40.

Huang, D.W., Sherman, B.T., Tan, Q., Collins, J.R., Alvord, W.G., Roayaei, J., Stephens, R., Baseler, M.W., Lane, H.C., Lempicki, R.A. DAVID Gene Functional Classification Tool: A novel biological module-centric algorithm to functionally analyze large gene list. Genome Biol *8(9)*, R183.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I., and Clamp, M. (2002) The Ensembl genome database project. Nucleic Acids Res *30(1)*, 38-41.

Hubert, L., and Arabie, P. (1985) Comparing partitions. Journal of Classification *2*, 193–218.

Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, 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., and Friend, S.H. (2000) Functional discovery via a compendium of expression profiles. Cell *102(1),* 109-126.

Hyvärinen, A. (1999) The fixed-point algorithm and maximum likelihood estimation for independent component analysis. Neural Processing Letters *10(1),* 1-5.

Hyvärinen, A., and Oja, E. (2000) Independent Component Analysis: Algorithms and Applications. Neural Networks 13(4-5), 411-430.

Ideker, T., Galitski, T., and Hood, L. (2001) A new approach to decoding life: systems biology. Annu Rev Genomics Hum Genet *2,* 343-72.

Irizarry RA, Bolstad BM, Collin F, et al. (2003a) Summary of Affymetrix GeneChip probe level data. Nucleic Acids Res *31(4),* e15.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., and Speed, T.P. (2003b) Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. Biostatistics *4(2),* 249-264.

Jain, A.K., and Dubes, R.C. (1988) Algorithms for clustering data. Prentice Hall, Englewood Cliffs, New Jersey, 1988.

Johnson, S.C. (1967) Hierarchical Clustering Schemes. Psychometrika *2,* 241-254.

Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G.R., Wu, G.R., Matthews, L., Lewis, S., Birney, E., and Stein, L. (2005) Reactome: a knowledgebase of biological pathways. Nucleic Acids Res 33(Database issue), D428-32.

Kanehisa, M., and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res *28(1),* 27-30.

Kankainen, M., and Holm, L. (2004) POBO, transcription factor binding site verification with bootstrapping. Nucleic Acids Res *32,* W222–W229.

Kankainen, M., and Holm, L. (2005) POCO: discovery of regulatory patterns from promoters of oppositely expressed gene sets. Nucleic Acids Res *33,* W427–W431.

Kass, R.E., and Raftery, A.E. (1995) Bayes factors. Journal of the American Statistical Association *90,* 773-795

Kaufman, L., and Rousseeuw, P.J. (1990) Finding groups in data. Wiley, New York.

Kaukoranta, T. (1999) Iterative and Hierarchical Methods for Codebook Generation in Vector Quantization, Ph.D. Thesis, TUCS Dissertations *22*, University of Turku, Turku, Finland, December 1999.

Kehagias, A.E., and Nidelkou, V.P. (2005) A dynamic programming segmentation procedure for hydrological and environmental time series. Stoch Environ Res Risk Assess *20*, 77–94

Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H., and Phillips, D.C. (1958) A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature *181(4610)*, 662-6.

Keogh, E., Chu, S., Hart, D., and Pazzani, M. (2003) Segmenting time series: a survey and novel approach, In: Data mining in time series databases. World Scientific Publishing Company, Singapore.

Keogh, E., and Smyth, P. (1997) A probabilistic approach to fast pattern matching in time series databases. In: Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD), pp. 24-30.

Kerr, M.K., and Churchill, G.A. (2001) Statistical design and the analysis of gene expression microarray data, Genet Res *77*, 123–128.

Kim, P.M., and Tidor, B. (2003) Subsystem identification through dimensionality reduction of large-scale gene expression data. Genome Res *13*, 1706-1718.

Knudsen, S. (2002) A Biologists Guide to Analysis of DNA Microarray Data. John Wiley & Sons, Inc. New York

Koeppen, M. (2000) The Curse of Dimensionality, 5th Online World Conference on Soft Computing in Industrial Applications (WSC5).

Kohonen, T. (1988) Self-Organization and Associative Memory, Springer-Verlag, New York.

Kohonen, T. (1995) Self-Organizing Maps, Springer-Verlag, Berlin, Germany.

Kontkanen, P., Lahtinen, J., Myllymäki, P., Silander, T., and Tirri, H. (2000) Supervised model-based visualization of high-dimensional data. Intelligent Data Analysis *4*, 213–227.

Kontkanen, P., Myllymäki, P., Buntine, W., Rissanen, J., and Tirri, H. (2002) An MDL framework for data clustering. Technical Report 2002-8, Helsinki University of Technology. Helsinki Institute for Information Technology (HIIT), Helsinki.

Koontz, W.L.G., Narendra, P.M., and Fukunaga, K. (1975) A Branch and Bound Clustering Algorithm. IEEE Transactions on Computers *C-24*, 908- 915.

Kraemer, D.F., and Woolson, R.F. (1987) A comparison of tests of homogeneity for sparse contingency tables, Comm. Statist. Comput. Simulation *16(2)*, 465–483.

Krasnov, A., Koskinen, H., Pehkonen, P., Rexroad, C.E. 3rd, Afanasyev, S., and Molsa, H. (2005) Gene expression in the brain and kidney of rainbow trout in response to handling stress. BMC Genomics *6(1)*, 3.

Kullback, S., and Leibler, R.A. (1951) On information and sufficiency. Annals of Mathematical Statistics *22(1)*, 79-86.

Kuncheva, L.I., Hadjitodorov, S.T., and Todorova, L.P. (2006) Experimental Comparison of Cluster Ensemble Methods, The 9th International Conference on Information Fusion, 2006.

Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. *313(5795)*, 1929-35.

Lancaster, H. O. (1961) Significance Tests in Discrete Distributions. Journal of the American Statistical Association, *56(294)*, 223-234.

Lance, G.N., and Williams, W.T. (1967) A general theory of classificatory sorting strategies 1. Hierarchical systems, The Computer Journal, *9(4)*, 373-380.

Lawal, H.B., and Upton, G.J.G. (1990) Comparisons of some chi-squared tests for the test of independence in sparse two-way contingency table. Biometrical J 32 (1), 59–72.

Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., and Wootton, J.C. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science UNITED STATES *262*, 208-214.

Lee, D.D., and Seung, H.S. (1999) Learning the parts of objects by non-negative matrix factorization. Nature *401(6755)*, 788-791.

Lee, D.D., and Seung, H.S. (2001) Algorithms for Non-negative Matrix Factorization", Advances in Neural Information Processing Systems 13: Proceedings of the 2000 Conference, 556-562, MIT Press.

Lee, M.D., and Pope, K.J. (2006) Model selection for the rate problem: A comparison of significance testing, Bayesian, and minimum description length statistical inference. Journal of Mathematical Psychology *50*, 193–202.

Lee, Y.S., and Mrksich, M. (2002) Protein chips: from concept to practice. Trends Biotechnol 20(12 Suppl), S14-8.

Li, C., and Wong, W.H. (2001b) Model-based analysis of oligonucleotides arrays: expression index computation and outlier detection. Proc Natl Acad Sci USA *98*, 31–36.

Li, C., and Wong, WH. (2001a) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biol *2*, 1–11.

Li, C., and Wong, WH. (2003) DNA-Chip Analyzer (dChip) In The analysis of gene expression data: methods and software. Edited by G Parmigiani, ES Garrett, R Irizarry and SL Zeger. Springer, New York. 120-141.

Li, W. (1990) Mutual information functions versus correlation functions. Journal of Statistical Physics, 60(5/6), 823-837.

Li, W. (2001) DNA segmentation as a model selection process. RECOMB01: Proceedings of the Fifth Annual International Conference on Computational Biology, pp. 204-210.

Li, W., Bernaola-Galvan, P., Haghighi, F., and Grosse, I. (2002) Applications of recursive segmentation to the analysis of DNA sequences. Computers & Chemistry *26*, 491-510.

Lipman, D.J., Altschul, S.F., and Kececioglu, J.D. (1989) A tool for multiple sequence alignment. Proc Natl Acad Sci USA *86*, 4412–4415.

Little, D.P., Braun, A., Darnhofer-Demar, B., Frilling, A., Li, Y., McIver, R.T. Jr and Koster, H. (1997) Detection of RET proto-oncogene codon 634 mutations using mass spectrometry. J Mol Med. *75(10)*, 745-50.

Liu, J.S., and Lawrence, C.E. (1999) Bayesian inference on biopolymer models. Bioinformatics ENGLAND *15*, 38-52.

Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E.L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol. *14(13)*, 1675-80.

MacQueen, J. B. (1967) Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press *1*, 281-297.

Martin, D., Brun, C., Remy, E., Mouren, P., Thieffry, D., and Jacq, B. GOToolBox: functional analysis of gene datasets based on Gene Ontology. Genome Biol *5(12)*, R101.

Marttinen, P., Corander, J., Toronen, P., and Holm, L. (2006) Bayesian search of functionally divergent protein subgroups and their function specific residues. Bioinformatics *22*, 2466-2474.

Maxam, A.M., and Gilbert, W. (1977) A New Method for Sequencing DNA. PNAS *74*, 560 - 564.

McCullagh, P. (2002) What is a statistical model? Ann Statist *30(5)*, 1225-1310.

McGee, M., and Chen Z. (2006) Parameter estimation for the convolution model for background correction of Affymetrix GeneChip data. Statistical Methods in Genetics and Molecular Biology *5*, Article 24.

McLachlan, G.L., and Basford, K.E. (1988) Mixture Models: Inference and Applications to Clustering New York: Marcel Dekker.

McNaught, K.S., Mytilineou, C., Jnobaptiste, R., Yabut, J., Shashidharan, P., Jennert, P., and Olanow, C.W. (2002) Impairment of the ubiquitin–proteasome system causes dopaminergic cell death and inclusion body formation in ventral mesencephalic cultures. J. Neurochem. *81*, 301–306.

McNaught, K.S., Perl, D.P., Brownell, A.L., and Olanow, C.W. (2004) Systemic exposure to proteasome inhibitors causes a progressive model of Parkinson's disease. Ann. Neurol. *56*, 149–162.

Mikheev, P.V.; Kheeroug, S.S.; Rogova, T.V. (1998) Visualization of hierarchical structure of multispectral remote sensing data. Geoscience and Remote Sensing Symposium Proceedings. IGARSS '98. 1998 IEEE International 4.

Myers, E.W., Sutton, G.G., Delcher, A.L., Dew, I.M., Fasulo, D.P., Flanigan, M.J., Kravitz, S.A., Mobarry, C.M., Reinert, K.H., Remington, K.A., Anson, E.L., Bolanos, R.A., Chou, H.H., Jordan, C.M., Halpern, A.L., Lonardi, S., Beasley, E.M., Brandon, R.C., Chen, L., Dunn, P.J., Lai, Z., Liang, Y., Nusskern, D.R., Zhan, M., Zhang, Q., Zheng, X., Rubin, G.M., Adams, M.D., and Venter, J.C. (2000) A whole-genome assembly of Drosophila. Science *287*, 2196-2203.

Myung, I. J., and Pitt, M. A. (1996) Applying Okham's razor in modeling cognition: A Bayesian approach. Psychonomic Bulletin & Review *4*, 79–95.

Neville, J., and Jensen, D. (2007) Bias-Variance Analysis for Relational Domains. In Proceedings of the 17th International Conference on Inductive Logic Programming, 2007.

Nelson, D.L. and Cox, M.M. (2004) Lehninger: Principles of Biochemistry, Fourth Edition. W. H. Freeman.

Nikkilä, J. (2005) Exploratory Cluster Analysis of Genomic High-Throughput Data Sets and Their Dependencies. Helsinki University of Technology. Dissertations in Computer and Information Science.

Nix, T. W., & Barnette, J. J. (1998) The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. RESEARCH IN THE SCHOOLS *5(2)*, 3-14.

Oh, M.-S., Raftery, A. (2003) Model-based Clustering with Dissimilarities: A Bayesian Approach Technical Report no. 441. University of Washington.

Olshen, A.B., Venkatraman, E.S., Lucito, R., and Wigler, M. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. Biostat *5*, 557-572.

Olson, C.F. (1995) Parallel algorithms for hierarchical clustering. Parallel Computing. *21*, 1313-1325.

Osiński, S. (2006) Improving Quality of Search Results Clustering with Approximate Matrix Factorisations. Springer Lecture Notes in Computer Science *3936*, pp. 167—178, Proceedings of the 28th European Conference on IR Research (ECIR 2006), London, UK.

Paatero, P., and Tapper, U. (1994) Positive Matrix Factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics *5*, 111-126.

Pan W. (2006) Incorporating gene functions as priors in model-based clustering of microarray gene expression data. Bioinformatics *22(7)*, 795-801.

Patrikainen, A., and Mannila, H. (2004) Subspace clustering of high-dimensional binary data - A probabilistic approach. Workshop on Clustering High-Dimensional Data and Its Applications, SIAM International Conference on Data Mining 2004, pp. 57-65.

Perks, W. (1947) Some observations on inverse probability including a new indifference rule. Journal of the Institute of Actuaries *73*, 285-334.

Pizarro, J., Guerrero, E., and Galindo, P.L. (2000) A Statistical Model Selection Strategy Applied to Neural Networks. ESANN'2000 proceedings - European Symposium on Artificial Neural Networks. Bruges (Belgium), 26-28 April 2000.

Raftery, E., Madigan, D., and Hoeting, J.A. (1997) Bayesian Model Averaging for Linear Regression Models," Journal of the American Statistical Association *92*, 179-191.

Rajapakse, M., and Wyse, L. (2003) NMF vs ICA for face recognition. Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis, 2003. ISPA 2003. Pp. 605-610.

Ramensky, V.E., Makeev, V.J., Roytberg M.A., and Tumanyan, V.G. (2000) DNA segmentation through the Bayesian approach. J Comput Biol UNITED STATES *7*, 215.

Ramer, U. (1972) An iterative procedure for the polygonal approximation of plane curves. Computer Graphics and Image Processing *1*, 244-256.

Rand, W. M. (1971) Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association *66*, 846–850.

Rao, C.R. (1964) The Use and Interpretation of Principal Component Analysis in Applied Research. Sankhya A *26*, 329 -358.

Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T.L., Wilson, C.J., Bell, S.P., and Young, R.A. (2000) Genome-wide location and function of DNA binding proteins. Science *290(5500)*, 2306-9.

Rissanen, J. (1987) Stochastic complexity. Journal of Statistics Society 4 (10), 223–239.

Rissanen, J. (1996) Fisher information and stochastic complexity. IEEE Trans. on Information Theory 42 (1), 40–47.

Roach, J.C., Boysen, C., Wang, K., Hood, L. (1995) Pairwise end sequencing: a unified approach to genomic mapping and sequencing. Genomics *26*, 345-353.

Rusakov D., and Geiger, D. (2002) Asymptotic Model Selection for Naive Bayesian Networks," UAI 2002.

Saeed, A.I., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., Braisted, J., Klapa, M., Currier, T., Thiagarajan, M., Sturn, A., Snuffin, M., Rezantsev, A., Popov, D., Ryltsov, A., Kostukovich, E., Borisovsky, I., Liu, Z., Vinsavich, A., Trush, V., and Quackenbush, J. (2003) TM4: a free, open-source system for microarray data management and analysis. Biotechniques *34(2)*, 374-8.

Salmenkivi, M., Kere, J., and Mannila, H. (2002) Genome Segmentation using Piecewise Constant Intensity Models and Reversible Jump MCMC, (European Computational Biology Conference 2002.) Bioinformatics *18 supplement 2*, S211-S218.

Sander J., Ester M., Kriegel H.-P., Xu X.: Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications, in: Data Mining and Knowledge Discovery, an Int. Journal, Kluwer Academic Publishers. Pp. 169-194.

Saviozzi, S., and Calogero, R.A. (2003) Microarray probe expression measures, data normalization and statistical validation. Comp Funct Genom *4*, 442–446.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270(5235)*, 467-70.

Schenker, A., Last, M., Bunke, H., and Kandel, A. (2003) A comparison of two novel algorithms for clustering web documents. 2nd IWWDA, 2003.

Schwarz, G. Estimating the dimension of a model. Annals of Statistics *6*, 461-464.

Sen, A., and Srivastava, M.S. (1975) On Tests for Detecting Change in Mean. The Annals of Statistics *3*, 98-108.

Seppänen, J.K., Bingham, E., and Mannila, H. (2003) A simple algorithm for topic identification in 0–1 data. In Knowledge Discovery in Databases: PKDD 2003; Cavtat-Dubrovnik, Croatia Edited by: Nada Lavrac, Dragan Gamberger, Hendrik Blockeel, Ljupco Todorovski. Springer; Pp. 423-434.

Shamir , R., and Sharan, R. (2000) CLICK: A Clustering Algorithm with Applications to Gene Expression Analysis. Proceedings ISMB 2000, pp. 307-316 (2000).

Shannon, C. E., and Weaver, W.W. (1949) The Mathematical Theory of Communication. University of Illinois Press, Urbana, IL.

Sheng, Q., Moreau, Y., and De Moor, B. (2003) Biclustering microarray data by Gibbs sampling. Bioinformatics 19(Suppl 2), II196-II205.

Sherlock, G., Hernandez-Boussard, T., Kasarskis, A., Binkley, G., Matese, J.C., Dwight, S.S., Kaloper, M., Weng, S., Jin, H., Ball, C.A., Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., and Cherry, J.M. (2001) The Stanford Microarray Database. Nucleic Acids Res *29(1)*, 152-5.

Smith, T.F., Waterman, M.S. (1981) Identification of Common Molecular Subsequences. Journal of Molecular Biology *147*, 195-197.

Snijders, A.M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A.K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J.P., Gray, J.W., Jain, A.N., Pinkel, D., and Albertson, D.G. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. Nat Genet *29(3)*, 263-4.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein D., and Futcher, B.. (1998) Comprehensive identification of cell cycle-regulated genes of

the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell UNITED STATES *9*, 3273-3297.

Spertus, E., Sahami, M., and Buyukkokten, O. 2005. Evaluating similarity measures: a large-scale study in the orkut social network. In Proceeding of the Eleventh ACM SIGKDD international Conference on Knowledge Discovery in Data Mining (Chicago, Illinois, USA, August 21 - 24, 2005) KDD '05. ACM Press, New York, NY, 678-684.

Stein, L., Sternberg, P., Durbin, R., Thierry-Mieg, J., and Spieth, J. (2001) WormBase: network access to the genome and biology of Caenorhabditis elegans. Nucleic Acids Res *29(1)*, 82-6.

Steinley, D. (2006) Profiling local optima in K-means clustering: Developing a diagnostic technique. Psychological Methods *11*, 178-192.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 *102(43)*, 15545-50.

Tate, M.W., and Hyer, L.A. (1973) Inaccuracy of the X2 test of goodness of fit when expected frequencies are small, J. Amer. Statist. Assoc. *68*, 836–841.

Terzi, E., and Tsaparas, P. (2006) Efficient Algorithms for Sequence Segmentation, SIAM Data Mining Conference (SDM).

Thorpe, G.W., Fong, C.S., Alic, N., Higgins, V.J., and Dawes, I.W. (2004) Cells have distinct mechanisms to maintain protection against different reactive oxygen species: oxidative-stress-response genes. Proc Natl Acad Sci U S A *101(17)*, 6564-6569.

Torgerson, W. S. (1958) Theory & Methods of Scaling. New York: Wiley.

Toronen, P. (2004) Selection of informative clusters from hierarchical cluster tree with gene classes. BMC Bioinformatics *5(1)*, 32.

ENCODE Project Consortium, Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., Kuehn, M.S., Taylor, C.M., Neph, S., Koch, C.M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J.A., Andrews, R.M., Flicek, P., Boyle, P.J., Cao, H., Carter, N.P., Clelland, G.K., Davis, S., Day, N., Dhami, P., Dillon, S.C., Dorschner, M.O., Fiegler, H., Giresi, P.G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K.D., Johnson, B.E., Johnson, E.M., Frum, T.T., Rosenzweig, E.R., Karnani, N., Lee, K., Lefebvre, G.C., Navas, P.A., Neri, F., Parker, S.C., Sabo, P.J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu, M., Collins, F.S., Dekker, J., Lieb, J.D., Tullius, T.D., Crawford, G.E., Sunyaev, S., Noble, W.S., Dunham, I., Denoeud, F.,

Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I.L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H.A., Sekinger, E.A., Lagarde, J., Abril, J.F., Shahab, A., Flamm, C., Fried, C., Hackermuller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A, Washietl, S, Korbel, J, Emanuelsson, O, Pedersen, JS, Holroyd, N, Taylor, R, Swarbreck, D, Matthews, N, Dickson, MC, Thomas, DJ, Weirauch, MT, Gilbert, J, Drenkow, J, Bell, I, Zhao, X, Srinivasan, KG, Sung, WK, Ooi, HS, Chiu, KP, Foissac, S, Alioto, T, Brent, M, Pachter, L, Tress, ML, Valencia, A, Choo, SW, Choo, CY, Ucla, C, Manzano, C, Wyss, C, Cheung, E, Clark, TG, Brown, JB, Ganesh, M, Patel, S, Tammana, H, Chrast, J, Henrichsen, CN, Kai, C, Kawai, J, Nagalakshmi, U, Wu, J, Lian, Z, Lian, J, Newburger, P, Zhang, X, Bickel, P, Mattick, JS, Carninci, P, Hayashizaki, Y, Weissman, S, Hubbard, T, Myers, RM, Rogers, J, Stadler, PF, Lowe, TM, Wei, CL, Ruan, Y, Struhl, K, Gerstein, M, Antonarakis, SE, Fu, Y, Green, ED, Karaoz, U, Siepel, A, Taylor, J, Liefer, LA, Wetterstrand, KA, Good, PJ, Feingold, EA, Guyer, MS, Cooper, GM, Asimenos, G, Dewey, CN, Hou, M, Nikolaev, S, Montoya-Burgos, JI, Loytynoja, A, Whelan, S, Pardi, F, Massingham, T, Huang, H, Zhang, NR, Holmes, I, Mullikin, JC, Ureta-Vidal, A, Paten, B, Seringhaus, M, Church, D, Rosenbloom, K, Kent, WJ, Stone, EA; NISC, Comparative Sequencing Program; Baylor College of Medicine Human Genome Sequencing Center; Washington University Genome Sequencing Center; Broad Institute; Children's Hospital Oakland Research Institute, Batzoglou, S, Goldman, N, Hardison, RC, Haussler, D, Miller, W, Sidow, A, Trinklein, ND, Zhang, ZD, Barrera, L, Stuart, R, King, DC, Ameur, A, Enroth, S, Bieda, MC, Kim, J, Bhinge, AA, Jiang, N, Liu, J, Yao, F, Vega, VB, Lee, CW, Ng, P, Shahab, A, Yang, A, Moqtaderi, Z, Zhu, Z, Xu, X, Squazzo, S, Oberley, MJ, Inman, D, Singer, MA, Richmond, TA, Munn, KJ, Rada-Iglesias, A, Wallerman, O, Komorowski, J, Fowler, JC, Couttet, P, Bruce, AW, Dovey, OM, Ellis, PD, Langford, CF, Nix, DA, Euskirchen, G, Hartman, S, Urban, AE, Kraus, P, Van, Calcar S, Heintzman N, Kim TH, Wang K, Qu C, Hon G, Luna R, Glass CK, Rosenfeld MG, Aldred SF, Cooper SJ, Halees A, Lin JM, Shulha HP, Zhang X, Xu M, Haidar JN, Yu Y, Ruan, Y, Iyer, VR, Green, RD, Wadelius, C, Farnham, PJ, Ren, B, Harte, RA, Hinrichs, AS, Trumbower, H, Clawson, H, Hillman-Jackson, J, Zweig, AS, Smith, K, Thakkapallayil, A, Barber, G, Kuhn, RM, Karolchik, D, Armengol, L, Bird, CP, de, Bakker PI, Kern AD, Lopez-Bigas N, Martin JD, Stranger BE, Woodroffe, A, Davydov, E, Dimas, A, Eyras, E, Hallgrimsdottir, IB, Huppert, J, Zody, MC, Abecasis, GR, Estivill, X, Bouffard, GG, Guan, X, Hansen, NF, Idol, JR, Maduro, VV, Maskeri, B, McDowell, JC, Park, M, Thomas, PJ, Young, AC, Blakesley, RW, Muzny, DM, Sodergren, E, Wheeler, DA, Worley, KC, Jiang, H, Weinstock, GM, Gibbs, RA, Graves, T, Fulton, R, Mardis, E.R., Wilson, R.K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D.B., Chang, J.L., Lindblad-Toh, K., Lander, E.S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga., Y., Zhu, B., de Jong, P.J. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447(7146)*, 799-816.

Tuimala, J. (2003) Preprocessing of data. In: Tuimala J, Laine MM, eds. DNA Microarray Data Analysis. Part I, pp. 38-41. [Espoo]: CSC - Scientific Computing.

Venkatraman, E. S. (1992) Consistency results in multiple change-point situations. Technical report, Dept. of Statistics, Stanford Univ.

Verbeek, J.J. Mixture Models for Clustering and Dimension Reduction. PhD thesis, Universiteit van Amsterdam/ASCI graduate school, 2004.

Verma, D., and Meila, M. (2003) A comparison of spectral clustering algorithms, technical report uw-cse-03-05-01, University of Washington.

Virmajoki, O. (2004) Pairwise Nearest Neighbor Method Revisited. University of Joensuu, Computer Science, Dissertations 9. Joensuu, 2004, 164.

Ward, J.H. (1963) Hierarchical grouping to optimize an objective function, Journal of American Statistical Association *58(301),* 236-244.

Watson, J.D., Crick, F.H. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171(4356),* 737-8.

Whitehead, A., and Crawford, D.L. (2005) Variation in tissue-specific gene expression among natural populations. Genome Biol *6(2),* R13.

Wishart, D. (1968) Mode analysis: A generalization of nearest neighbour which reduces chaining effects. In Proceedings of the Colloquium in Numerical Taxonomy, pp. 282–308, University of St. Andrews, Fife, Scotland, Academic Press.

Woo, Y., Affourtit, J., Daigle, S., Viale, A., Johnson, K., Naggert, J., and Churchill, G. (2004) A comparison of cDNA, oligonucleotide, and Affymetrix GeneChip gene expression microarray platforms. J Biomol Tech *15(4),* 276-84.

Wray, G.A., Hahn, M.W., Abouheif, E., Balhoff, J.P., Pizer, M., Rockman, M.V., and Romano,L.A. (2003) The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol *20,* 1377–1419.

Xu, W., Liu, X., and Gong Y. (2003) Document clustering based on non-negative matrix factorization. SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval. Pp. 267-273.

Yang, Y.H., Dudoit, S., Luu, P., and Speed, T. (2001) Normalization for cDNA microarray data, in: M.L. Bittner, Y. Chen, A.N. Dorsel, E.R. Dougherty (Eds.), Microarray Optical Technologies and Informatics, SPIE, Society for Optical Engineering, San Jose.

Yeung, K.Y., and Ruzzo, W.L. (2001) Principal component analysis for clustering gene expression data. Bioinformatics. *17(9),* 763-74.

Yu, L., Lai, K.K., Wang, S., and Huang, W. (2006) A Bias-Variance-Complexity Trade-Off Framework for Complex System Modeling. In M. Gavrilova et al. (Eds.): ICCSA 2006, LNCS *3980,* pp. 518 – 527.

Zhang, N. (2005) Change-point detection and sequence alignment: statistical problems of genomics. Ph.D. Thesis, Statistics Department, Stanford University, Stanford.

Zhang, X., Odom, D.T., Koo, S.H., Conkright, M.D., Canettieri, G., Best, J., Chen, H., Jenner, R., Herbolsheimer, E., Jacobsen, E., Kadam, S., Ecker, J.R., Emerson, B., Hogenesch, J.B., Unterman, T., Young, R.A., and Montminy, M. (2005) Genome-wide analysis of cAMP-response element binding protein occupancy, phosphorylation, and target gene activation in human tissues. Proc Natl Acad Sci U S A *102(12)*, 4459-64.

Zhou, Y., Shie, F.S., Piccardo, P., Montine, T.J., and Zhang, J. (2004) Proteasomal inhibition induced by manganese ethylene-bis-dithiocarbamate: relevance to parkinson's disease. Neuroscience *128*, 281– 291.

Kuopio University Publications G. - A.I.Virtanen Institute

G 36. Valonen, Piia. A multimodal NMR study of apoptosis induced by HSV-tk gene therapy in a rat experimental glioma model.
2005. 87 p. Acad. Diss.

G 37. Järvinen, Aki. α-Methylated polyamine analogues: tools to study polyamine metabolism and polyamine oxidase.
2005. 63 p. Acad. Diss.

G 38. Pirttilä, Terhi. Expression and functions of cystatin C in epileptogenesis and epilepsy.
2006. 103 p. Acad. Diss.

G 39. Mikkonen, Jarno Eelis. Short-term dynamics of hippocampal fast brain rhythms and their implications in the formation of functional neuronal networks in vivo.
2006. 68 p. Acad. Diss.

G 40. Wahlfors, Tiina. Enhancement of HSV-TK/GCV suicidegene therapy of cancer.
2006. 65 p. Acad. Diss.

G 41. Keinänen, Riitta et al. (eds.). The eleventh annual post-graduate symposium of the A. I. Virtanen Institute Graduate School: AIVI Winter School
2006. 57 p. Abstracts.

G 42. Nissinen, Jari. Characterization of a rat model of human temporal lobe epilepsy.
2006. 93 p. Acad. Diss.

G 43. Nairismägi, Jaak. Magnetic resonance imaging study of induced epileptogenesis in animal models of epilepsy.
2006. 77 p. Acad. Diss.

G 44. Niiranen, Kirsi. Consequences of spermine synthase or spermidine/spermine N1-acetyltransferase deficiency in polyamine metabolism - Studies with gene-disrupted embryonic stem cells and mice.
2006. 72 p. Acad. Diss.

G 45. Roy, Himadri. Vascular Endothelial Growth (VEGFs) - Role in Perivascular Therapeutic Angiogenesis and Diabetic Macrovascular Disease.
2006. 81 p. Acad. Diss.

G 46. Räty, Jani. Baculovirus surface modifications for enhanced gene delivery and biodistribution imaging.
2006. 86 p. Acad. Diss.

G 47. Tyynelä, Kristiina. Gene therapy of malignant glioma. Experimental and clinical studies.
2006. 114 p. Acad. Diss.

G 48. Malm, Tarja. Glial Cells in Alzheimer's Disease Models.
2006. 118 p. Acad. Diss.

G 49. Tuunanen, Pasi. Sensory Processing by Functional MRI. Correlations with MEG and the Role of Oxygen Availability.
2006. 118 p. Acad. Diss.

G 50. Liimatainen, Timo. Molecular magnetic resonance imaging of gene therapy-induced apoptosis and gene transfer: a role for 1H spectroscopic imaging and iron oxide labelled viral particles.
2007. 81 p. Acad. Diss.