Bioinformatics methods
for
analysis of gene expression data

Tor-Kristian Jenssen
Department of Computer and Information Science
Norwegian University of Science and Technology

May 7, 2002
Abstract

Recent technical advances in molecular biology have had a profound impact on biology and medicine. It is unlikely that this progress could have been made without the support of modern computers. For instance, the massive amounts of sequence data produced by The Human Genome Project, and other genome sequencing efforts, have been stored, shared, and analyzed through use of electronic equipment. After completion of sequencing projects, research effort will be shifted towards the analysis of the biological function of the thousands of genes in the genome(s), their mRNAs, and their proteins. One of the keys to understand the function of genes and their products is the level of expression in various cell types and states. High-throughput technologies for measuring gene expression, such as spotted cDNA microarrays, synthesized oligonucleotide microarrays, and SAGE, are maturing. Huge amounts of gene expression data have already been generated, and more will follow. Similarly as with sequence data, gene expression data will be analyzed with highly computerized methods. However, in contrast, analysis of gene expression will benefit considerably from integration of external sources of information, including clinical information of patients, published literature, and sequence data.

In summary, this thesis presents computational methods for various steps in analysis of gene expression data measured by cDNA microarrays and similar technologies. This includes methods to assess data quality, compare measurements from different technological platforms, and extract and integrate background knowledge in analysis.
Contents

Preface iii
List of Papers v

I Background 1

1 Introduction 3
  1.1 Introduction to molecular biology 3
    1.1.1 The Genome 3
    1.1.2 Genes, RNAs, and proteins 4
    1.1.3 Gene expression and function 5
  1.2 Methods for mRNA detection 8
    1.2.1 Roboticly spotted DNA microarrays 9
    1.2.2 Synthesized oligonucleotide DNA microarrays 10
  1.3 Analysis of gene expression data 10
    1.3.1 Data quality and preprocessing 11
    1.3.2 Unsupervised data analysis 13
    1.3.3 Supervised data analysis 17
    1.3.4 Survival analysis 19
    1.3.5 Combined approaches 19
  1.4 Text analysis for molecular biology 19
    1.4.1 Natural language processing 20
    1.4.2 Applications to biomedical text resources 21

2 Aims of the Study 23

3 Summary of Papers 25

4 Discussion and Conclusion 27
  4.1 Analysis of gene expression data 27
    4.1.1 Data quality 27
    4.1.2 Cross-platform compatibility 28
    4.1.3 Interpreting gene expression patterns 28
4.2 Methods for incorporating existing meta-data .............. 29
  4.2.1 Gene information extraction ............................ 29
  4.2.2 Concept-based information retrieval .................. 30
4.3 Conclusion .................................................. 31

II Papers .......................... 43
Preface

This thesis is submitted to the Norwegian University of Science and Technology (NTNU) in partial fulfillment of the requirements for the degree Doktor ingeniør.

The work has been carried out at the Department of Computer and Information Science (IDI) at NTNU, Trondheim, Norway, the Decision Systems Group (DSG) at Brigham and Women’s Hospital and the Division of Health Sciences and Technology, Harvard Medical School/Massachusetts Institute of Technology, Boston, MA, USA, and at the Department of Tumour Biology at the Norwegian Radium Hospital, Oslo, Norway. The research visit at DSG was supported by grant 134422/410 from the Norwegian Research Council.

Acknowledgments

First of all, I would like to thank my co-authors for the, from my point of view, inspiring and enjoyable collaborations. Special thanks to Eivind Hovig who has taught me a lot about genetics and biology in general, and to Arne Halaas who accepted the formal responsibilities as my supervisor the last year.

I would like to thank past and present people at the IDI for making my years at the department a great time. In particular, many thanks to the very kind and helpful administrative staff.

Finally and most importantly, I thank my parents and Cecilie, for love, encouragement and support.
List of Papers


**Paper IV** Kuo WP\(^1\), Jenssen T-K\(^1\), Butte AJ, Ohno-Machado L, Kohane IS, *Analysis of matched mRNA measurements from two different microarray technologies*, to be published in *Bioinformatics*

\(^1\) These authors contributed equally.

**Paper V** Jenssen T-K, Mette Langaas, Kuo WP, Sørensen BS, Myklebost O, Hovig E, *Analysis of repeatability in spotted cDNA microarrays*, Submitted for publication

**Paper VI** Jenssen T-K, Kuo WP, Stokke T, Hovig E, *Associations between gene expressions in breast cancer and patient survival*, Submitted for publication
Part I

Background
Chapter 1

Introduction

Computer and information science methods are becoming more and more important in biomedical research. Huge databases with DNA and amino acid sequences, gene and protein annotations and gene expression data are stored electronically and made available through the world wide web. The same data are analyzed almost exclusively with the use of computerized methods. The topic of this thesis is computer methods to aid the analysis of gene expression data for elucidation of gene function. Thus, the thesis lies in the intersection between computer science and molecular biology and with the hope that it should be read by people with either background, trivia from both fields are included in the following introduction.

1.1 Introduction to molecular biology

Self-replication is a fundamental component of all known forms of life. In order to sustain the existence of a species, individuals must be able to create new copies of the same species. In the process of reproduction, a complete set of instructions for making a new organism is passed on from parent to progeny. The same genetic information, only modified by evolution, is passed on from generation to generation.

1.1.1 The Genome

The genetic information of an organism is called its genome. This information is encoded as a sequence of the four letters A, G, C, and T. Structurally this information is stored in deoxyribonucleic acid (DNA) molecules. Each DNA molecule has the form of a right-handed double helix consisting of two complementary chains (strands) of sequentially joined deoxyribonucleotide units.

Each nucleotide contains a 2-deoxyribose sugar, one or more phosphate groups, and a nitrogen-containing base. In each strand, nucleotides are
joined by a phosphodiester linkage between the 5' carbon of one deoxyribose group to the 3' carbon of the next. There are four types of bases used in DNA, adenine (A), guanine (G), cytosine (C), and thymine (T). The genetic information is always read from the 5' end towards the 3' end of a DNA strand. The two strands are antiparallel in that the 5' end of one strand is connected with the 3' end of the other, and vice versa.

The complementarity of the two strands lies in that an A in one strand will always be connected by two hydrogen bonds with a T in the opposite strand, and vice versa, and a C in one strand will always be connected by three hydrogen bonds with a G in the opposite strand, and vice versa. When joined in a double helix, the paired bases are on the inside and the sugars and phosphates on the outside. The complementarity of the base pairs ensures a high level of fidelity in the process of DNA replication taking place in cell division and the production of egg and sperm cells, i.e., when passed on to the next generation.

The human genome contains approximately 3.1 billion base pairs packaged in chromosomes and located in the cell nucleus. There are 22 autosomes numbered 1-22 and two sex chromosomes X and Y. Humans, as other higher organisms, are diploid meaning that (normally) all somatic cells, except red blood cells, which have no nucleus, have 23 pairs of chromosomes. Trisomies occur, but most people have 22 pairs of autosomes (chromosomes 1-22). In addition, females have one pair of two X chromosomes, while males have one pair with one X chromosome and one Y chromosome. Egg cells contain one copy of each of the 23 pairs of the female, while sperm cells contain one copy of each autosome and either one X chromosome or one Y chromosome. At conception, the child thus receives one copy of chromosomes 1-22 and X from the mother and one copy of chromosomes 1-22 plus either an X chromosome or a Y chromosome from the father.

### 1.1.2 Genes, RNAs, and proteins

Although genetic information is passed from parent to child in units of chromosomes, the basic unit of heredity is the gene. Conceptually, genes are sequences of As, Cs, Gs, and Ts containing the information required to synthesize specific proteins. “Proteins are the main catalysts, structural elements, signalling messengers and molecular machines of biological tissues” [1]. In other words, proteins are the molecules that effectuate most biological processes in the cells.

Protein molecules are linear polymers of amino acids connected by peptide bonds. There are 20 common amino acids with different chemical properties. In addition to the peptide bonds connecting each amino acid with the next, bonds may also be formed between amino acids on different parts of the polymer, giving rise to a three-dimensional shape (conformation). The shape of a protein is largely determined by the amino acid sequence, and in
1.1. INTRODUCTION TO MOLECULAR BIOLOGY

turn determines the chemical properties of the protein. Rather simplified, synthesis of biochemically active proteins consists of three steps: transcription, translation, and protein folding. It should however be noted that some proteins are further modified, by, for instance, phosphorylation, glycosylation or bonding to other proteins or RNAs.

In the first step, the DNA sequence of the gene is transcribed and an intermediate product called messenger ribonucleic acid (mRNA) molecule is synthesized by RNA polymerase enzymes. Messenger RNA (and other RNAs) resemble single stranded DNA molecules but have the base uracil (U) in all positions where DNA would have thymine (T). DNA transcription involves a complex machinery, but, roughly, starts when an RNA polymerase finds a start signal, proceeds in the 5' to 3' direction and ends at a stop signal. In this process one of the two DNA strands, the template strand, is copied into an RNA transcript that is equivalent (except for the U for T substitution) to the non-template strand. In humans and other higher eukaryotes, the raw RNAs are further processed by RNA-splicing producing mRNAs.

In the second step of protein synthesis, mRNAs are read linearly by ribosomes and transfer RNAs (tRNAs). A ribosome is a large complex of RNA and protein molecules that binds together each mRNA with two tRNAs and catalyzes the formation of peptide bonds between neighbouring amino acids. The mRNA sequence is read in triplets called codons. With few exceptions, the 64 possible codons have a universal meaning across all species, each codes for a unique amino acid or function as a stop signal (there are three stop codons) for the translation process. The translation of this genetic code is carried out by the tRNA molecules, which 'transport' the correct amino acid to the ribosome upon reading a given codon.

The last step before the protein reaches a functional form is folding from a linear polymer to a three dimensional structure. The shape or conformation of the protein determines its ability to react chemically and hence also to determine its function. Thus, in order for the cell to work correctly, it is necessary that the proteins fold correctly to the right conformation. Unlike transcription and translation, protein folding does not follow a linear sequence of steps and the native structure of a protein is often but one of many physically possible conformations. Sometimes proteins do not fold as they should, despite several correction mechanisms. Misfolded proteins are normally digested before they can harm the cell.

1.1.3 Gene expression and function

According to recent estimates, the human genome contains somewhere between 30 thousand to 40 thousand genes [2, 3]. Far from all of these have been found, and fewer still have been characterized. The genes range widely in length, from just a few hundred to several hundred thousand base pairs.
Most human genes contain both protein-coding (exons) and non-coding regions (introns). During RNA splicing, regions of the RNA transcript corresponding to introns are removed and exon-corresponding regions are spliced to form the mRNA. Many genes have alternate splice variants, implying that the same gene may give rise to more than one protein. Precise estimates of the number of distinct proteins in the human body are difficult to obtain, but recent results indicate that each gene, on average, has approximately three splice variants, implying that there are over 90 thousand different proteins.

**Gene expression diversity**

Of the over 200 different cell types in the human body only three does not contain the same DNA. The vast majority of the cells contain the same set of genes but exhibit a highly diverse composition of proteins, accounting for the differences of function and structure between different cell types, and also between different states of the same cell. Thus, gene expression, or the rate at which the various proteins are produced, is the key to cell diversity. The number of active copies of a protein is regulated by several mechanisms: Transcriptional control, RNA processing control, RNA transport control, translational control, mRNA degradation control, and several modes of protein activity control [4]. The cell regulates the level of active copies of various proteins over time and in response to internal and external signals.

**Gene and protein function**

The most striking effects of gene function can be found in genes that are inherited in distinct variants, including non-functional, resulting in synthesis of different variants of the corresponding protein, or no protein at all. A number of genes are named and known by detrimental syndromes caused by their malfunctioning or other easily identified phenotypic traits linked to particular variants [5, 6]. Many of these are extreme cases in which the expression of the corresponding gene is not modulated by 'normal' mechanisms. In general, gene function, or more correctly, the biological effect of the corresponding product(s), is much more differentiated.

Gene and protein function can be described along many dimensions [7, 8, 9]. On the molecular level, protein function can be described in terms of structural class and which types of bonds the molecule can form. The latter determines with which other molecules the protein can interact. Most proteins are involved in a limited number of cellular process - for many only one - and are thus spatially restricted to specific parts of the cell and temporally restricted to specific parts of the cell cycle. On another level, proteins can be classified according to what they 'do'; for instance, whether
they catalyze chemical reactions, are structural elements, bind to DNA, or convey chemical signals, and so on. Finally, proteins and genes may be classified according to their phenotypic effects.

The genes are themselves not biochemically active. Their biological importance and functions are effectuated by, and hence, derived from the function of their proteins. Indications of protein function can be found from protein structure and from the relative levels of protein species (abundance) [10]. There are several limitations to protein-based methods in the context of studying gene function on a genome-wide level [11]. For instance, current technologies to measure protein abundance are neither sufficiently high-throughput nor sufficiently accurate to obtain precise measurements for all proteins present in complex biological samples, such as extracted from a cell. Since important parts of the control of gene expression is exerted on the mRNA level, valuable information about gene function can be obtained by measuring abundance of mRNA species. It should, however, be kept in mind that many questions cannot be answered by studying gene expression alone - not even on the level of protein abundance. For instance, co-expression is indicative of protein interaction but much more direct evidence can be found from studies of protein structure and protein complexes.

Gene databases

A number of comprehensive collections of information about gene and protein function exist. Due to the rapid progress in discovering and characterizing novel genes and proteins, the world-wide-web has become one of the primary sources of up-to-date information. Below, a few of the most central databases are described.

The Online Mendelian Inheritance in Man (OMIM) [12] database currently (as of January 25th, 2002) contains descriptions for 9,846 genes or phenotype loci. This database, which “is a continuously updated catalog of human genes and genetic disorders”, is only one of many databases containing descriptions of human genes and their biological roles. The LocusLink [13] database is another comprehensive online database with information about human genes, as well as other genetic loci in Homo sapiens and selected model organisms. Together with each locus, LocusLink provides, if known, information about the corresponding protein, function, phenotype links, etc. The database also provides meta-information pertaining to the status of its content as well as a number of links to external data. In particular, each locus is linked to a number of sequence-related resources, such as various chromosome maps and mRNA, protein and genomic sequence data.

The International Nucleotide Sequence Database Collaboration comprising the three sister databases, GenBank [14, 15], the EMBL nucleotide database [16], and the DNA Databank of Japan [17]) organize and make available all publicly available DNA sequences. Each sequence is stored in a
separate record also containing annotations such as source, references, and biological features and feature qualifiers. The structured annotations follow a fixed format defined by the DDBJ/EMBL/GenBank Feature Table.

Expressed sequence tags (ESTs) [18] are cDNA representations of fragments of expressed mRNA. Raw EST sequence data are collected in subdivisions of the nucleotide databases. Many ESTs can be linked to characterized genes and proteins, but a large number of ESTs can not be identified in such a way. The UniGene [19] database represents one of several efforts [20, 21, 22] to provide order among the many ESTs that have been sequenced. The database contains ‘clusters’ of ESTs grouped by sequence similarity, with each cluster putatively representing a unique gene.

Compared to the genomic sequence data, ESTs provide a much more direct link to proteins. When joined together into a full-length mRNAs, ESTs can be used to predict novel proteins or mapped to characterized proteins.

A number of protein-related databases are available. The Protein Data Bank (PDB) [23] maintained by the Research Collaboratory for Structural Bioinformatics is the primary source for protein structures. The SWISS-PROT database [24] is a protein sequence database with an extensive level of annotation making it the de facto catalog of proteins for a number of organisms. The TrEMBL [25] supplement to SWISS-PROT contains a non-redundant set of translations of all coding sequences in the EMBL nucleotide sequence database that do not correspond to existing SWISS-PROT entries.

In addition to these comprehensive general databases, there are a number of specialized collections of gene or protein information directed either toward specific organisms or specific diseases. Along a quite different direction, motivated by the need to create a consistent way to name and describe genes and proteins, several nomenclature and terminology databases have been created.

1.2 Methods for mRNA detection

Several methods exist for detection of mRNA levels. In addition to the methods briefly described below, mRNA expression can be measured by for instance, Northern blots, various methods based on the polymerase chain reaction (RT-PCR) [26, 27], sequencing of cDNA-libraries [18, 28], and serial analysis of gene expression (SAGE) [29, 30, 31]. Except for de novo cDNA sequencing, all methods used to measure mRNA require prior knowledge of the DNA sequence of the genes of interest. Thus, large-scale assays of gene expression are natural successors of comprehensive sequencing projects.

Most methods used to detect specific mRNA molecules rely on the fundamental principles of hybridization between complementary nucleic acids. This was discovered by first noting that duplex DNA will de-naturalize in
aqueous solution when the temperature is raised to approximately 100 degrees Celcius. The 'melting' temperature depends on properties of both the sequence and solution. If the temperature is lowered, the single strands will re-naturalize (hybridize) to form stable duplexes again. Similar hybridization reactions can be obtained to form heteroduplexes with RNA and DNA, as well as to form homoduplexes with two strands of RNA. The hybridization process will happen spontaneously if the single stranded nucleic acids are allowed to collide and stable duplexes will be formed if there is sufficient base-pair complementarity, depending on the temperature. By adjusting the temperature (and other properties of the solution), it is possible to control the stringency and use hybridization to nucleic acids with known sequence (and identity) to interrogate very specifically biological samples with unknown composition of DNA or RNA.

1.2.1 Robotically spotted DNA microarrays

Many technologies for quantitative detection of mRNAs that rely on positioning DNA probes of known identity (at least sequence) on a surface, hybridizing labelled RNA or cDNA in solution, and subsequently detecting hybridization signals at each position of the surface. One popular variant used porous membranes, typically nylon filters, and radioactivity for detection of hybridization signals [32, 33]. The use of radioactivity and type of surface puts limits on the density with which the probes can be positioned. A more recent variant of this principle used solid surfaces, typically glass microscope slides, and fluorescence for detection.

This microarray technology uses high-precision robots to position probes of cDNA clones or synthesized oligonucleotides on glass-slides [34, 35, 36]. Hybridization signals from each position (spot) is detected by confocal laser-scanning systems. The use of a glass surface and fluorescence permits the probes to be 'printed' at a very high density; several implementations routinely produce arrays with up to 20 thousand probes, and some even more. This is made possible by several factors. The probe sequences can be covalently attached to the surface. The glass is impermeable, which improves the hybridization conditions, allows smaller quantities of probe to be used, and allows the deposited probes to follow a highly regular geometry.

In addition, the glass has low intrinsic fluorescence, which is beneficial in the signal detection. Moreover, many scanner devices can detect light of different wavelengths permitting the use of simultaneous hybridization of two differently fluor-labelled mRNA samples. This allows more reliable measurements of the relative levels of mRNA abundance in the two samples, than if the two samples were hybridized separately. The measured gene expression is commonly reported as the estimated ratio of hybridization signal from the two samples.
1.2.2 Synthesized oligonucleotide DNA microarrays

Instead of positioning pre-formed probes, the probes can be synthesized directly, at specified positions, on the glass surface [37, 38, 39]. This technology, commonly referred to simply as oligonucleotide (micro)arrays, relies on light-directed synthesis of probes. High-resolution photolithographic masks direct the solid-phase DNA synthesis and can produce very densely spaced probes of arbitrary composition. The method, however, cannot be used to reliably synthesize nucleotides of arbitrary length. Typically, the arrays contain probes that are 20 to 25 bases in length.

For the purpose of measuring gene expression - these arrays can also be used for sequence analysis - the short 25-mer probes have reduced specificity. To improve specificity each sequence to be detected is represented by a probe set containing pairs of perfect match probes and mismatch probes, typically there are 20 such pairs. In each pair, the mismatch probes are identical to the perfect match probes except in one position, usually the middle. The perfect match probes are designed to be complementary to a target gene or EST sequence.

The estimate of gene expression is then calculated from the hybridization signals of all perfect match and mismatch probes. Several ways to carry out this calculation has been proposed. Without detailed knowledge of the actual sequences used in the probes, this information is unfortunately proprietary to the Affymetrix company producing these chips, it is not easy to tell which method is the better. Thus, in many applications of Affymetrix produced oligonucleotide chips, the gene expression measurement for a given target sequence represented by a single probe set is simply estimated as the difference between the average across the perfect match probes and the average across the mismatch probes. Better estimates are likely to result if sequence information could be applied and perhaps weighting the probes according to uniqueness. In such a scheme one would assign higher weight to target sequence specific probes than probes that also have similarity to regions of other target sequences.

1.3 Analysis of gene expression data

Large-scale screenings of gene expression levels have been applied in a number of biological settings [11]. These applications range in complexity from simple two-sample analyses in which the main question may be to identify differentially expressed genes to large clinically relevant screens of gene expression in up to hundred samples from patient-derived tissues [40, 41, 42, 43]. In the latter type of experiments, the goal is often to identify genes by whose expression levels it is possible to classify patients, or more generally samples, into relevant groups, for instance, to distinguish patients likely to survive or respond to a treatment from those who are not. Another type
of experiment in which large-scale gene expression have been used is time-course studies. In this type of experiments, the investigator aims to identify genes whose expression levels vary and co-vary across different time points after some reference state; the reference state could be the start of a cell cycle [44], some other point of time in the cell cycle in which cells are conveniently arrested, or immediately after a perturbation of the cell state, for instance, heat-shock or drug treatment [45].

1.3.1 Data quality and preprocessing

Since the first microarrays were seen, the technologies have been refined and improved in order to improve the quality of the data produced. However, data quality remain a central issue, and it may be considered a symptom that many practitioners advocate the use of experiment replication in order to enhance the reliability of the measurements [46, 47]. Without trying to give formal definitions, a few central data quality criteria can be listed:

sensitivity Relates to the ability to (reliably) detect low signals, i.e., rare transcripts.

specificity Relates to the fidelity of a measured signal, i.e., the degree to which a measured signal represents the quantity of interest.

repeatability Relates the how well a measurement can be repeated in an identical experimental setup.

reproducibility Relates to how well a measurement can be reproduced in a different experimental setup.

Early studies focused on sensitivity. It was shown that even rare transcripts could be detected, see for instance [48]. Few of these early studies, however, investigated how reliably rare transcripts could be detected in 'normal' settings.

Specificity, reproducibility, and repeatability are intertwined issues that all relate to the reliability of the data. Several reports evaluating these issues have been published. Not surprisingly, the results vary, but probably less than the interpretations. Proponents of the technology, in particular investigators reporting their own work, claim to have shown microarrays to be highly reproducible, while independent reports, often on other investigator's data, have been less optimistic.

Image analysis

The first part of analysis consists in obtaining estimates of expression signals for each spot on an array. Typically, two image files are generated from the scanner, one for each channel (mRNA sample). The two files are then
analyzed using specialized software, often provided with the scanner. For each spot and each channel, this analysis often starts by defining a spot-region of pixels in the image file and a corresponding background-region of pixels. Then intensity of spot pixels and background pixels are 'averaged' in order to obtain an estimate of gene expression based on background-corrected spot intensity. Most often this correction consists in subtracting the estimated background intensity from the estimated spot intensity. There are a number of choices in this process that may affect the final estimates of gene expression, such as how spot and background areas are defined and which averaging procedure is used across the pixel measurements. These topics have not been investigated in this thesis.

Filtering

Several reports have shown that mRNA species with many copies per cell (high abundance) are much more reliably, and hence, more consistently, measured with any microarray technology, than mRNA species of low abundance. A natural explanation may be that there is inherent noise in the hybridization and signal detection processes, which for the weak signals of low-abundant transcripts result in lower signal to noise ratio compared to the stronger signals of high-abundant transcripts. This has been reported also in one-dye (one-sample) platforms, but in two-dye platforms, from which ratios are calculated, these effects are amplified and the ratio estimates are confounded by uncertainty of two numbers rather than only one. As a result, it has become common practice to 'filter' out weak signals, for which it is assumed that the signal-to-noise ratio is too low for the data to be useful. This filtering process includes the identification and removal of array elements prior to further analyses.

A common approach to filtering microarray data consists in calculating a 'data-quality' score based on the raw measurements, setting some cutoff or threshold for this score and subsequently remove all array elements falling below this quality score. Several ways to calculate a quality score (or index) have been used [49, 50].

Normalization

With respect to cDNA microarrays, normalization refers to a process where the raw ratio measurements are adjusted with the aim of correcting for systematic biases in the microarray experiment. This process is by some referred to as calibration. There are several factors that can introduce systematic biases to the estimates of gene expression ratios, including differences in the two fluorescent dyes used and differences between pen tips used to transfer cDNA probes to the arrays, and differences in scanning between the two channels.
Normalization is a topic of great interest because it has much impact on the generality of the ratio estimates, and thus, the question of whether two ratios from different microarrays can be compared in a meaningful way. The topic has been investigated in detail by several [51, 52, 53].

**Calibration**

In addition to the issue of consistency within a certain technology, it would be beneficial if data produced with one technology could be compared, at least qualitatively, but optimally quantitatively, to data produced with another technology for gene expression measurement. It, however, appears that some ground remains to be covered before this can be done reliably. A central explanation to many of these problems is that all major technologies for measuring gene expression rely on a number of complex steps, many of which are highly non-linear and very difficult to calibrate.

Both robotically spotted and synthesized microarrays rely on the specificity of the hybridization process between complementary nucleic acids. Many fundamental questions regarding the dynamics of this process, for instance, the dependency of temperature and composition of the solution in which the hybridization takes place remain open [54].

### 1.3.2 Unsupervised data analysis

The possibility of measuring gene expression for thousands of genes in a single experiment is a great asset, but poses several challenges for analysis and interpretation. For the following, it will be assumed that the gene expression data has been organized in an $n \times d$ matrix $X$, with rows corresponding to genes and columns corresponding to samples. In other words, the data matrix contains measurements for $n$ genes and $d$ samples.

If the number of samples is small, or if one wish to focus on a small number of samples or genes, traditional scatter-plots may provide useful information of the data [55]. Scatter-plots provide a useful means to visually identify groups of genes or samples. Without an *a priori* hypothesis of which genes or samples to select, a simple method such as this will have less value when $n$ or $d$ are large. Visual identification of groups may then be aided by transforming the data to a lower-dimensional space, in which data-points can be usefully plotted.

Principal-component analysis (PCA) is a standard statistical approach to data-reduction [56, 55]. With a principal component transformation it is usually possible to reduce the dimensionality of the data to as few as a handful dimensions without much loss of information about the co-variance structure. Applications to microarray data have been reported [57, 58].

More commonly than relying on visual identification of 'natural' groups of genes (or samples), underlying structure in gene expression data is de-
tected by the use of clustering algorithms. Clustering is based on some computable definition of similarity or distance. The objective is to place similar objects (data-points in a multi-dimensional space) together and separate dissimilar objects into different clusters, or groups. A measure of quality of a particular clustering, i.e., partition of the set of objects, can often be defined in terms of the definition of similarity (or distance), but even for a given definition of quality it is often not possible to find the optimal clustering.

**Measures of similarity**

Several formal definitions of similarity are possible. Most definitions apply to numerical data but valid definitions for categorical data also exist. For many purposes it may be desirable to use a definition in terms of a distance metric. Given any three vectors \( x, y, \) and \( z \) in an \( m \)-dimensional space, a measure of distance \( d \) is a distance metric, or simply metric, if the following properties are satisfied:

\[
\begin{align*}
    d(x, y) &\geq 0 \\
    d(x, y) &= d(y, x) \\
    d(x, x) &= 0 \\
    d(x, z) &= d(x, y) + d(y, z)
\end{align*}
\]

(1.1) \hspace{1cm} (1.2) \hspace{1cm} (1.3) \hspace{1cm} (1.4)

Two common metrics are Euclidian distance and statistical distance.

A distance measure satisfying the first three of these conditions is called a semi-metric (distance).

The Pearson (linear) and Spearman (rank-order) correlation coefficients are examples of similarity measures that are not semi-metric.

**Hierarchical clustering**

Hierarchical clustering is probably the clustering method most widely applied to gene expression data. The result is a dendrogram, or rooted tree, with the root representing the whole dataset and the leaves representing individual observations. Each internal node represents a subset, and successively as the tree is traversed from root towards the leaves, nested subsets are found. Each level of the tree represents a partition of the dataset.

In applications to gene expression data, the most commonly used strategy for hierarchical clustering is the method often referred to as the average linkage method. Combined with Euclidian distance as the measure of similarity the method can be described by the following algorithm.

1. Initialize by forming one cluster for each data point.
2. Repeat until there is only one cluster left:

   (a) Calculate the average distance for all pairs of clusters.
   
   (b) Merge the two clusters with the smallest average distance.

There are several variations to this simple theme of iteratively merging the two clusters with the highest similarity. Some pertains to how the point-wise distances are used to compute the distance between clusters (weighted average linkage, complete linkage and single linkage). Another class of strategies do not use the point-wise distances of points within the clusters but represent the clusters by 'cluster-centers' from which cluster similarities are computed.

Ward's method for hierarchical clustering takes a different approach. While the previously described methods can be described as 'local' approaches, Ward's method uses a 'global' criterion, defined as some objective function, simultaneously computed from all clusters at each iteration [59]. At each iteration, the algorithm is then to merge the two clusters whose merge will optimize the objective function. Typically this function is defined so as to maximize within cluster similarities and minimize between cluster similarities. For instance, the objective function may be defined as the summed within cluster variance relative to the summed variance between cluster centers.

**K-means and K-medoids**

If there is reason to believe that the objects to be clustered belong to disjoint subclasses, hierarchical clustering methods may be less appropriate. The so-called K-means and the more general K-medoid methods are clustering methods that naturally produce disjoint clusters. Briefly, the K-means method proceeds by the following steps:

1. Decide upon a number K of clusters to find and initialize K clusters by defining K cluster centers.

2. Repeat until some pre-defined convergence criterion is reached:

   (a) For each data point (re-)assign it to the nearest cluster.
   
   (b) Update each cluster center as the centroid of all data points assigned to the corresponding cluster.

   Alternatively, the centroids can be updated each time a new data point is re-assigned.

   A typical criterion of convergence is to terminate at the first iteration when no data points are re-assigned to a different cluster.
The K-medoids method proceeds by similar steps and the most important difference lies in allowing more general definitions of cluster center and similarity.

If indications can be obtained with respect to a natural number of clusters expected in the data, it may be sensible to attempt to group the data into a fixed number K of groups. Otherwise, several runs of the basic algorithm can be used and the K optimizing some global criterion can be selected and the corresponding partition reported as the final result.

**Self-organizing maps**

The method of self-organizing maps (SOMs) proposed by Kohonen [60] can be viewed as (unsupervised) adaptation of neural networks and can also be used to partition high-dimensional data, i.e., to create disjoint clusters. A SOM consists of a user-specified number of reference vectors in the input space and a transformational map from the input space to a low-dimensional representation space, typically two-dimensional to enhance visualization of the resulting clusters. The clustering process proceeds by iteratively adapting the location of the reference vectors. Several applications to microarray data have been reported [61, 62].

Briefly, the self-organization follows the following steps:

1. Initialize the network by selecting a topology and a transformational map. This includes selecting the number of reference vectors and the way they are connected. Typically, the topology is chosen such that the images of the reference vectors in the representation space define regular rectangular or hexagonal grids.

2. Repeat until convergence or a pre-defined maximum number of steps have been reached:
   
   (a) Randomly select a data point.
   
   (b) Find the reference vector closest to this data point. Usually Euclidian distance is used but other metrics can also be applied.
   
   (c) Adjust the closest reference vector to be closer to the selected data point.
   
   (d) Adjust reference vectors in a neighbourhood of the closest reference vector also to be closer to the selected data point.

   Usually, both the size of the adjustments and the size of the neighbourhoods are decreased as the algorithm proceeds.

Mathematically, the SOM method can be viewed as a generalization of K-means clustering. One advantage with the former is that it is less sensitive to the number of reference vectors chosen.
1.3. **ANALYSIS OF GENE EXPRESSION DATA**

**Other clustering methods**

Clustering is a fundamental way of data analysis. In addition to several variants and generalizations of the previously described methods, an interesting extension involves the simultaneous clustering along two dimensions, e.g., genes and samples. Relationships between subsets of genes and subsets of samples can be found by clustering each dimension sequentially, but a combined simultaneous clustering along both axes may provide even more interesting relationships to be found. Recently, at least two strategies applied to gene expression data have been reported: coupled two-way clustering [63, 64] and biclustering [65].

Typical clustering methods can be described as global approaches, as they try to impose a structure on all observations in a dataset. Gene-shaving is an iterative method that can be used to identify interesting (typically by similarity) subgroups without simultaneously clustering all objects [66]. One 'cluster' is singled out in each step and observations are allowed to belong to more than one cluster.

**Comparison of cluster structures**

Different clustering methods will detect different structures in data. Although it may be impossible and even inappropriate to attempt to rank the different methods, it is frequently of interest to compare different clustering results. For instance, if the k-means algorithm is run with \( k_1 \) and \( k_2 = k_1 + 1 \), say \( k_1 = 5 \), and a unique mapping can be found for clusters from \( k_2 \) partition into the \( k_1 \) partition, this would indicate validity of the structures imposed by the two partitions, with one being a specialization of the other. Analogously, if similarities are found between partitions imposed by different clustering algorithms or different parameter settings, this would indicate that 'real' structure in the data has been detected.

The Jaccard and Kappa statistics are two measures of similarity for binary vectors that can be adapted to compare two partitions [67]. The labels attached to clusters are usually chosen quite arbitrarily and instead of comparing cluster labels for data points it is preferable to check whether relationships between pairs of observations are preserved across the two partitions being compared.

Comparison of hierarchical structures is more involved. A conceptually straightforward procedure to compare two dendrograms is to form all possible partitions and calculate similarity of pairs of partitions having the same number of groups.

1.3.3 **Supervised data analysis**

Clustering and other unsupervised methods for data analysis can in one way be viewed merely as methods for data reduction or summarization - they can
describe structure in the data but provide no explanation of the structure found. Incorporating prior knowledge into the analysis allows more directed studies to be conducted and stronger hypotheses may be tested.

In a simple two-sample comparison, a natural first question is to identify genes whose expression differs the most. As with other biological data, gene expressions are often converted to ratios in order to measure difference in terms of fold-change. In general, the aim is to study differences in gene expression from two groups of samples represented by the two samples. Often more than one sample is collected for each group, either by replication or by analyzing material from several samples from each group. If such data are available, there are several statistical methods that could, and should, be used to identify those genes with the most interesting gene expression differences across the two groups. Essentially, the analysis proceeds by averaging the expression measurements within each group and comparing the averages. Rather than ranking the genes by difference of averages, statistical methods should be used to rank the differences according to significance. Several statistical tests are available for these purposes [68, 69].

In many studies there may be data from samples from more than two groups. If the experimental design of the study is appropriate, an analysis of variance (ANOVA) may be conducted.

In addition to assessing the levels of gene expression across groups of samples for the genes in a study, it is often of interest to assess whether such differences can be used to classify new samples, i.e., to use expression data in prediction. Some of the promises of such classification analyses lies in the the possibility of using gene expression for classification of phenotypes, but it may also be possible to use gene expression to predict gene function of uncharacterized (functionally) transcripts. Already, there have been studies that have tried to classify genes according to level of expression in collections of samples [70]. Several machine learning and statistical methods for predictive analyses are applicable to gene expression data given the existence of an a priori classification of the data.

**Support-vector machines**

The statistical-learning theory formulated by Vapnik [71] introduced the support-vector machine (SVM). The SVM method is a conceptually simple approach for binary classification where the two classes are described by lying on either side of a hyperplane. Unlike Fisher's linear discriminant, the (ideally) separating hyperplane is not calculated in the input space directly but in a higher dimensional space. SVMs have been used with good results on gene expression data [70].
Other classification methods

A number of other methods for multi-variate classification exist. Artificial neural networks have been widely applied in other machine learning tasks and have recently been applied also to gene expression data [72].

1.3.4 Survival analysis

In clinical studies there may be additional information not only about the incidence of an event of interest but also the time at which the event occurred. Typical examples of such events may be the death of a patient, development of a symptom, or relapse of a disease. Statistical analyses of survival times may then be used to estimate and compare survival functions and hazard ratios for different groups of patients. The patient groups can be defined by clustering methods applied to gene expression data of any subset of genes, or in a univariate setting, by categorization of gene expression of single genes.

Kaplan-Meier [73] plots of survival curves are useful visual aids to assess survival experience the different groups. The logrank hypothesis test can be used to investigate whether survival curves are significantly different. If the groups are ordered a slight modification should be used to test for trend.

Multivariate survival analyses can be carried out with the Cox regression analysis (also called proportional hazards regression analysis). The dependent variable to model is the hazard function and the result is an expression of the hazard as a function of the covariates. Standard variable selection procedures (forward and backward) may be used for feature (covariate) selection.

1.3.5 Combined approaches

Many interesting analyses of gene expression are studies combining several modes of analysis. These includes studies combining unsupervised methods with supervised methods, but, in particular and perhaps conceptually more interesting, studies combining analysis of gene expression data with external information, such as, for instance, sequence data, information from annotation databases, as well as, information extracted from the literature.

Based on the assumption that similarly expressed genes are regulated by the same, or similar, transcription factors, several studies have reported findings of plausible regulatory elements in groups of genes co-clustered by gene expression data [74, 75, 76].

1.4 Text analysis for molecular biology

The amount of text data available in electronic form is rapidly growing. The transition from paper to digital media has revolutionized distribution and
access to information - molecular biology and other biomedical sciences are no exceptions. Through the world wide web a growing number of scientific journals are publishing electronic versions online, sometimes even before the printed versions are available. However, timely and efficient access alone is not enough to use this information productively. Relevant information has to be correctly related to the problem at hand before it can be found and put to use. These issues have spurred a lot of research into retrieval and extraction of biomedical information, many motivated by challenges of interpreting expression measurements for thousands of genes. Both information retrieval (IR) [79, 80] and information extraction (IE) [81, 82] rely on some form of text analysis.

Briefly, IR can be said to have the goal of returning a subset of relevant pieces of text from a larger set based on a user query, while IE can be said to have the goal of gathering occurrences of pre-specified pieces of information, such as descriptions of events or entities, from texts written in natural language. Although important differences exist, the two fields are quite related, both with each other as well as with other research fields processing natural language.

Both IR and IE are commonly evaluated in terms of precision and recall. Precision is defined as the number of correct answers given by the system divided by the total number of answers given by the system. Recall is defined as the number of correct answers given by the system divided by the total number of correct answers. For IR, an answer given by the system corresponds to a returned document, and for IE, an answer corresponds to a detected occurrence. The correct answers corresponds to relevant documents and actual occurrences for IR and IE, respectively. Other performance measures have also been used.

The remainder of this section will be devoted to a brief introduction to some general issues related to text analysis followed by a brief overview of applications of IR and IE within the biomedical domain.

1.4.1 Natural language processing

Computer systems for IR and IE must be able to have some 'understanding' of the text written in natural language. This understanding may be shallow with respect to the level of knowledge employed, such as that of an indexing procedure recognizing words within a text, or deep, such as that of a machine translator. Jurafsky and Martin lists six categories of language knowledge [83], of which five are relevant to IR and IE:

**Morphology** The study of meaningful components of words

**Syntax** The study of structural relationships between words

**Semantics** The study of meaning
1.4. TEXT ANALYSIS FOR MOLECULAR BIOLOGY

Pragmatics The study of how language is used to accomplish goals

Discourse The study of linguistic units larger than a single utterance

IR systems tend to rely on a more shallow understanding of the text. This depends on the indexing language but some may rely on only morphological knowledge for indexing and query matching. IE systems frequently use also knowledge from the other categories [82].

General natural language understanding is difficult. Besides the complexity, natural language is inherently ambiguous. The latter may be less of a problem with respect to IR and IE as applied to scientific texts than in general. Nevertheless, even text understanding in limited and fairly precise scientific domains require many tasks to be solved. This has lead to highly modularized systems, where each module is designed to solve a smaller sub-task [84]. A list of common tasks is given below:

Pre-processing Includes splitting the text into smaller units and possibly filtering of some parts.

Tagging Determination of grammatical categories (or part-of-speech) for single words.

Parsing Determination of phrase structures for sequences of 'tagged' words.

Semantic interpretation This includes lexical semantics (the determination of the meaning of individual words), as well as determination of the meaning of sentences and still larger units of text.

A variety of methods have been used to solve these tasks. A comprehensive presentation of statistical methods can be found in Manning and Schütze [85]. A general description of language processing that also treats (finite) state machines, formal rule systems, and logic-based representations of language was given by Jurafski and Martin [83].

1.4.2 Applications to biomedical text resources

The already large biomedical literature is rapidly growing along with the rapid increase in mapping and characterization of genes. This has been an important additional motivation for IR and IE applications in these domains.

Information Retrieval

One of the primary entry points to the biomedical literature can be found in the PubMed service [86] from the National Library of Medicine, USA. Through this web-interface, users can search the MEDLINE database of citation records from most biomedical publications from 1966 up to present.
These records contain, among other fields, abstracts (for most articles), titles, author information, and MeSH (Medical Subject Headings) keywords. In addition to being a central IR service for biomedical citations, PubMed (MEDLINE) is a frequently used source of text data for IE applications in this field.

GeneCards[87, 88] is an online database aimed at providing efficient access to information about genes. In this database, each gene has a 'document' termed a gene card. Each such card contains information about the corresponding gene gathered from a number of source databases. Retrieval is based on index lookup and the returned documents are ranked. The system also uses several methods for query reformulation to aid the user if a search resulted in zero hits.

MedMiner[89] is another internet tool for retrieval of gene-related information. Briefly, this system can be described as a text-filtering system, but MedMiner also includes highlighting and other presentation techniques.

Information Extraction

A short review of IE research in molecular biology was made by Andrade and Bork [90]. In the following overview, both work reviewed there and more recent projects will be described.

Many entities in molecular biology, such as genes, proteins and chemical substances, have complex names. Detecting and disambiguating such names are important first steps in extracting relevant information [91, 92, 93, 94, 95]. Extraction of entity references has interest itself, as well as being essential to extraction of relationships between entities and to text-mining methods using a document-entity index.

Efforts to extract specific relationships between entities include extraction of binding relationships [96], regulatory relationships [97, 98, 99], as well as more general classes of protein interactions [92, 100, 101]. Also extraction of relationships between entities of different types have been investigated [102, 103, 104].

Other protein-related information 'extraction' efforts have attempted to use text-mining to gather protein annotations[105, 106] and information about the cellular location of proteins[107].

Genes do not physically interact with other genes. There are, nevertheless, several other fundamental ways in which two genes may have a common relationship. Indirect methods to extract, or infer, relationships between pairs of genes have been shown to be promising[108, 109, 110, 111].
Chapter 2

Aims of the Study

As briefly described above, there are several critical steps in the process of measuring gene expression, both in general for all commonly used mRNA detection methods and in particular for the cDNA microarray technology. Moreover, interpretation of gene expression data requires inclusion of background information from a wide variety of sources, in addition to knowledge of mathematical methods to analyze data.

The general aim of the study was to develop and implement novel methods to aid the analysis of gene function through the study of gene expression measurements. Rather than focusing on a single step in the process, this goal has been pursued by identifying critical parts where a contribution could be made. Conceptually, however, the work presented herein can be divided into two distinct areas, each with specific aims:

I) Text mining of biomedical literature.

II) Statistical analysis of gene expression data from high-throughput gene expression technologies such as DNA microarrays.

Within the first area, the aim was to develop methods to extract information from the published literature pertaining to known genes, structure this information, and make it available for studies of gene function in-large.

Within the second area, the aim was to develop methods to analyze the quality and biological significance of gene expression data.
Chapter 3

Summary of Papers

Papers I-III presented a novel method for extracting structured information about human genes from published literature and other free-text sources. The method was applied to citation records in the MEDLINE database.

Papers IV and V presented novel methods for assessing quality and biological significance of gene expression levels measured with the use of DNA microarrays.

Paper VI presented a novel method applying Kaplan-Meier analysis to a large number of genes and subsequently mapping Kaplan-Meier derived p-values to chromosomal positions in order to identify regions of human chromosomes enriched for genes being significant with respect to survival of breast-cancer patients.

**Paper I** described a gene network extracted from the MEDLINE database by indexing human genes and linking genes that co-occurred in the free-text fields of citation records of biomedical articles.

The structured information represented by the extracted network was put to use in a novel method for analysis of gene expression data. This method was validated by re-analyzing two large, publicly available data sets. The results confirmed findings from the original publications, indicating that the automated analysis was able to identify similar patterns and explanations as the original authors but in significantly less time. More interestingly, for one of the data sets, the automated analysis was able to point to an important observation missed by the original authors. This observation was confirmed independently by other works.

**Paper II** described further analyzes of the gene network extracted from the MEDLINE database. In this paper, an in-depth analysis of the relation defined by MEDLINE record co-occurrence was presented. The analysis was directed towards the family of Peroxisome proliferative activated receptor (PPAR) genes.
Paper III described a different application of the index-information underlying the gene network in enhancing information retrieval. The novel method for document retrieval is based on a parameterized heuristics to set-covering, where the set to be covered is a set of genes. The results of this paper indicated that information retrieval based on set-covering may be a viable alternative to standard Boolean operators implemented in conventional search interfaces, while at the same time being much simpler to use than more advanced query interfaces.

Paper IV described a large-scale analysis of the correlation between gene expression measurements obtained with microarrays having photolithographically synthesized oligonucleotide probes and microarrays with robotically spotted cDNA probes. The main results of this paper was: 1) that these two technologically different platforms are unlikely to give comparable measurements, and 2) to provide a sound methodology for similar comparisons.

Paper V described a method for evaluation of the quality of cDNA microarrays using measurements of gene expression ratios from multiply spotted clones. The paper also presented a novel method for adaptively flagging and removing measurements from spots where the signal-to-noise ratio is to low.

Paper VI described an analysis of the significance with respect to patient survival of gene expression levels measured in mRNA samples from breast cancer patients. This analysis included a novel application of the logrank test for survival based on gene expression data, including a comparison with data on chromosomal aberrations. The analysis also included a novel application of the tools developed in Paper I.
Chapter 4

Discussion and Conclusion

4.1 Analysis of gene expression data

4.1.1 Data quality

Reproducibility and repeatability are two central dimensions along which to assess quality of measurements. Based on the large number of possible sources of variation between two different cDNA microarray experiments, even when carried out by the same person with the same equipment, we considered it best to assess repeatability through multiple measurements on a single microarray slide. In paper V we presented a method to assess data quality in microarray experiments by evaluating variation of measurements between array elements supposed to measure expression of the same gene. At the time, this was, to the best of our knowledge, the first use of this ANOVA-type analysis to assess data quality. The main objection to this method is probably that the assumptions in the ANOVA model may be too strong. First, the assumption of normality of log-transformed expression ratios is clearly debatable. Second, we further assumed that each log-transformed ratio measurement of gene \( i \) was distributed as \( N(\mu_i, \sigma_i) \), with all \( \sigma_i = \sigma \), i.e., that the variations \( \sigma_i \) around each mean \( \mu_i \) were equal. Apparently, this is an even stronger assumption. More work is need in order to assess the appropriateness of these assumptions.

In paper V we also suggested how the repeatability analyses could be used to direct a quality-based data filtering strategy. Previous works have used several criteria as basis for microarray data filtering. As far as we know, paper V was the first report that suggested the use of single-array data quality criteria to direct the filtering process for each microarray individually. An important limitation with this approach is that the set of multiply spotted genes is representative with respect to the total set of genes printed on each microarray, in order to avoid biases in the filtering process. To ensure this, a minimum number of genes must be represented in multiples, reducing the number of genes that can be simultaneously measured. We believe the
benefits far outweigh the disadvantage, as we consider this the most direct way to measure data quality. Furthermore, since the technology is rapidly developing and soon allows more array elements than the number of distinct genes, this is not a major limitation.

4.1.2 Cross-platform compatibility

For the scientific community, it would be desirable if gene expression data from different studies could be collected in publically available repositories [112]. Careful annotation will however alone not be sufficient for broad cross-study utilization of data. An important issue that needs to be resolved is that the different technological platforms measures mRNA levels on different scales. This is not an issue of the relative merits with respect to ease of use, relative cost, and applicability in various experimental contexts, but the question of how data obtained from different platforms may be used in a common analysis.

In paper IV we addressed the issue of measurement agreement between in situ synthesized oligonucleotide arrays and spotted cDNA microarrays. The results of this paper indicated that the two technological platforms did not consistently measure the same genes as highly expressed, nor the same genes as having low expression. As expected, ratios from the cDNA microarrays showed no correlation to oligonucleotide measurements. Some correlation existed between the background-corrected intensities from the target mRNA samples of the cDNA arrays and the oligonucleotide measurements. However, this correlation was much lower than would be required to use data from these two platforms interchangeably in numerical analyses. An important limitation of this analysis was the fact that the two experiments had been conducted in different laboratories. In particular, the two sets of mRNA samples interrogated had not been cultivated identically. Thus, the result is not as strong as it would have been if the mRNA samples had been identical and subsequently measured on different platforms. More experimental work is needed to address this question.

4.1.3 Interpreting gene expression patterns

In paper VI we presented a novel application of the logrank test for survival time analysis to associate gene expression patterns to survival status of breast cancer patients. This enabled the identification of several previously uncharacterized genes whose expression levels exhibited very significant associations to survival. The results were validated by the identification of several previously reported cancer markers. The previously uncharacterized genes as well as the genes with no previous link to breast cancer are interesting candidates to be included in future studies of this cancer.

By combining the p-values from logrank tests with chromosome data, we
were able to locate genes with expression patterns significantly associated to survival in a number of previously reported amplicon regions. This strategy indicates how gene expression data may complement genotype assays to further enhance the understanding of the biological roles of various genes and genotypic variants. Although conceptually straightforward, the chromosome plots with log-transformed p-values well illustrated this combined approach.

To further analyze the associations between gene expression and survival, we used the tools presented in paper I to combine information from log-transformed p-values with meta-data from the literature. This resulted in the identification of several genes for which the combination of literature information and gene expression data resulted in a high relevancy-score. In particular, we could identify two pairs of subunits of two different proteins which ranked highly.

4.2 Methods for incorporating existing meta-data

4.2.1 Gene information extraction

Co-occurrence networks

Efficient interpretation of gene expression data analyses commonly depend on a largely manual process of putting together different pieces of information from various databases and the literature. The number of genes that are commonly under study as well as the amount of relevant literature makes it impossible to a priori have in-depth knowledge or a posteriori to survey but a small fraction of what is potentially relevant. Even if focusing on a relatively small subset of, say, 50 to 100 genes, it is a huge task to obtain a comprehensive view of all existing knowledge potentially relevant for the study in question. The approach presented in paper I represents a principally new approach for automating the process of putting together new experimental data with existing knowledge. Although the underlying methods are, in principal, very simplistic, the potential in automating the process of ‘annotating’ gene clusters (any arbitrary set of genes) with pre-extracted information from the literature is clearly demonstrated. Clearly, more advanced text-mining methods will further enhance both the precision and the level of detail of these annotations.

A heuristic similar to the one we used in paper I was also used by Stapley and Benoit to extract a literature network for genes genes from the Saccharomyces Cerevisiae (baker’s yeast) organism. They indexed co-occurrences of yeast genes in a pre-selected subset of MEDLINE citation records. On one important point, however, Stapley and Benoit arrived at a conclusion different from what suggested by our analyses: Specificity of gene terms. We found a large number of gene symbols for human genes that were problematic with respect to use in indexing. In particular, many short gene
symbols coincided with commonly used acronyms. This severely hampered the precision of our indexing procedure. Another problem lies in the fact that a number of gene symbols have been used for different genes. This is potentially a much more difficult problem than gene symbols coinciding with other acronyms, as it is harder to use contextual clues to resolve these ambiguities. Also, the fact that many genes have alternative names and symbols represent a challenge with respect to achieving high recall. In order to correctly index all references to a given gene, it is obviously necessary to have prior knowledge of all names used to refer to the genes. Our solution was to collect gene symbols and names from several sources, in addition to the official nomenclature database. More work is needed in this respect to make this compiled nomenclature collection as complete as possible.

An issue that we did not try to resolve is related to the complexity of the (full) gene names. Although genes with long and complex names tend to be referred to by the gene symbol abbreviation, some textual occurrences may be missed. Correctly detecting full names is further complicated by the many possible ways these names can be written. Comprehensive tables of alternative spellings and word-orders should be made to enable full-name indexing to complement our largely symbol-oriented indexing procedure.

**Alternative network construction methods**

In paper II we suggested an alternative approach to create connections between genes from literature mining. This alternative approach linked genes based on information about citation structure in the literature. The premise was that two genes have things in common if they can be linked to reference articles with similar citation structures. This approach was also compared to the co-occurrence indexing strategy used in paper I. The results indicated there was some overlap between the two methods, but that important differences existed. One explanation for the differences can be found in the fact that our co-occurrence indexing was based on titles and abstracts while citation structure is likely to reflect overall document content. Another explanatory factor is that the reference document sets were unlikely to represent a complete picture of the selected genes.

A related use of reference document sets were proposed by Shatkay et al. They used document similarity instead of citation structure to define the links between reference documents. Also this approach required the use of carefully maintained reference document sets.

**4.2.2 Concept-based information retrieval**

In paper III we presented a novel heuristic to set-covering and application of the heuristic to concept-based information retrieval. The heuristic represents an alternative to Boolean operators in combining search terms and was
4.3 Conclusion

The work presented in this thesis include contributions to several important steps in the process of obtaining and analyzing gene expression data.
Bibliography


Part II

Papers