Visualization and Exploration of Transcriptomics Data

Nils Gehlenborg

Sidney Sussex College

A dissertation submitted to the University of Cambridge
for the degree of Doctor of Philosophy

European Molecular Biology Laboratory,
European Bioinformatics Institute,
Wellcome Trust Genome Campus,
Hinxton, Cambridge, CB10 1SD,
United Kingdom.

Email: nils@ebi.ac.uk

October 12, 2010

To Maureen.

This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgements.

This dissertation is not substantially the same as any I have submitted for a degree, diploma or other qualification at any other university, and no part has already been, or is currently being submitted for any degree, diploma or other qualification.

This dissertation does not exceed the specified word limit of 60,000 words as defined by the Biology Degree Committee.

This dissertation has been typeset in 12 pt Palatino using LaTeX2ε according to the specifications defined by the Board of Graduate Studies and the Biology Degree Committee.

October 12, 2010 Nils Gehlenborg

Visualization and Exploration of Transcriptomics Data

Summary

October 12, 2010

Nils Gehlenborg
Sidney Sussex College

During the last decade, high-throughput analysis of transcriptomes with microarrays and other technologies has become increasingly mature and affordable, which has led to rapid growth of the number and size of available transcriptomics data sets. This in turn has created new challenges for bioinformatics to provide adequate methods for data exploration and visualization, which are the topic of this dissertation.

The first research question that I address is the efficient retrieval of transcriptomics data sets from large databases, which is motivated by the rapid growth of public repositories for transcriptomics data. This makes the exploration of such repositories increasingly challenging, but also provides opportunities for biological discovery. I describe a knowledge-driven approach for exploration of transcriptomics data repositories based on ontology visualization and data-driven approaches based on gene set enrichment analysis and generative probabilistic models.

The second research question of this dissertation deals with the visualization of large transcriptomics data sets. This work has been driven by the observation that there is a growing number of data sets with hundreds or thousands of samples but a lack of suitable methods to visualize such data sets. To address this problem, I first present an analysis of the visualization tasks in this context, and then describe the design of an interactive visualization method based on pixel-oriented visualizations and tree maps. This design study is complemented with a description of the implementation of the method and a discussion of practical aspects that need to be taken into account when visualizing data sets of such scale.

This dissertation includes several case studies in which I describe the application of the proposed methods to a range of real-world data sets and discuss my findings.

Preface

This dissertation marks the end of my student life, which has been a beautiful ten year journey that took me halfway around the globe, and during which I have met many inspiring people who have shaped my view of the world.

My decision to pursue a PhD grew out of the wonderful experience I had as an undergraduate researcher in the group of Kay Nieselt in Tübingen and I'm very thankful that she gave me the opportunity to take some first steps in the big, great world of science. I'm also very indebted to my former mentors at the Institute for Systems Biology, Inyoul Lee and Daehee Hwang, who inspired me with their perseverance and work ethic and for whose support over the years I am ever so grateful.

For the last four years, I have had Alvis Brazma as my advisor to whom I am most thankful for his support and advice as well as for allowing me the freedom to develop my scientific interests and define my own area of research. I also thank the members of my thesis advisory committee, Nick Luscombe, Gos Micklem and Lars Steinmetz for their guidance and encouragement.

I'm indebted to the members of the Functional Genomics team at the EBI, who supported so many aspects of my research during last four years. I want to acknowledge in particular the help of Nikolay Kolesnikov, who worked with me on the ArrayExpress Explorer interface, Margus Lukk, who provided me with the data set that ultimately led to the development of the Space Maps visualization, Gabriella Rustici, who was so helpful with everything related to actual wet lab biology, James Malone and Tomasz Adamusiak, who answered many, many questions about the Experimental Factor Ontology, Ele Holloway, who solved a few curation mysteries for me, Misha Kapushesky, who provided me with access to the ArrayExpress Atlas data, and Helen Parkinson, who helped out whenever I had a questions about ArrayExpress, annotation or ontologies. I also want to thank my fellow PhD students and office mates Katherine Lawler

and Ângela Gonçalves for sharing their knowledge with me. In particular, I want to thank Garth Ilsley for the many conversations we had over the last four years and for joining me for some time in A3-118. Richard Coulson also deserves special mention for many insightful discussions about biology and, even more so, for fun evenings at the pub followed by mandatory late night Indian dinners.

Much of the research I undertook during the four years of my PhD was in collaboration with the group of Samuel Kaski in Helsinki. I thank Sami for making me feel so welcome during my visits, and for the great discussions we had during our meetings. Very special thanks go to José Caldas, from whom I have learned a lot, and whom I greatly respect as a person, a scientist and a colleague. Furthermore, I'm indebted to Ali Faisal, Jaakko Peltonen and Helena Aidos for being such good collaborators and co-authors. I also have fond memories of good times spent outside the lab with Leo Lathi and, in particular, with Gayle Leen.

If I had to name one single reason why it took me four instead of three and a half years to finish my PhD, I would blame it on the time spent on the organization of the "Visualization of Biological Data" workshop. Nonetheless, I'm enormously grateful for having been part of this, as well as for the opportunities that arose from it, and I'm deeply grateful to Seán O'Donoghue and Jim Procter for going through this experience with me. In particular, I want to thank Seán, and also Anne-Claude Gavin, for their help with the review on visualization of systems biology data. Of course, this would not have been possible without the support of our co-authors Nitin Baliga, Alexander Goesmann, Hiroaki Kitano, Oliver Kohlbacher, Heiko Neuweger, Reinhard Schneider, Dan Tenenbaum, and in particular Matt Hibbs, whose comments and feedback were especially helpful.

Cydney Nielsen and Miriah Meyer have been such wonderful friends and colleagues, and I appreciate our many insightful discussions about visualization in biology. I also thank Joe Parry, my local connection to the visual analytics community, for teaching me a lot about visualization in the intelligence analysis field, and about good real ale pubs in Cambridge. I'm grateful also to Bang Wong, Alan Blackwell and Roy Ruddle for conversations about visualization that shaped my views and influenced my work, as well as for coming to the EBI to speak at the "EBI Interfaces" seminars. In this context, I also want to thank Francis Rowland, Eamonn Maguire and Jenny Cham for helping me to organize these seminars and for growing a sizable visualization community on the Genome Campus.

Finally, I want to thank all of my friends, and especially my fam-

ily, in both Germany and the United States, for being so encouraging and helpful during my life as a student. And despite all the support I have received from the many people mentioned above, I could not have made it to this point without the encouragement and love of my wife, Maureen. I'm immensely grateful to her for being such an understanding and supportive partner.

Contents

List of Figures

List of Tables

List of Acronyms

AP	Average Precision
bp	Base Pairs
cDNA	Complementary DNA
CPU	Central Processing Unit
DAG	Directed Acyclic Graph
DCG	Discounted Cumulative Gain
DNA	Deoxyribonucleic Acid
dPCA	Discrete Principal Component Analysis
EBI	European Bioinformatics Institute
EFO	Experimental Factor Ontology
ES	Enrichment Score
JSON	JavaScript Object Notation
FDR	False Discovery Rate
GEO	Gene Expression Omnibus
GPU	Graphics Processing Unit
GSEA	Gene Set Enrichment Analysis
HTML	Hypertext Markup Language
IDCG	Ideal Discounted Cumulative Gain
JSON	JavaScript Object Notation
JSONP	JSON with Padding
JOGL	Java Bindings for OpenGL
kDa	Kilo Dalton
KS	Kolmogorov-Smirnov
LDA	Latent Dirichlet Allocation
MAGE-ML	Microarray Gene Expression Markup Language

MAGE-OM	Microarray Gene Expression Object Model
MAGE-TAB	Microarray Gene Expression Tabular
MDS	Multi-Dimensional Scaling
MeSH	Medical Subject Headings
MIAME	Minimum Information About a Microarray Experiment
MINiML	MIAME Notation in Markup Language
miRNA	Micro RNA
mRNA	Messenger RNA
ms	Milliseconds
NAD$^+$	Nicotinamide Adenine Dinucleotide
NCBI	National Center for Biotechnology Information
NDCG	Normalized Discounted Cumulative Gain
NeRV	Neighborhood Retrieval Visualizer
NetCDF	Network Common Data Format
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PDAC	Pancreatic Ductal Adenocarcinoma
RAM	Random Access Memory
REST	Representational State Transfer
RNA	Ribonucleic Acid
SNAP	Synaptosome-Associated Protein
SOFT	Simple Omnibus Format in Text
TF-IDF	Term Frequency - Inverse Document Frequency
tRNA	Transfer RNA
VaR	Value and Relation
VBO	Vertex Buffer Object
XML	Extensible Markup Language

Chapter 1

Introduction

The overarching theme of this dissertation is the exploration of large amounts of transcriptomics data to identify patterns and formulate hypotheses that can be followed up in more detailed analyses, either computationally or experimentally. The work presented here is motivated by the rapidly increasing amount of available high-throughput gene expression data and the challenges and opportunities arising with it.

This introductory chapter provides relevant background material about transcriptomics data, such as measurement technologies used to generate the data and explains how these data are represented, stored and accessed. Furthermore, some of the standardization efforts that enable the approaches described in the remaining chapters are considered. This is followed by a discussion of the challenges for exploration and visualization of these data that have arisen in recent years. Based on this discussion, the main research questions that will be addressed in this dissertation are introduced as well as key visualization and data exploration concepts that motivate the novel approaches introduced in this dissertation.

1.1 Transcriptomics Data

The system-wide analysis and study of gene expression on the transcript level is commonly referred to as "transcriptomics". Research in this field is a data intensive undertaking that requires an understanding not only of the underlying biology and the measurement technologies employed to study it, but also of the representation, storage and analysis of the data. These topics are the subject of the following sections.

1.1.1 Gene Expression

Proteins are the key components required for structure and function of cells in all living organisms. The amino acid sequence of proteins, and thus also most aspects of their 3-dimensional structure, is encoded in genes on the genome. The process in which a protein is produced according to a genomic gene sequence is called *gene expression*.

Gene expression is a highly dynamic process controlled by a regulatory machinery that selectively activates and deactivates the production of proteins depending on a wide range of factors, such as cell type, developmental stage and environmental conditions. The regulation of gene expression allows organisms to adapt to their environment and to create specialized tissues and organs from just a single genome sequence.

In the interest of space and due to the focus of this dissertation, the following sections discuss only the expression of protein coding genes in eukaryotes. While the gene expression systems of eukaryotes and prokaryotes are similar, they also have some significant differences. Furthermore, there are differences between the expression of protein coding and non-coding genes, such as microRNA (miRNA; Bartel, 2004) and transfer RNA (tRNA) genes.

1.1.1.1 Transcription

Gene expression consists of a series of tightly regulated steps, the first of which is *transcription*, the production of a *transcript* according to the gene sequence. The transcript is a messenger RNA (mRNA) molecule produced by an *RNA polymerase* with the help of several other proteins that mediate polymerase binding and initiation of the transcription process. The generated mRNA sequence is a copy of the DNA sequence of the gene.

Transcription is regulated by a range of mechanisms. The simplest and most direct of which is the activation or repression of transcription by binding of so-called *transcription factors*. Transcription factors are proteins that contain a DNA-binding domain and bind to specific sequences in the *promoter* of a gene and modulate the activity of the RNA polymerase. A repressor transcription factor may prevent the RNA polymerase or other required proteins from binding, while an activator may increase the chances of the RNA polymerase binding to the DNA and initiating the transcription process (Lee and Young, 2000).

Another mechanism by which transcription of genes can be controlled is through changes in the structure of the DNA. In the nucleus, DNA is wrapped around complexes of *histones*, forming so-called *nucleosomes*, which are the basic building blocks of the *chromatin structure*. Strands of chromatin give rise to the supercoiled structure of the chromosomes. Where the genomic DNA is densely packed, the transcriptional machinery is less likely to bind and initiate the transcription process. However, opening of the chromatin structure can increase the rate of transcription and this process is controlled by a variety of modifications of the histone complexes, such as methylation as well as acetylation and deacetylation (Margueron et al., 2005). More recently it has been shown that not only changes in the chromatin structure, but also changes in the 3-dimensional structure of chromosomes influence transcription levels (Dekker, 2008).

Once the gene sequence has been transcribed, the mRNA undergoes several post-transcriptional modifications in a process called *matura-*

tion. Only the nucleotide sequence of *mature* mRNA is translated into the amino acid sequence of the protein. Removal of intronic sequences from the transcript is a significant step during the maturation of the mRNA. This process is performed by the *spliceosome*, a complex of several proteins that detects the splice sites, cuts out the introns and then ligates the ends of the adjacent exons. The selective removal of exons, which is called *alternative splicing* and creates so-called *transcript isoforms*, is another hallmark of the splicing process and part of the regulatory system. *Polyadenylation*, the process in which a short sequence of adenines is attached to the 3′ end of the transcript is another step in the maturation of the transcript. Polyadenylation is required for export of the mRNA from the nucleus for translation and also for mRNA stability (Coller et al., 1998), which also has an impact on transcript levels. Another post-transcriptional modification that protects the mRNA from degradation is the attachment of a 7-methylguanosine cap to the 5′ end of the transcript (Shatkin, 1976).

1.1.1.2 Translation

Once an mRNA transcript has been fully processed, it is exported from the nucleus into the cytoplasm where the mRNA sequence is translated into a polypeptide with a corresponding amino acid sequence. The translation process is performed by ribosomes, which recruit tRNA molecules that transfer the required amino acids to the growing polypeptide. The polypeptide chain is imported into the endoplasmic reticulum or cytosol and eventually folded into its three-dimensional structure. Ultimately, the protein is exported into the Golgi apparatus, where it receives post-translational modifications (Walsh et al., 2005) and is then transported to other destinations in the cell.

Several mechanisms are known to control gene expression after transcription and during translation. A major class of post-transcriptional regulators are short non-coding RNA molecules, such as miRNAs. MiRNAs are around 22 base pairs (bp) long single stranded RNA molecules that bind to specific sites in mature mRNA in the cytoplasm. They either

promote degradation of the mRNA or prevent it from being translated (Eulalio et al., 2009). Other mechanisms that interact with the translational machinery affect either the translation of all transcripts through modification of general translation initiation factors or they affect only the translation of particular mRNA transcripts through binding of specific proteins to regulatory sites in the untranslated regions of the mRNA. These mechanism have been reviewed in detail for instance by Lackner and Bähler (2008). Since they do not affect measurements of transcript levels, they are not discussed further and mentioned here only for completeness.

1.1.2 Measurement Technologies

The relative quantification of mRNA transcripts in a sample has been possible since the mid-1970s with blotting methods originally developed by Southern (1975) and later also through reverse-transcription, quantitative real-time PCR (Polymerase Chain Reaction; reviewed by Van Guilder et al., 2008). However, the field of transcriptomics was only established in the mid-1990s with the invention of high-density microarray platforms. For the first time, it became possible to measure the relative abundance of thousands of transcripts simultaneously to gain a global view on the transcriptional state of an organism.

1.1.2.1 DNA Microarrays

Microarrays are a highly parallel measurement technology that is based on the hybridization of transcripts from a biological sample to complementary DNA oligonucleotides located on a solid surface. The oligonucleotides located on the array are so-called *probes*, which are arranged into a spatially addressable array of *features*. Each feature consists of a very large number of identical probes, whose sequence is associated with and unique for a particular gene. In some cases there are multiple features for a single gene.

Even though a wide range of different microarray platforms are

available today, the general principle applied to study gene expression with this technology has remained essentially the same: Either total or poly-adenylated mRNA is extracted from a biological sample, such as a tissue biopsy or a cell culture. After purification, the mRNA is reverse transcribed in the presence of nucleotides that can be linked to fluorescent dyes. The resulting fluorescent complementary DNA (cDNA) molecules are then hybridized to the sequences located on the microarray. Following washing steps to remove unbound cDNA, the microarray is scanned while a laser excites the fluorescent dye associated with the cDNA. With the help of image processing software the brightness of each feature can be quantified and used to derive a relative abundance measurement for the transcripts of each gene that is represented by a feature on the array.

The key assumption is that the amount of fluorescence measured for a feature correlates with the amount of bound cDNA and that the amount of cDNA correlates with the abundance of transcripts in the original sample. For instance, Park et al. (2004) performed a comprehensive study that compared different microarray platforms with quantitative real-time PCR and showed that this assumption holds for most common microarray platforms. Nonetheless, the measurements obtained from microarrays are affected by several sources of error, such as cross-hybridization, i.e. binding of cDNAs to probes that they are not supposed to bind to, different levels of dye incorporation for different cDNA sequences or the formation of secondary structures in cDNA or probe sequences that prevent complementary binding.

1.1.2.2 RNA Sequencing

Serial Analysis of Gene Expression (SAGE, Velculescu et al., 1995) is an early sequencing-based technology to measure the abundance of mRNA transcripts. The key idea behind SAGE is that short sequence tags excised from mRNA transcripts are sufficient for the identification of the source gene and that large numbers of tags can be sequenced efficiently by concatenating them. Ultimately, SAGE yields quantitative transcript counts.

While in some cases this is an advantage over microarray technologies, application of SAGE is limited for instance by the large amount of mRNA required Ye et al. (2002).

With the advent of next-generation sequencing technologies a new sequencing-based approach to measure transcript abundance has become possible. This approach, called *RNA-seq*, is based on the deep-sequencing of transcripts from biological samples (Wang et al., 2009). In general, this approach implies the extraction and purification of mRNA from a sample, followed by the deep-sequencing of the sequences, which yields reads of a length between 30 and 400 bp, depending on the sequencing technology. The reads obtained from the sequencing step are then typically aligned to an annotated reference genome, where the location of genes and their introns and exons is mostly known.

Several approaches have been devised to derive gene expression levels from mapped sequencing reads. One option is to infer expression levels from the total number of reads that fall into a gene, normalized by the length of the gene. This and related approaches requires that all reads map to a unique location in the genome. Reads that map to multiple locations either have to be removed from the data set or assigned to a single location. If reads map to a known splice junction, similar precautions are necessary in order to correctly quantify transcript isoforms.

However, the technology and data processing techniques are still in their infancy and some issues have been identified recently. Biases are introduced, for instance, by the PCR amplification required for the construction of sequence libraries used by some next-generation sequencing platforms (Kozarewa et al., 2009; Mamanova et al., 2010). Additional complications exist on the data analysis level. For instance, Degner et al. (2009) report that the mapping rate of reads to the genome can be influenced by single-nucleotide polymorphisms in the genome sequence, which is a problem for the detection of allele-specific expression differences. Another issue is that with RNA-seq technologies genes with longer transcripts inherently are more likely to be called differentially expressed due to the

fragmentation required to cover the length of the transcript with reads (Oshlack and Wakefield, 2009).

Even though next-generation sequencing technologies show great promise for transcriptomics applications, it will take some time until investigators will be able to use them to generate reliable and biologically meaningful results on a large scale. However, it is expected that in the long term sequencing-based transcriptomics technologies will take over the current role of microarray-based technologies. This is primarily due to their increased flexibility, for instance when measuring transcript isoforms, and their ability to determine transcript abundances at a higher resolution than typically possible with microarrays.

1.1.3 Experimental Design

Depending on the biological question that is being studied, different experimental designs can be applied in transcriptomics studies. There are numerous issues relating to the power of statistical tests and the employed measurement technologies that need to be taken into account when an experiment is designed. Many of these issues are discussed in Kerr and Churchill (2001), Yang and Speed (2002), Churchill (2002) and Townsend (2003). However, since the work presented in this dissertation is based on the assumption that appropriate design decisions were made to address statistical issues, here only those aspects of the experimental design are considered that are relevant to the scientific question, such as "What is the transcriptional response after a treatment of cell line X with drug Y?" Thus the experimental designs are described independently of the measurement technologies introduced in Section 1.1.2, which in practice have a significant influence on the experimental design.

It is important to point out that in order to reduce the variation in the data in almost every transcriptomics study several replicates are analyzed for each biological state that is being investigated (Allison et al., 2006). Both biological and technical replicates are commonly used. Where this dissertation mentions a "sample", the set of replicates for a corre-

sponding biological state or a summary derived from measurements of these replicates is referred to, unless otherwise noted.

The transcript levels corresponding to the biological states that are measured in a transcriptomics study are influenced by a number of factors. Such factors are for instance the organ from which the sample was taken (the factor is "organ"), the time that has passed after a treatment (factor "time") or the gene that was knocked out (factor "genotype"). In studies in which only a single such factor is suspected to be relevant for the observed phenotype, two major classes of experimental designs can be applied. These depend on whether there is an intrinsic ordering of the studied samples or not:

Ordered Samples Studies that fall into this class are, for instance, time-series and dose-response studies as well as studies of organism or organ development. Time-series data are often collected to study the cellular response to a chemical or physical stimulus. For instance, chemical stimuli are treatments with drugs or other chemical agents, while physical stimuli may be heat shocks or injuries through an external force. When developmental processes are studied, samples are taken at various stages during the development of an organism. It can be assumed that due to ordering of the samples in this class of experimental designs a *reference sample* exists. Usually this is the first sample in the series, e.g. time point 0, against which all other samples can be compared to obtain relative transcript expression levels.

Unordered Samples Two subclasses comprise this class of experimental designs: those that include a *reference sample* and those that do not. The latter case is common, for instance in studies that aim to create an "atlas" of gene expression across a range of tissues of an organism, as described for example by Su et al. (2002). In practice an artificial reference sample is sometimes created, such as by pooling samples all used in the study (Townsend, 2003). Experimental designs with unordered samples that

include a reference sample are essentially binary comparisons between a control and one or more phenotypes of interest, such as studies that involve series of gene knockouts that can be compared to a wild-type organism (Hughes et al., 2000) or studies that compare healthy tissue with tumor tissue.

If more than one factor is considered to be relevant in a study "multi-factorial" designs are used in which every sample is associated with two or more factors. For individual factors in these designs the aforementioned classes still apply. An example of such a multi-factorial design is a study by Hwang et al. (2009), in which the effect of infection with different prion strains on temporal expression patterns in mouse strains was investigated.

The key difference between the experimental designs described above is the way in which the generated data are interpreted. This has an influence on the selection of suitable analysis and visualization methods and is important for the work presented in Chapters 2, 3 and 4.

1.1.4 Data Representation

Figure 1.1 illustrates the key steps in the acquisition and preparation of microarray data for further analysis. The figure also explains the representation of the experimental design as meta information.

Quantified transcript levels are usually represented as floating point values. Depending on the measurement technology used, the value can be either only positive or both positive and negative. The latter case occurs when transcript levels are *log*-transformed ratios of transcript levels from two conditions that are being compared. The *log*-transformation is applied to obtain a distribution around 0. The *transcript* or *gene expression values* collected in a single *study* are typically represented in form of an *expression matrix*, in which the rows correspond to genes and the columns correspond to the biological states that are being investigated in the study.

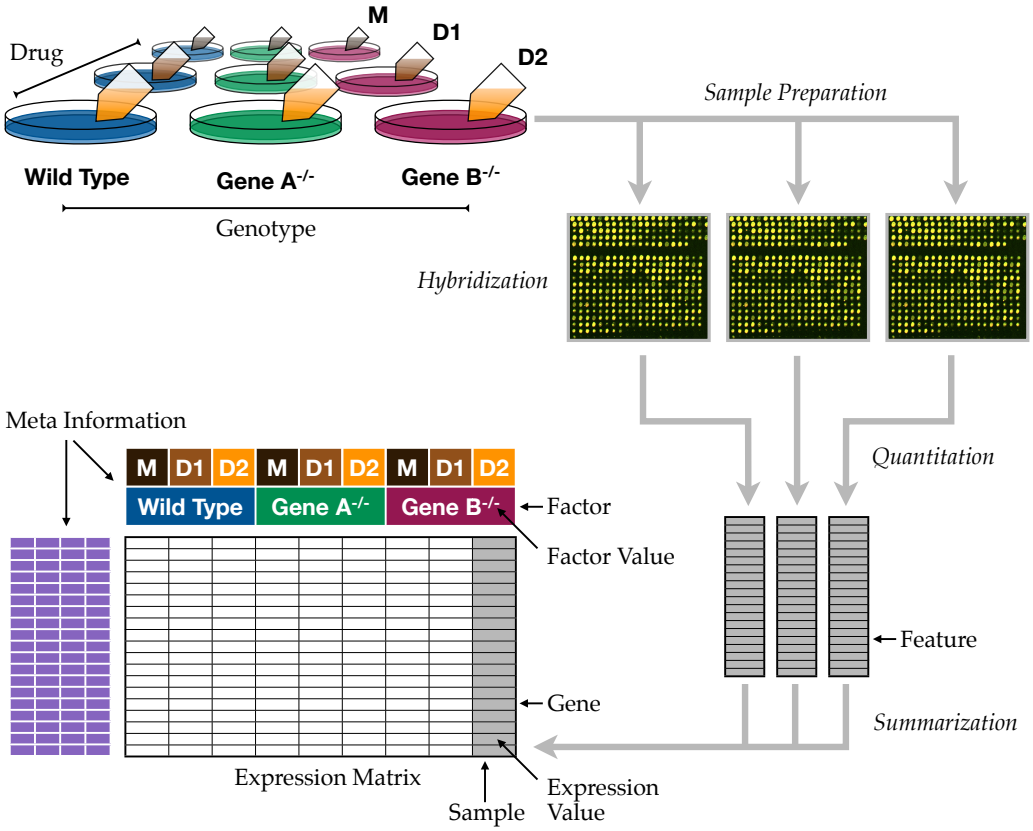

Figure 1.1: Acquisition and representation of transcriptomics data using microarrays. The experimental design shown in the example contains two factors, *Drug* (with values *M(ock)*, *D1*, *D2*) and *Genotype* (*Wild Type*, *Gene* $A^{-/-}$, *Gene* $B^{-/-}$). Combinations of the drug and genotype instances represent the biological states that are being investigated in this example. For each biological state samples are obtained and prepared for hybridization on microarrays. After hybridization the microarrays are scanned and image processing techniques are used to quantify the relative amount of transcript sequences for each feature. In the next step the feature measurements corresponding to each gene are summarized within each array and summaries are computed across replicates. Meta information associated with the columns of the expression matrix reflects the factors and factor values used in the experimental design. Meta information associated with the rows of the matrix provides further details about the genes for which transcript levels were measured. *(Adapted from Causton et al., 2003, Chapter 1; Microarray image source: smd.stanford.edu)*

Thus, the row vectors are also referred to as *gene expression profiles*, and the column vectors of the matrix are so-called *sample expression profiles*.

In order to yield reliable results in subsequent analysis steps, the expression values in the expression matrix have to be *normalized*. Both

within and between array normalization methods are used to remove systematic noise and other biases. Common methods for normalization are, for instance, quantile normalization (Bolstad et al., 2003) and variance-stabilizing normalization (Huber et al., 2002). In particular, in earlier studies missing values in the matrix are not uncommon due to problems in the hybridization and quantification steps. Several approaches are available to deal with missing values, either by removal of affected rows and columns or by imputation of the missing values (e.g. Troyanskaya et al., 2001; Oba et al., 2003).

The transcript levels obtained after these processing steps are only indirectly indicative of the amount of transcripts that were present in the biological sample. Transcript levels are usually only interpreted relative to other transcript levels from the same study. Additionally, the values are comparable on a quantitative level only if acquired on the same microarray platform (Shi et al., 2006).

Both rows and columns of the expression matrix are annotated with *meta information* about the studied genes and samples. This meta information is essentially a vector of attributes for every gene and another vector of attributes for every sample. In the case of genes, this information often includes terms from the Gene Ontology (Ashburner et al., 2000), chromosomal locations, database identifiers or sequences used for the corresponding probes on the microarray. The description of the samples consists primarily of labels for each of the *factors* that describe the biological state. These labels are called *factor values* and can be of any type, i.e. nominal, ordinal, interval or ratio values. More information about the meta information associated with the expression matrix and how it is used in the context of data visualization is provided in Section 4.2.1.

Both meta information and details about how the study was performed are crucial for the interpretation of the data. Efforts to standardize this information are described below.

1.1.4.1 Standards

The availability of large collections of data sets with standardized, expert curated meta information lays the foundation for this dissertation, in particular for the approaches presented in Chapters 2 and 3.

Early in the development of the transcriptomics field a need to standardize the description of microarray experiments was observed. Published microarray data usually lacked a sufficiently detailed description to allow their interpretation and verification by independent researchers who had not been involved in the original study. Driven by this observation the "Minimum Information About a Microarray Experiment" (MIAME) guidelines were developed as an agreement about the content and structure of the description of a microarray study (Brazma et al., 2001) and have since become a de facto standard.

A major challenge in the description of gene expression studies is the enormous number of possible states of the cell, as described in Section 1.1.1. Furthermore, the situation is complicated by the use of a wide range of array platforms and protocols, both for wet lab work and data processing.

The MIAME standard addresses the complexity of a microarray experiment by dividing the description of a study into six parts. The *experimental design* part describes the relationships between the hybridizations performed in the study as well as the type of experiment, for example: time-series, dose-response and normal-versus-disease. Furthermore, this part contains a description of the experimental factors and the relationship between the biological samples and the microarrays used in the study. These are described in more detail in the *array design* part, which contains information about the sequences of probes used, the transcript they have been designed for and their physical arrangement on the microarray slide. The third part of the MIAME guidelines addresses *samples*, in particular their source, the treatments that they have undergone before or after they were taken and the methods used to prepare the samples for hybridization on the array. This step of the experiment and the protocols used are de-

scribed in detail in the *hybridization* part of the MIAME standard. The fifth part describes the *measurements*, which are the result of the experiment and consist of the scanned images, quantifications based on the images and the normalized gene expression matrix obtained after combining quantifications of multiple replicates. Finally, the *normalization controls* part specifies relevant parameters for normalization of the measurements and the algorithms used.

The practical implications of experimental descriptions that are compliant with the MIAME guidelines are manifold. For instance, MIAME enables the creation of standardized microarray data repositories (see Section 1.1.5) and file formats that represent the information in a machine-readable format for storage and data exchange purposes. It has to be noted that whereas the MIAME guidelines specify the content and structure of the experiment description, they do not provide a file format for data storage and exchange. However, an important recommendation of the MIAME guidelines is the use of controlled vocabularies to describe many aspects of the experiments (see Section 1.1.4.3).

1.1.4.2 File Formats

To date, several file formats have been developed that provide a MIAME-compliant representation of microarray gene expression data. MAGE-ML (Microarray Gene Expression Markup Language) by Spellman et al. (2002) is described in XML and is designed after the Microarray Gene Expression Object Model (MAGE-OM). It was later found that this format is too complex for a large portion of its intended audience (Brazma, 2009). With MAGE-TAB (Microarray Gene Expression Tabular) Rayner et al. (2006) proposed a simpler format that has been developed to address this shortcoming. MAGE-TAB is a tab-delimited file format that defines a series of description and data file types that can be edited with common spreadsheet software but it is still structured, machine-readable and formally defined based on the MAGE-OM.

Two other notable formats are SOFT (Simple Omnibus Format in

Text) by Edgar et al. (2002) and MINiML (MIAME in Markup Language, www.ncbi.nlm.nih.gov/geo/info/MINiML.html). While these formats support MIAME-compliant representation of the experiment description, this is achieved through free text summaries, rather than through the use of controlled vocabularies. This characteristic makes it hard to automatically and reliably extract key information about an experiment from SOFT and MINiML files, which, to some degree, hinders their use in large-scale meta analysis and data mining approaches (Rayner et al., 2006). Despite the challenges, several groups have used text-mining approaches followed by expert curation to extract phenotype and contextual information from free text descriptions (Butte and Kohane, 2006; Lukk et al., 2010).

Overall, the widespread use of the MIAME-compliant data representations has a positive effect on the reproducibility of scientific results, even though there are still limitations, as found in a recent study by Ioannidis et al. (2009).

1.1.4.3 Ontologies

As pointed out in the previous section, the use of controlled vocabularies to describe experiments is an important aspect of the MIAME guidelines. For instance, it greatly simplifies automated integration of data sets that were generated independently and by different labs. Controlled vocabularies also enable more powerful queries in transcriptomics repositories (see Chapter 2).

Ontologies are a typical source of controlled vocabularies and add a further layer of structure by introducing ontology classes and relationships between them. They also allow computational inference based on the relationships defined by the ontology (Malone et al., 2010).

The best known ontology in the biomedical field is probably the Gene Ontology (Ashburner et al., 2000), which is typically used to annotate genes and their products. However, in the context of transcriptomics data it is also important to have well-defined descriptions of the cellular states and phenotypes that have been studied. For this purpose, several

ontologies have been developed over the last couple of years to model different facets of the biomedical domain. Some examples are the Cell Ontology (Bard et al., 2005), the Disease Ontology (Dyck and Chisholm, 2003) and the Plant Ontology (Avraham et al., 2008). The Experimental Factor Ontology (EFO) is an application ontology specifically developed to describe various aspects of transcriptomics experiments. This ontology combines ontology classes from parts of existing ontologies and references back to the original classes, which enables interoperability with those ontologies and other resources referring to them (Malone et al., 2010).

1.1.5 Repositories

Standardized representation allows high-throughput automated mining of the data (Brazma et al., 2001) but also enables the creation of repositories for transcriptomics data and the exchange of data between public archives.

1.1.5.1 ArrayExpress

The ArrayExpress Archive (www.ebi.ac.uk/arrayexpress) at the European Bioinformatics Institute (EBI) is the archive component of ArrayExpress (Parkinson et al., 2009) and at the end of July 2010 contained data from over 12,900 studies. About a third of these studies were submitted directly to ArrayExpress, while the rest is imported from other repositories, mostly from the Gene Expression Omnibus (GEO; see Section 1.1.5.2). The ArrayExpress Archive is the second largest repository for transcriptomics data after GEO operated by the National Center for Biotechnology Information (NCBI). With one exception, all transcriptomics data used in this dissertation was obtained from ArrayExpress.

The ArrayExpress Archive contains data from functional genomics studies, primarily gene expression data generated on microarrays, but also next generation RNA sequencing data and genotyping data. In the case of transcriptomics data, both raw data and processed data are stored in the repository, if available. The content and format of the raw data are depen-

dent on the platform used to perform the measurements. Essentially, the raw data represent quantified fluorescence signals for each feature on the microarray as well as quantifications of background fluorescence signals and other information needed to derive transcript levels. Processed data is stored in the form of a data matrix containing data that usually have been normalized and prepared for further analysis. For every study MIAME-compliant descriptions are available in MAGE-TAB or MAGE-ML format.

Data sets are either submitted directly by the authors of a study or obtained from other transcriptomics databases through data import pipelines, e.g. from GEO or the Stanford Microarray Database (Hubble et al., 2009). Whereas the latter is mostly automated, the former involves manual checking and processing of the submissions by expert curators to ensure that the annotation of the data sets is complete and of sufficient quality.

The data sets in ArrayExpress can be accessed through a range of different interfaces. The primary user interface is a web-based front-end which supports keyword queries against the meta data of the studies as well as a direct access to data sets whose accession number is known, for instance from an associated publication that reported on the results of the study. The web interface offers comprehensive descriptions of the studies as well as links to download the actual data files. A web service provides programmatic access to the query functionalities of the repository front-end.

Besides downloads through the web interface, the content of the ArrayExpress Archive can be retrieved through an R/Bioconductor package (Gentleman et al., 2005; Kauffmann et al., 2009) and for bulk downloads all data sets are stored on an FTP server in MAGE-ML and MAGE-TAB formats (Parkinson et al., 2007).

1.1.5.2 Gene Expression Omnibus

The National Center for Biotechnology Information (NCBI) is operating the Gene Expression Omnibus (GEO; www.ncbi.nlm.nih.gov/geo) archive

for high-throughput functional genomics data (Barrett et al., 2009). GEO is the largest public repository for transcriptomics data and at the end of July 2010 contained data from over 17,900 studies.

The core features of GEO are very similar to those of ArrayExpress. The GEO repository can be queried for data sets using keywords through a web-based user interface. Like in ArrayExpress, the data sets can be queried from, and imported directly into, R/Bioconductor (Sean and Meltzer, 2007) and are also available for bulk download from an FTP server. GEO can store data sets in a format that is fully MIAME-compliant, but also accepts submissions that themselves do not contain all information required by the MIAME guidelines (Edgar and Barrett, 2006). Unlike ArrayExpress, GEO uses SOFT and MINiML as primary data formats for submission and storage (see Section 1.1.4.2).

A distinguishing feature of the GEO repository is availability of curated, so-called *DataSet*s (Barrett et al., 2007), which are data structures obtained from the original submissions. Several visualization and basic data exploration tools are provided for these *DataSet*s in the web interface, such as computation of differentially expressed genes between two factor values, as well as visualization of expression matrices as clustered heatmaps and visualization of transcript or feature expression profiles. The latter are also available in ArrayExpress, however as part of the derived ArrayExpress Atlas database, which also provides information about differentially expressed genes (see Section 1.1.5.4).

1.1.5.3 Other Repositories

Besides ArrayExpress and GEO, several smaller repositories exist that offer similar features and, in many cases, are MIAME-compliant, but have a more focussed purpose. These repositories are often organism-specific and cater to the needs of particular scientific communities, such as The Arabidopsis Information Resource (TAIR; Swarbreck et al., 2008) and NASCarrays (Craigon et al., 2004) for Arabidopsis. Other repositories contain features that are not available in GEO or ArrayExpress, such a inte-

gration with the GenePattern transcriptomics data analysis pipeline (Reich et al., 2006) in the Stanford Microarray Database (Hubble et al., 2009).

1.1.5.4 Derived Databases

A key aspect of the MIAME guidelines is the explicit recommendation for the use of controlled vocabularies. The use of these controlled vocabularies and the availability of large collections of well annotated data sets in aforementioned repositories, primarily in ArrayExpress, enables the automated analysis of large collections of data sets and their application in the construction of derived databases (Brazma, 2009).

In the last couple of years, several value-added secondary databases have been built based on the publicly available data. Typically, the goal is to use expression profiles to associate genes with particular conditions, such as disease states, developmental stages, genetic modifications or drug treatments, and vice versa. Conditions from different studies are usually summarized through the use of ontologies or other classifications that are often hierarchically organized. The general assumption is that combining data from a large number of studies will produce more reliable results and provide a broader picture. The meta analysis results provided by these databases fall into one of two categories: either the data from the original studies are reanalyzed for each study individually, or the data are combined and treated like a single study.

Major value-added databases that fall into the former category include Oncomine (www.oncomine.org) by Rhodes et al. (2007), Genevestigator (www.genevestigator.com) by Hruz et al. (2008) and the ArrayExpress Atlas (www.ebi.ac.uk/gxa) by Kapushesky et al. (2010). The latter is derived from a curated subset of the ArrayExpress Archive and contains information about differentially expressed genes across a wide range of experimental conditions such as specific tissues and diseases.

Databases that treat the data like a data set collected in a single study are for example the Human Gene Expression Map by Lukk et al. (2010) and GeneSapiens (www.genesapiens.org) by Kilpinen et al.

(2008), which provide access to meta analysis results of data collected on Affymetrix GeneChips for human samples. Since the approach taken by these databases is based on the assumption that expression levels measured in different studies can be directly compared, they are generally limited to well-defined subsets of the publicly available data, such as particular microarray platforms.

1.2 Challenges and Opportunities

During the last decade, the transcriptomics field has developed into a mature domain of biomedical research. This is evident in the wide range of infrastructure that supports the generation, analysis and storage of transcriptomics data, as described in the previous sections. At the same time high-throughput analysis of transcriptomes with microarrays and other technologies has become increasingly affordable, which has also contributed to rapid growth of the number and size of available transcriptomics data sets.

Figure 1.2 illustrates the growth of the ArrayExpress Archive from 2003 to the end of 2009 in terms of submitted studies and hybridizations experiments as part of these studies. The number of publicly available data sets has grown continuously and has increased by at least two orders of magnitude since 2003.

The major opportunity arising from access to large numbers of transcriptomics data sets with MIAME-compliant descriptions is the possibility to automatically integrate these data sets and analyze them together to discover patterns that are not detectable in individual data sets. Such patterns are for instance sets of genes that behave very similarly under a wide range of conditions, which may be used to assign functions to uncharacterized genes (e.g. Hibbs et al., 2007). Similarly, large collections of transcriptomics data sets can be analyzed to identify gene activation patterns shared by multiple diseases and to suggest possible drugs to treat these diseases (e.g. Suthram et al., 2010). Another approach is to combine

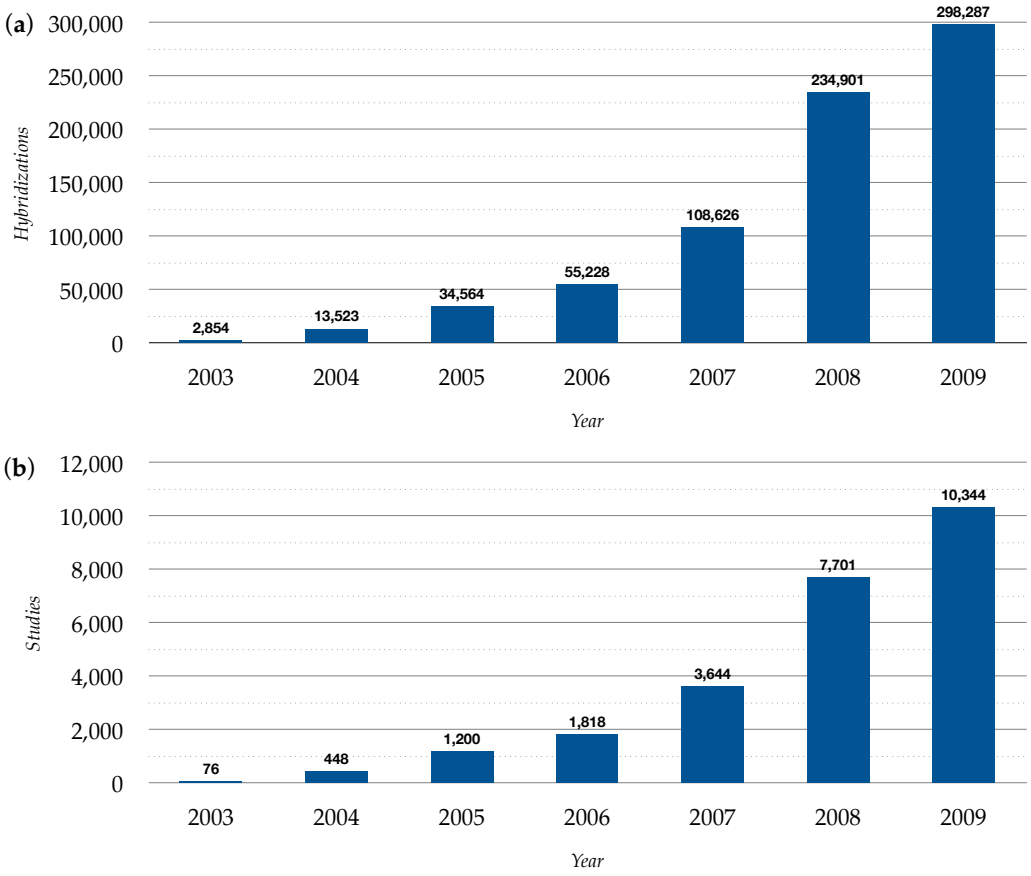

Figure 1.2: Growth of the ArrayExpress Archive. The height of the bars represents the total number of (**a**) hybridizations and (**b**) studies for which data has been deposited in the ArrayExpress Repository by the end of the indicated year. The large increase between 2007 and 2008 can at least partly be explained by the implementation of a data import pipeline from GEO into the ArrayExpress Archive. *(Source: www.ebi.ac.uk/microarray/doc/stats)*

data sets obtained from the same measurement platform and treat the data like a single very large data set that can be used to identify major gene expression patterns in a global expression space that contains a very wide range of different conditions (e.g. Lukk et al., 2010).

While the increased scale of data sets is creating opportunities by providing a more complete picture of gene expression, exactly this increased scale is also a key challenge for exploration of these data. In data exploration, or *exploratory data analysis* (Tukey, 1977), the main goal is to

generate hypotheses based on patterns observed in the data. Such patterns can be identified computationally, for instance with the help of unsupervised classification methods from data mining or machine learning.

1.2.1 Information Visualization

A complementary approach to computational pattern discovery with clustering or association rule learning methods is to visualize the abstract data through transformation into an appropriate graphical representation. When data is represented graphically the human visual system is used for pattern detection. Visualization has long been an important tool in exploratory data analysis because it enables unbiased study of the data without making assumptions about their structure or underlying models (NIST, 2010, see Chapter 1). Many methods from exploratory data analysis such as box-and-whisker plots (McGill et al., 1978), scatter plots and histograms are well known and frequently used. A convincing example, which demonstrates the importance of visualization in data analysis, is a collection of four small data sets with identical summary statistics (mean, variance, correlation, linear regression), known as *Anscombe's Quartet* (Anscombe, 1973). The data sets are shown in Table 1.1 and the corresponding visualizations are shown in Figure 1.3. The visualization of these data sets illustrates that despite their apparent statistical similarity, the data sets are actually quite different.

1.2.1.1 Definition

Card et al. (1999) define information visualization as "the use of computer-supported, interactive, visual representations of abstract[1] data to amplify cognition." According to this definition the goal of visualization is to support humans in using or acquiring knowledge. Visualizations achieve this,

[1]The visualization community distinguishes between *information visualization* and *scientific visualization*. As indicated by the definition, information visualization deals with *abstract* data whereas scientific visualization deals with *physical* data that is inherently geometric, such as tomographic scans of the human body or protein structures.

Table 1.1: Anscombe's Quartet (Anscombe, 1973). In each of the four data sets mean $\mu_{X_i} = 9.0$, variance $\sigma^2_{X_i} = 11.0$, $\mu_{Y_i} = 7.5$, $\sigma^2_{Y_i} = 4.12$, correlation $cor(X_i, Y_i) = 0.816$ and the linear regression line is $Y_i = 3 + 0.5X_i$ for $i \in \{1, 2, 3, 4\}$.

X_1	Y_1	X_2	Y_2	X_3	Y_3	X_4	Y_4
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

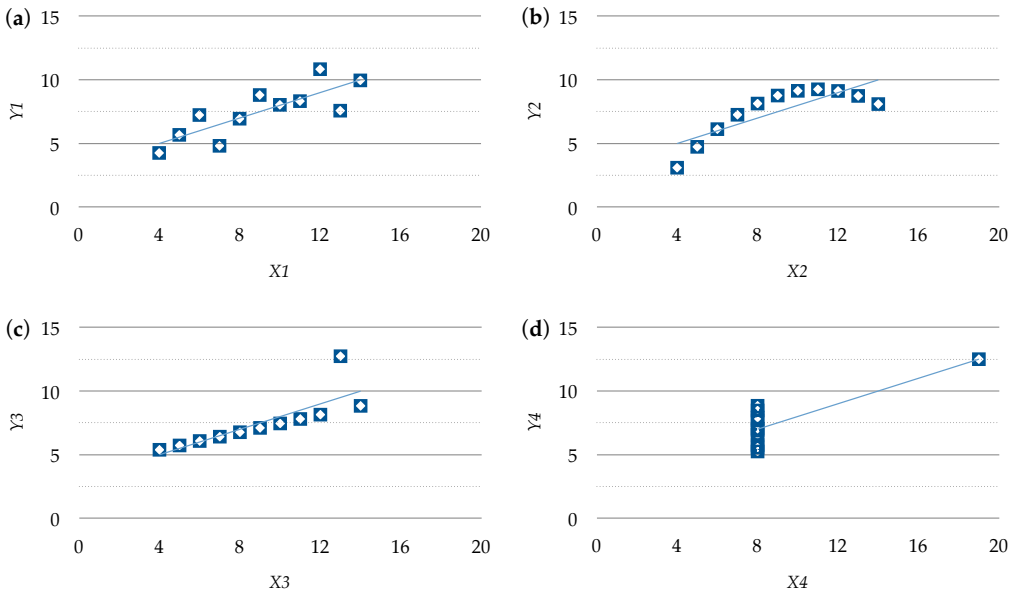

Figure 1.3: Plots of the data sets in Anscombe's Quartet corresponding to the data shown in Table 1.1. The equation of the regression line in (**a**), (**b**), (**c**) and (**d**) is $Y_i = 3 + 0.5X_i$ for $i \in \{1, 2, 3, 4\}$.

for instance, by providing increased cognitive resources, reducing the time required to search for information, enhancing the recognition of patterns and enabling perceptual inference (Fekete et al., 2008). In practice, this is accomplished through the use of graphical representations that organize information, highlight its key characteristics, support the identification of patterns, trends and outliers and allow for visual comparisons (Hearst, 2009, see Chapter 10).

To emphasizes the need to consider how human perception and cognition work in order to design useful visualizations Card et al. (1999) have adapted Richard Hamming's famous quote

The purpose of computing is insight, not numbers.

for the visualization field as

The purpose of visualization is insight, not pictures.

According to Ware (2004, see Chapter 1), the psychological foundations of visualization can be derived primarily from two perceptual properties: *Preattentive processing* and the principles of *Gestalt theory*.

1.2.1.2 Preattentive Processing

Visual properties that can be perceived in less than 250 milliseconds (ms) are processed preattentively (Healey et al., 1993) and do not require sequential scanning of the visualization. Two examples and one counter example of preattentive processing of colors and shapes are shown in Figure 1.4, but several further visual properties are known to be processed preattentively as well, such as size (Healey and Enns, 1999). Fundamental to the role of preattentive processing in information visualization is that it is independent of the number of objects being displayed. This fact can be exploited for the design visualizations that contain very large amounts of data and can be evaluated rapidly.

Figure 1.4: Preattentive processing of colors and shapes after Healey et al. (1993). The presence of the blue square in (**a**) and the presence of the red circle in (**b**) can be recognized preattentively, whereas the red circle in (**c**) cannot, because color and shape distinction does not take place simultaneously in the visual system.

1.2.1.3 Gestalt Theory

Gestalt theory was introduced by German psychologists in the early 20th century and is based on the observation that visual stimuli are perceived in a way that results in structures that are as simple as possible. Based on this observation, Gestalt psychologists have derived a set of laws[2] that explain some key aspects of human image understanding. Most relevant for information visualization are the *law of proximity*, stating that objects that are close together are perceived as a group, the *law of similarity*, stating that similar-looking objects are perceived as forming a group, the *law of continuity*, stating that objects are grouped when a straight or smooth line is potentially connecting them and the *law of closure*, stating that contours that are not closed will be perceived as being closed and segmenting the space into an inner and outer region. Further laws address symmetry, size, common fate of moving objects, and familiarity of shapes (see Chapter 8; Tovée, 1996). Figure 1.5 provides examples for some of these laws.

In practice, the laws of Gestalt theory can be applied in the design of visualization methods. For example, when sets of related objects are placed close to each other they will be perceived as clusters (law of proximity; applies to scatter plots, e.g. see Section 1.2.2.1), or when lines overlap they will still be perceived as continuous lines (law of continuity; applies to profile plots, e.g. see Section 1.2.2.2). When objects are placed within

[2]Sometimes also called "principles" or "rules".

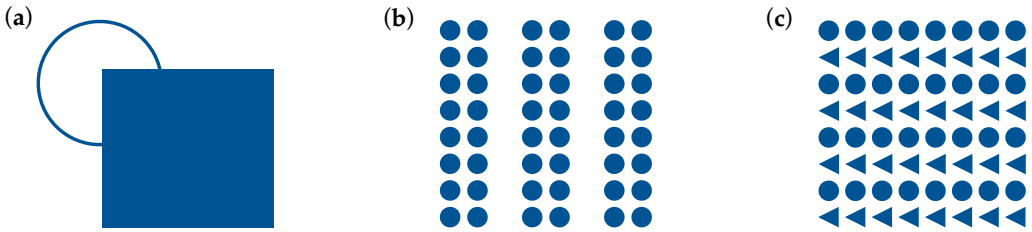

Figure 1.5: Illustrations of key laws of Gestalt theory after Ware (2004). (**a**) *Law of closure*: The structure is perceived as a rectangle covering a circle rather than as a rectangle and a broken ring. (**b**) *Law of proximity*: Three groups of dots are perceived rather than individual dots. (**c**) *Law of similarity*: The objects are perceived as rows of circles and triangles, rather than as columns.

the boundaries of another object, they will be perceived as being parts of that object (law of closure; applies to tree maps, e.g. see Chapter 2).

1.2.1.4 Visual Encoding

Visual comparison of data is an important goal of visualization. Many different possibilities exist to encode data, but not all are appropriate to support quantitative comparisons (Mackinlay, 1986). Bertin (1983) defined a vocabulary of visual encodings for data that consists of three distinct categories that are used in combination: *marks*, *retinal variables* and *position*. While *position* refers simply to the coding of data by placing objects in distinct locations in space, *marks* can either be lines, points and areas. *Retinal variables* consist of size, shape, orientation, color, gray scale and texture. These properties and their combinations form the core repertoire for visual encoding of data and several studies have examined how they can be applied effectively (e.g. Cleveland and McGill, 1984). One important finding from these studies is that relative comparison of quantities work much better when the quantities are encoded as the length of a vertical line, such as in a bar chart, rather than in the size of an area, such as in a pie chart (Spence and Lewandowsky, 1991).

1.2.1.5 Interaction Techniques

One very important aspect of information visualization is interaction, as highlighted in the definition provided by Card et al. (1999). This applies particularly to visualization methods that are employed in data exploration since there is usually so much data that a single perspective or view on the data is not sufficient.

A very common and simple interaction technique is the *selection* of one or more objects to retrieve details about them. For instance, when a scatter plot is used, it is often desirable to obtain information about certain data points, such as a group of outliers, which could be provided in as a "tool tip" once the mouse is moved over a data point. Visualizations that are designed to display very large amounts of data often support *zooming* and *panning*[3] . Zooming enlarges an area of the visualization and panning is used to shift the enlarged view horizontally or vertically in order to move areas of the visualization into the visible frame. A technique that is often used in combination with zooming and panning is *animation*, which allows these operations to be performed gradually, rather than abruptly. The advantage of animations is that the investigator can retain the context of the data that are being focussed on (Robertson et al., 1993). This has also been found to be beneficial when applied to transitions between two different graphical representations of the same data (Heer and Robertson, 2007).

Zooming and panning are direct *filtering* methods because they allow the investigator to focus on a subset of the data directly through interaction with the visualization. In this case, the filtering is typically applied to attributes of the data that have been encoded spatially. However, it is often desirable to filter on attributes that are encoded non-spatially, which is usually achieved indirectly with components of the user interfaces that are not part of the visualization, i.e the graphical representation of the data. Here the visualization would still be considered interactive,

[3]In 3-dimensional visualizations *rotation* is another important interaction technique, however, 3-dimensional visualizations are beyond the scope of this dissertation.

because the visualization can be updated or modified by the investigator.

When multiple views of the data are used, they are often *linked* using an interaction technique called *brushing*. Brushing highlights objects in all linked views that have been selected in the currently active view. This approach helps investigators to identify patterns and it improves understanding since multiple aspects of the data can be examined in combination (Keim, 2002). Another form of linking can be achieved by automatically updating related views to reflect panning or zooming in the active view so as to maintain the same focus in all views.

Interaction techniques and their importance for visual data exploration are highlighted in the famous "visual information seeking mantra" coined by Shneiderman (1996):

Overview first, zoom and filter, then details-on-demand.

1.2.2 Visualization of Transcriptomics Data

An expression matrix is essentially a collection of high-dimensional vectors that represent either gene or sample expression profiles. Several common visualization methods exist to visualize such multi-dimensional data. Scatter plots, profile plots and heat maps (see Figure 1.6) are the ones most commonly used for transcriptomics data and have been implemented in wide range of visualization and analysis tools (Gehlenborg et al., 2010).

The following sections give a brief overview of these methods. A more comprehensive discussion of the advantages and disadvantages of scatter plots, profile plots and heat maps and how they are used is presented in Chapter 4.

1.2.2.1 Scatter Plots

Scatter plots are primarily used to examine dependencies between two variables, but used in combination with dimensionality reduction methods, they can also be applied to multivariate data. For example, to gain

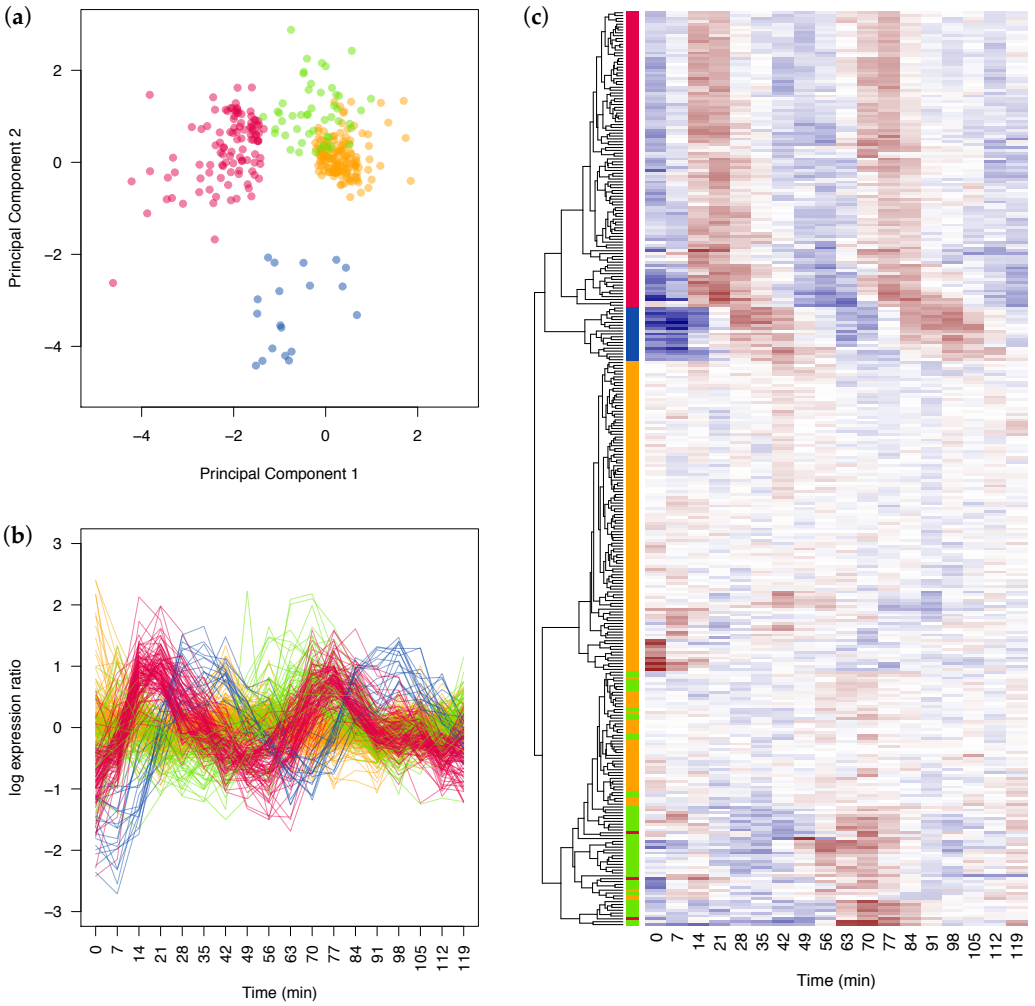

Figure 1.6: Visualization of an expression matrix. The data shows the $\log_2$-transformed expression levels of 320 transcripts from *Saccharomyces cerevisiae*, collected over 18 time points throughout the cell cycle after synchronization (Spellman et al., 1998). Colors indicate cluster membership based on a k-means clustering (k = 4). (**a**) Scatter plot showing a projection of the profiles on the first two principal components obtained by PCA. The marks for the data points in this plot are not fully opaque but semi-transparent, which helps to distinguish regions of varying density within clusters, since marks plotted on top of each other appear darker. (**b**) Profile plot. The red cluster contains genes active in the G1 phase and the blue cluster contains genes that are active in the S phase of the cell cycle, for the yellow and the green cluster the phase assignment is not clear. (**c**) A heat map of the profiles. Colors represent abundance levels (red = higher than control sample, blue = lower and white = no change). Here the rows of the heat map have been re-ordered according to a hierarchical clustering, which is shown on the left in form of a dendrogram. The color bars between the dendrogram and heat map indicate the k-means clusters, allowing comparison of the two clustering results.

insight into the global patterns in a gene expression matrix a dimensionality reduction method is applied to obtain a two-dimensional (sometimes three-dimensional) representation of the expression profiles, which are then visualized in a scatter plot revealing clusters and outliers in the data. Some frequently applied dimensionality reduction methods for this purpose are Principal Component Analysis (PCA; Hotelling, 1933) and Multi-Dimensional Scaling (MDS; Kruskal, 1964), which are implemented in many tools. Besides PCA and MDS, many other suitable dimensionality reduction methods exist (Venna and Kaski, 2007b), but are less commonly applied. Scatter plots combined with dimensionality reduction methods are an excellent tool to gain insight into the overall structure of large sets of gene expression profiles. However, due to the dimensionality reduction the original expression profiles are no longer accessible in the visualization, it is not possible to directly extract information about the relationship between expression levels and the conditions under study.

1.2.2.2 Profile Plots

Profile plots visualize the expression levels of a large number of transcripts across all samples. Thus, they provide insight into the patterns of correlation between samples and expression levels. For instance, at a glance, the investigator can determine whether a transcript is expressed constitutively in all conditions or whether it is only expressed in a single condition, such as a particular tissue or phase of the cell cycle. Furthermore, it is possible to generate hypotheses about trends, such as increasing expression levels for a transcript over time after a stimulus, or differential expression of a transcript, for instance between samples of diseased and normal tissue. Since many profiles are shown in the same plot the investigator can interpret such observations in the context of the overall data set.

A profile plot can also be queried visually for transcripts with a particular behavior, such as low expression levels in one set of samples and high levels in another set, or, for profiles that are similar to that of a transcript of interest. A major disadvantage of profile plots is that, due to their

construction, profiles overlap, severely limiting the number of profiles that can be visualized effectively at the same time.

1.2.2.3 Heat Maps

Heat maps (Eisen et al., 1998; Wilkinson and Friendly, 2009) are the most commonly used visualization method for expression matrices (Weinstein, 2008) and can be generated with most tools. Like profile plots, heat maps visualize the abundance of each transcript in each sample but in a heat map the profiles do not overlap, which means that a larger number of profiles can be visualized effectively. However, the size of the heat map grows with the number of profiles, so that the available screen space is a limiting factor.

A key aspect of heat map visualization is the reordering of the rows, which ensures that similar profiles are placed close to each other. Typically this reordering is done using hierarchical clustering (Wilkinson and Friendly, 2009), and a dendrogram showing the hierarchy is usually arranged immediately adjacent to the heat map, as illustrated in Figure 1.6(**c**). This combined view helps the investigator to see groups of genes that have a similar expression pattern. The dendrogram clearly conveys to the investigator exactly which genes are clustered together, and also which genes are outliers with a very unusual expression pattern. The heat map allows the investigator to see in more detail which features of the expression pattern are shared by gene clusters. Typically, for example, the investigator may see that one group of genes have a peak expression at about the same time in an experiment.

1.2.2.4 Integration with Network Visualizations

Transcripts level are frequently integrated with network data and analyzed in this context. For instance, visualization of expression profiles combined with a gene regulatory network can explain expression patterns, such as downregulation of genes when a repressing transcription factors

is upregulated. Several techniques have been developed for such integrated visualizations. These range from simple coloring of nodes in the network combined with animation to show changes across conditions or time points to methods that use sophisticated visualizations of complete expression profiles in each node. While increasingly important for many biomedical studies, these techniques are beyond the scope of this dissertation and have been reviewed recently by Gehlenborg et al. (2010).

1.3 Research Questions for the Dissertation

The general topic of the research presented in this dissertation is the discovery of patterns in transcriptomics data. The growing amount of transcriptomics data is the major challenge that is being addressed in this dissertation. Computers and computational methods are better prepared to deal with the increasing size and number of data sets than human investigators, since computers can be upgraded with new generations of processors and more memory and storage capabilities. However, since human cognitive and perceptive skills and capabilities do not change over time (Thomas and Cook, 2005, see Chapter 1), visual methods for pattern discovery require special attention. But visualization is not the only interface between data and humans in data analysis. For instance, humans need to be able to interpret the results produced by computational pattern discovery methods, which can be comprehensive and complex. Thus it is important to take into account the interpretability of the results in the design computational methods. This dissertation addresses some of these challenges and introduces exploratory approaches for two distinct organizational levels of transcriptomics data: individual data sets and collections of data sets.

1.3.1 Exploring Collections Data Sets

The first research question that this dissertation addresses is how information retrieval and visualization methods can be used for efficient exploration of large collections of transcriptomics data sets.

In Chapter 2 a knowledge-driven approach for exploration of transcriptomics data repositories based on ontology visualization is introduced. The method is based on a mapping of data sets in the ArrayExpress Archive to terms in the *Experimental Factor Ontology* (EFO; Malone et al., 2010). A tree structure is extracted from the ontology and this weighted tree is then visualized as a *tree map* (Johnson and Shneiderman, 1991), which provides access to the data sets. The systems is fully web-based and implemented using JavaScript and web services.

In Chapter 3 two novel data-driven approaches based on *gene set enrichment analysis* (GSEA; Subramanian et al., 2005) and generative probabilistic models are presented. The first method uses GSEA to effectively convert studies in the repository into vectors of differentially expressed gene sets. Based on these vectors a *topic model* (Blei et al., 2003b) is built, in which every study represented as a combination of topics. The topics are probability distributions over gene sets. The topic model provides a straightforward way to perform probabilistic queries in the collection of studies, which ranks the studies based on their similarity to a query study. Information extracted from the topic model is employed in novel visualizations to visualize the space of studies in the repository, as well as the activity of biological processes in those studies. A manual classification of a subset of the studies is applied to evaluate the performance with standard information retrieval measures.

The second method presented in Chapter 3 uses the same basic approach as described above, but a refined probabilistic model that is an extension of the topic model used in the original method. Furthermore, the system is extended to handle data from mouse and rat in addition to human data. Rather than using a manual approach, evaluation of the model and comparison with other approaches is performed with the help of the

EFO as an external gold standard. Finally, a web-based interface to query the collection of data sets and to identify interesting connections between studies is presented. Some findings are described and discussed in several case studies that are also part of Chapter 3.

1.3.2 Visualizing Large Data Sets

The second research question of this dissertation deals with the visualization of large transcriptomics data sets. This work has been driven by the observation that there is a growing number of data sets with hundreds or thousands of samples and a lack of suitable methods to visualize such data sets.

In Chapter 4 an analysis of the visualization tasks in transcriptomics data analysis is presented. Based on the evaluation of commonly used visualization methods, such as heat maps, profile plots and scatter plots, the design of a novel interactive visualization method called *Space Maps* is presented. The Space Maps technique is a pixel-oriented visualization method that combines the *Value and Relation (VaR) display* by Yang et al. (2004, 2007) with hierarchical representations of the expression profiles that are similar to tree maps. The glyphs represent expression profiles and the arrangement of the glyphs in 2-dimensional space represents relationships between the profiles. Additionally, several case studies of the Space Maps technique are presented in Chapter 4.

Contributions and Publications

The material presented in Sections 1.1.2.2 and 1.2.2 was published previously in similar form as part of the following manuscript: N. Gehlenborg, S. I. O'Donoghue, N. Baliga, A. Goesmann, M. A. Hibbs, H. Kitano, O. Kohlbacher, H. Neuweger, R. Schneider, D. Tenenbaum and A.-C. Gavin. Visualization of Omics Data for Systems Biology. *Nature Methods* 7(3):S56-S68, 2010.

Chapter 2

Ontology-guided Visual Exploration of a Repository

2.1 Introduction

The large collections of transcriptomics data sets that are made available in repositories such as the ArrayExpress Archive and the Gene Expression Omnibus (GEO) can be re-used to support biomedical research. On the one hand, the data can, for example, be used to find genes whose expression profiles are correlated across a wide range of conditions (see e.g. Hibbs et al., 2007; Kapushesky et al., 2010). On the other hand, such collections can be queried for conditions of interest, which is the topic of this chapter.

For instance, it is possible to query collections of data sets to answer questions such as "Which diseases have been studied with transcriptomics technologies?", "Which data sets are available for a particular knock-out strain?" or "How extensively has a particular cell line been studied?" and to obtain the corresponding data sets. Answering such questions is only possible if the data are integrated and accessible through an appropriate query interface.

When querying collections of transcriptomics data sets for condi-

tions, there are two primary sources of information that can be used for retrieval, which define two distinct approaches for search and exploration.

Meta Information The meta information associated with a data set is a comprehensive description of the experimental setup, including details about what conditions the investigators studied and how the study was performed. The description is typically provided as a combination of free text and terms from controlled vocabularies or ontologies (see Section 1.1.4). However, the meta information associated with a data set almost always contains only information about what was known before the study[1]. In this dissertation, approaches using only meta information for retrieval of data sets from collections are called *knowledge-driven* approaches.

Expression Data The expression matrix contains transcript levels measured under the conditions investigated in the study (see Section 1.1.4). To distinguish from the knowledge-driven approaches introduced above, approaches in which retrieval is based on the measurements in the expression matrix are referred to as *data-driven* approaches in this dissertation.

Besides these primary sources of information it is also possible to employ derived data sources, such as publications discussing findings obtained through analysis of a data set. However, in this dissertation only the aforementioned primary sources are considered.

The choice between a knowledge-driven or a data-driven approach depends on the goals of the investigator. For instance, if the investigator is interested in retrieving studies that all investigated a particular condition or biological state, say "lung adenocarcinoma" or "HeLa cell line", then a knowledge-driven approach is appropriate. Examples for such approaches are the query interfaces of ArrayExpress (Parkinson et al., 2009)

[1] Exceptions are data sets for which some of the findings were submitted as part of the experimental description.

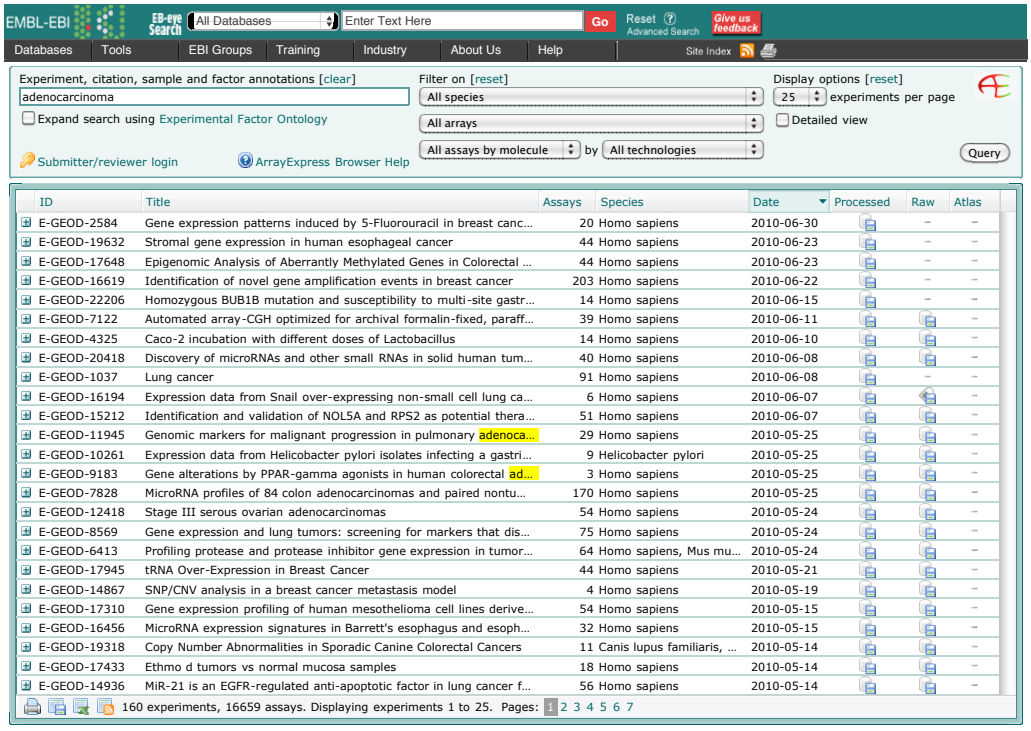

Figure 2.1: The ArrayExpress query interface showing results for query "adenocarcinoma". *(Source: www.ebi.ac.uk/microarray-as/ae/browse.html?keywords=adenocarcinoma, retrieved on 6 July 2010)*

and GEO (Barrett et al., 2009), as well as related search engines (e.g. Zhu et al., 2008). However, if the investigator is interested in retrieving studies in which similar expression patterns are observed as in a study of interest, then a knowledge-driven approach would be of little value. In such cases a data-driven approach is required that considers the content of the expression matrices.

The ArrayExpress Archive (www.ebi.ac.uk/arrayexpress) currently contains data from almost 13,000 studies, in which the transcriptome of hundreds of diseases, genotypes, tissues and other experimental conditions have been analyzed (see Section 1.1.5.1). This chapter introduces a novel knowledge-driven approach approach for exploration of the repository based on a tree map visualization of the Experimental Factor Ontology (EFO, see 1.1.4.3).

The ArrayExpress Archive user interface shown in Figure 2.1 sup-

ports lookup by accession number and offers extensive functionality to query the repository with keywords. Keywords are matched against the meta information of the data sets. If desired by the user, querying can be supported by *query term expansion*, which attempts to match the query terms against the EFO and, if successful, also includes synonyms and child terms of the ontology term in the search. For instance, query term expansion applied to a query "cancer" maps the query to EFO term "cancer" (EFO_0000311) and queries with the original query "cancer", its synonyms "malignant neoplasia", "malignant tumor" and "malignant tumour" as well as its child terms "carcinoma", "central nervous system cancer", "chordoma", "lymphoid neoplasm", "mesothelioma" and "sarcoma", their synonyms, as well as all of their descendants and the synonyms of their descendants[2]. For more fine grained control of the keyword search, a simple query language can be used to limit the queries to particular fields of the meta information, such as the experimental factor values or the species name (N. Kolesnikov, personal communication). However, it is important to note that the ArrayExpress Archive query interface is completely knowledge-driven without consideration of the expression data associated with the studies in the archive.

Search results are presented as a table (see Figure 2.1) and can be filtered and sorted by species, number of arrays used in the study, submission date, availability of raw or processed data and several other attributes. Details about retrieved studies are displayed directly in the results table by expanding the corresponding row as shown in Figure 2.2.

The search and filtering capabilities of the query interface make it well suited to retrieve studies related to a specific query provided by the investigator. However, since it relies on keyword search, the interface is not well suited to providing an overview of the available data sets or to browsing through the content of the archive without defining a specific query.

Browsing a collection of data sets rather than searching for a specific

[2]Terms and synonyms based on EFO version 121.

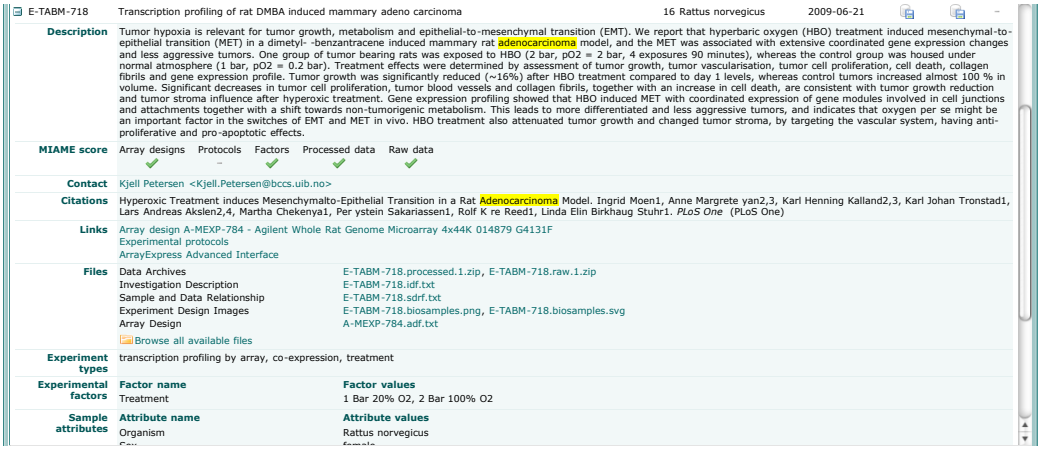

Figure 2.2: The ArrayExpress Archive user interface showing details for a study retrieved by querying for "adenocarcinoma". The query term "adenocarcinoma"is highlighted with a yellow background in the results. Details about the study include a brief free text summary, contact details for the authors, a reference to a publication related to the data set, details about the experimental factors and factor values as well as further information about the samples. Links for download of MIAME-compliant descriptions of the data are provided. *(Source: www.ebi.ac.uk/microarray-as/ae/browse.html?keywords=adenocarcinoma, retrieved on 6 July 2010)*

data set can be useful in several ways. For instance, Marchionini (1995, see Chapter 6) suggests that browsing helps for instance to gain an overview, to discover new information and to identify new keywords that could be used for a more targeted search. It also frees resources in the cognitive system by shifting the workload to the perceptual system to identify useful information by filtering, rather than using the cognitive system to generate keywords that might help to retrieve data sets of interest. This is summarized in the idea of "recognition over recall", which is one of the major results of cognitive science. It states that it is easier for humans to recognize something than to generate a mental representation of it (Hearst, 2009). This is also emphasized by Aula and Siirtola (2005), who differentiate between "browsing" and "searching" as follows: "Considered in cognitive terms, searching is a more analytical and demanding method for locating information than browsing, as it involves several phases, such as planning and executing queries, evaluating the results, and refining the queries, whereas browsing only requires the user to recognize promising looking links."

Browsing requires a structure that can be used to organize the collection of data sets. Since there is no inherent structure in the studies contained in the ArrayExpress Archive, an external structure is required to enable browsing. The EFO is a highly structured and well-defined description of the experimental factor values used to annotate these studies. Thus the EFO is a suitable structure to organize the content of the archive.

A graphical representation of the EFO structure is desirable, since unlike a textual representation, it can provide an overview of the whole collection of data sets and reveal patterns that otherwise would remain hidden. Tree maps (Johnson and Shneiderman, 1991) have been used extensively to visualize hierarchical structures that organize quantitative information. Similarly, the use of other types of maps in exploratory interfaces has been studied as well, for instance, by Lin (1997) and Chen and Hearst (1998), who used self-organizing maps (Kohonen, 1982) to create graphical overviews of document collections.

To enable browsing in the ArrayExpress Archive the *ArrayExpress Explorer* was developed. The ArrayExpress Explorer is a web-based tool that is built around an interactive tree map visualization of the EFO. The tree map visualization provides an overview of the data sets in the ArrayExpress Archive at different levels of resolution and supports ontology-guided browsing of thousands of data sets.

2.2 Methods and Data

The EFO models the relationships between experimental variables used in studies that have been deposited in ArrayExpress. The ontology is a directed, acyclic graph (DAG) that contains different types of relationships between parent and child terms. Most relationships are either of type "is_a" or of type "part_of". Since some of the terms in the EFO have more than one parent term, the ontology is not a tree structure. However, in order to visualize the EFO as a treemap, every term except the root must have exactly one parent term. To achieve this, the EFO is converted into a

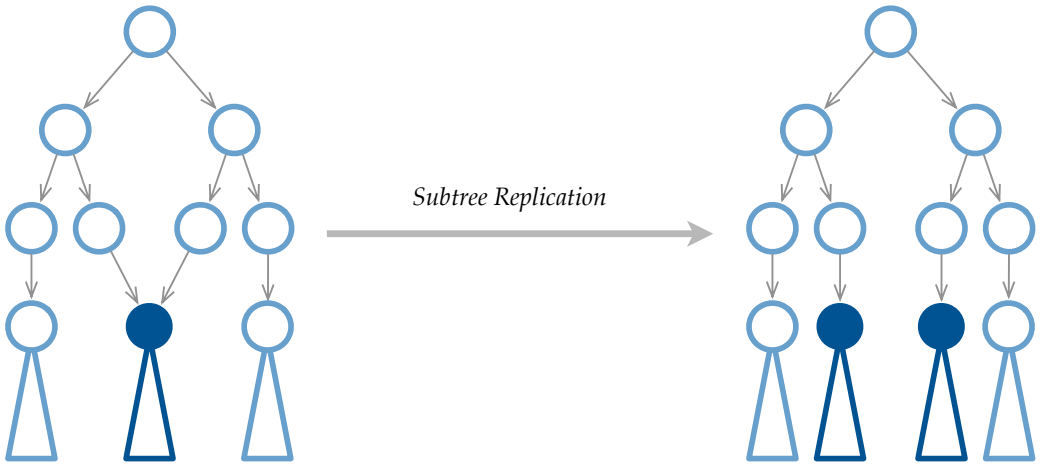

Figure 2.3: Ontology subtree replication. In order to obtain a tree structure from the ontology, subtrees with more than one parent node are replicated for each parent node.

tree by following all "is_a" relationships and by replicating subtrees with a root with more than one parent. This is illustrated in Figure 2.3.

Some of the high-level structure of the EFO has been created primarily to provide compatibility with other biomedical ontologies. As a result, many of the high-level terms used to structure the lower levels of the ontology are very technical and provide no value for the exploration of the ArrayExpress Archive. To resolve this issue, the high-level terms of the EFO are restructured to provide easier access to investigators who are not familiar with the EFO and to remove parts of the EFO that are not useful for exploration, as shown in Figure 2.4. The simplified structure was obtained by attaching the relevant branches directly to the root of the ontology and by removing all others. A similar approach has been taken by the ArrayExpress Atlas database to use the EFO to support querying of differentially expressed genes (J. Malone, personal communication).

In the next step, the studies associated with each ontology term are determined. A term t is associated with a study s if the string representation t_s of the term is contained in any of the experimental factor value descriptions of the conditions that are part of study s. For instance, if t_s is "cancer" it matches the following fictitious experimental factor value descriptions "cancer" and "lung cancer", but not "cancerous lesion". After

Figure 2.4: Restructured Experimental Factor Ontology. Colors correspond to the coloring used for the ontology branches in later figures in this section. (**a**) Original high-level structure of the EFO expanded to show all terms that make up the revised high-level structure. Most terms shown here are roots of subtrees. (**b**) Revised high-level structure for data exploration purposes.

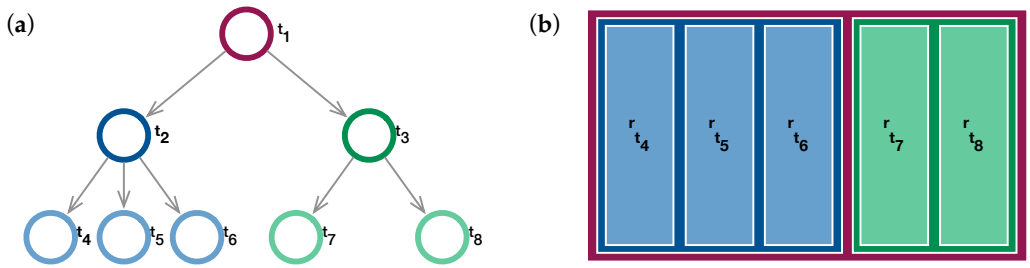

Figure 2.5: Tree map construction. (**a**) A tree consisting of eight nodes. (**b**) A tree map representing the tree shown in (**a**). The color of the rectangles r_t corresponds to the color of the nodes t shown in (**a**). In this example all leaf nodes have the same weight and thus the same size.

the mapping every term t of the ontology is associated with t_c studies and $t_c \geq 0$.

The resulting tree is visualized as a tree map (Johnson and Shneiderman, 1991). Tree maps are space-filling visualizations in which every term t of the tree is represented by a rectangle r_t as illustrated in Figure 2.5. If the term t has a number of children u_i, their rectangles r_{u_i} are placed within the rectangle r_t. The quantity t_c, i.e. the number of studies associated with every term t is encoded in $size(r_t)$, which is the area of rectangle r_t. In the tree map construction $size(r_t) = \sum_i size(r_{u_i})$ if t is an inner node or $size(r_t) \geq 0$ if t is a leaf.

In order to create a tree map visualization of the weighted tree ob-

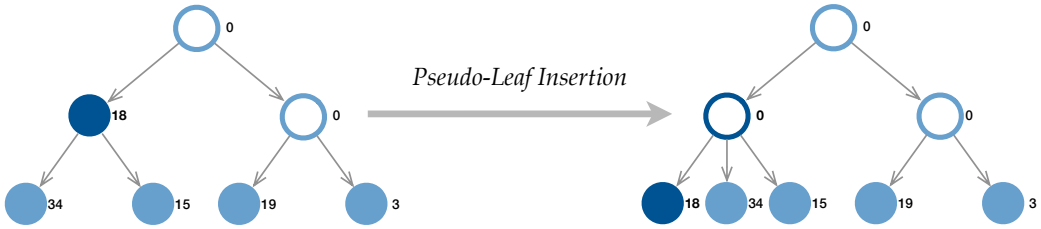

Figure 2.6: Ontology tree pseudo-leaf insertion. Filled circles are nodes associated with one or more data sets, circle outlines are nodes associated with no data sets. The numbers next to the circles represent the number of data sets associated with the corresponding node. Inner nodes that have been associated with one or more data sets cannot be visualized directly in a tree map. To accommodate such cases, a pseudo leaf is inserted by duplicating the original inner node as a leaf of itself and by removing the association between the inner node and the data sets.

Figure 2.7: Ontology tree map branch color scheme. A qualitative color scheme is used to differentiate the twelve branches defined by the top-level terms of the modified ontology shown in Figure 2.4. A sequential color scheme using tints of the colors, i.e. varying lightness, from the qualitative color scheme is used to represent the depth of a node. Here the depth of the node is defined as the length of the longest path from the node to a leaf. Lighter colors indicate less depth, i.e. the subtree beneath the node is rather flat and does not provide much structure. See "cell line" (flat structure) and "organism part" (deep structure) in Figures 2.17 through 2.20 on pages 54-55 for an example.

tained by mapping data sets to the ontology tree structure an additional step is required. Since $size(r_t)$ is defined as the sum of sizes of the child rectangles, a tree map cannot accommodate cases in which an inner node has been associated with one or more studies. To resolve such cases, a pseudo-leaf node is created for every inner node that is associated with with one or more data sets, as illustrated in Figure 2.6.

A second variable, and sometimes also a third variable, associated with a node or term t can be encoded in the color of the corresponding rectangle r_t. These variables can be either nominal or quantitative. For nominal variables a qualitative color scheme is used, while for quantitative variables a sequential or diverging color scheme is used. The color schemes used for the tree map visualization of the EFO are shown in Figure 2.7.

A series of tree map algorithms have been devised to lay out the rectangles r_t. Among the better known algorithms are the *slice-and-dice algorithm* (Johnson and Shneiderman, 1991), the *squarified tree map algorithm* (Bruls et al., 2000) and the *mixed tree map algorithm* (Vliegen et al., 2006). All tree map algorithms take an input area representing the root node r_{root} and subdivide this area to represent the rectangles r_t of all other nodes in the

tree. The main differences between these algorithms are the aspect ratios[3] of the generated rectangles r_t, the stability of the location of the rectangles when the t_c change and the preservation of the order of the input data (Tu and Shen, 2007). In the ideal case the layout algorithm produces aspect ratios close to 1, the layout is not affected by small changes in the t_c and the order of the input data is preserved as much as possible. However, no known algorithm has all of these properties and in practice an algorithm has to be chosen that represents the best trade off for the given problem. For the visualization of the ontology tree and the associated data set counts the *squarified tree map algorithm* was chosen. As indicated by its name, the algorithm optimizes the aspect ratio of the generated rectangles, which is helpful for label placement and user interaction and is also aesthetically pleasing.

2.2.1 Implementation

The system consists of a server-based back-end and a client-based front-end as illustrated in Figure 2.8. The back-end provides access to the ArrayExpress Archive query interface, computes the mappings from data sets to ontology terms and provides the EFO tree structure. The modified structure shown in Figure 2.4 is created in the front-end for greater flexibility. All components of the back-end are implemented as web-services that support a Representational State Transfer (REST) architecture, i.e. they are RESTful, and provide JavaScript Object Notation (JSON) data structures. The data structures are processed by the browser-based JavaScript front-end. Three key types of data structures used by the front-end are provided by different web services that use JSON with padding (JSONP) to support cross-domain queries, i.e. queries to servers that are not the same as the one hosting the front-end:

[3]The aspect ratio of a rectangle is the length of its longer side divided by the length of its shorter side, i.e. a square has an aspect ratio of 1 and a A4 page (297 mm $\times$ 210 mm) of paper has an aspect ratio of approximately 1.414.

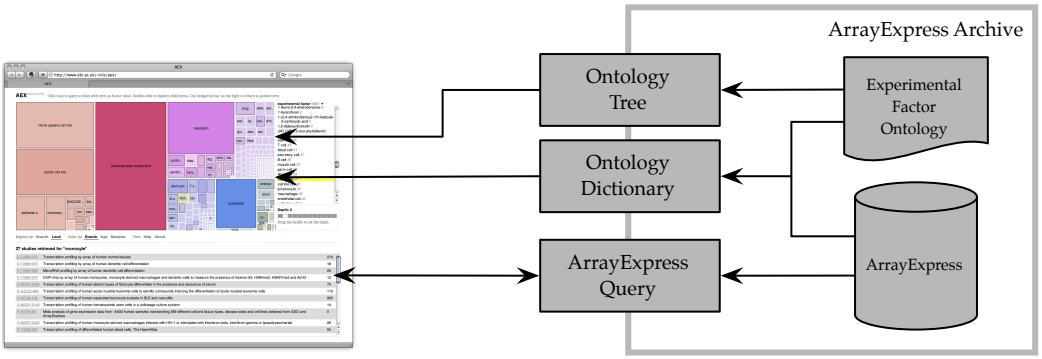

Figure 2.8: ArrayExpress Explorer implementation. The front-end on the left accesses the web services implemented in the back-end to retrieve the ontology tree structure and the number of studies mapped to each term. Furthermore, the front-end queries the ArrayExpress Archive through a third web service when more information about studies associated with a term have to be retrieved.

Ontology Tree Provides the ontology tree structure derived from the ontology DAG. The tree corresponds to the full EFO structure. This service is called once during the initialization phase to create the tree map.

Ontology Attributes Provides a dictionary that contains additional information about each term t in the ontology such as the string representation t_s, the number of associated studies t_c, median number of samples in all associate studies, the age of the oldest associated study in days. This service is called once during the initialization phase to create the tree map.

Archive Query The default query web service provided by the ArrayExpress Archive. This web service is also used by the existing ArrayExpress Archive keyword-based query interface described in Section 2.1. This service is called every time the investigator wants to retrieve the list of studies associated with a term.

The front-end part consists of the tree map visualization and related user interface components. The tree map is implemented in JavaScript and builds on the squarified tree map visualization provided by the Protovis library (Bostock and Heer, 2009). The JQuery library (www.jquery.com) is

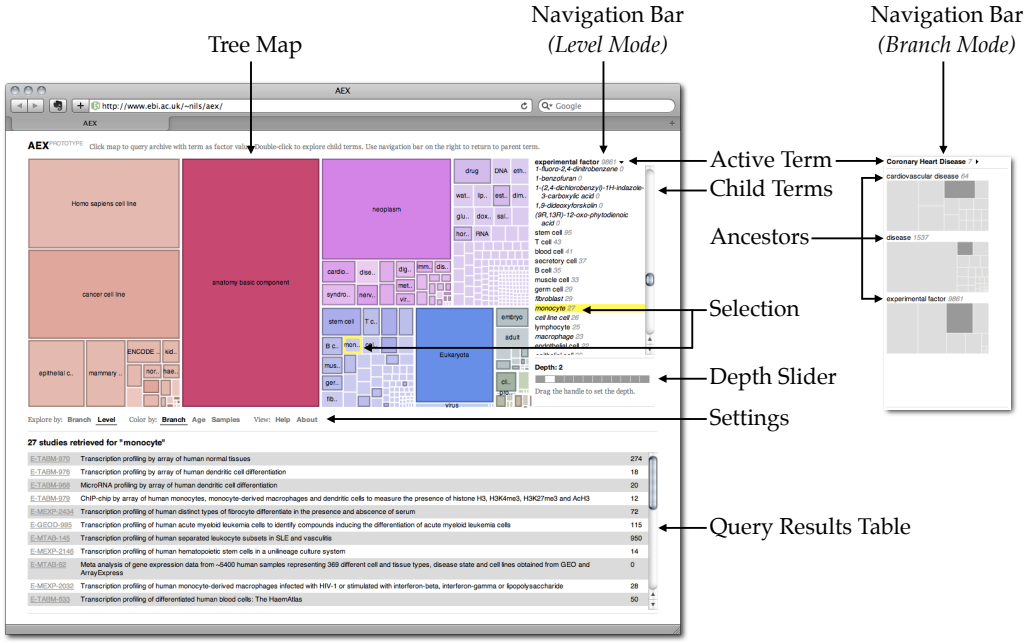

Figure 2.9: ArrayExpress Explorer user interface. The main components are the tree map, the navigation bar and the query results table showing the studies associated with a term. The investigator can choose coloring and exploration mode (branch or level) in the settings bar beneath the tree map. *(Source: www.ebi.ac.uk/~nils/aex, retrieved on 8 August 2010)*

used to implement user interactions, animations and access to the aforementioned web services.

2.3 Results

The components of the user interface are illustrated in in Figure 2.9. The main components of the user interface are the tree map, the navigation bar and the query results table. A prototype implementation of the *ArrayExpress Explorer* is available at www.ebi.ac.uk/~nils/aex.

2.3.1 Tree Map and Query Results Table

The r_t of the tree map can either be colored by ontology branch, as shown in Figures 2.11 through 2.20 or by a quantitive variable, as shown in Fig-

ures 2.21 and 2.22. When colored by branch, the hue[4] indicates the branch and the lightness is used to indicate the depth of the subtree beneath the node, as illustrated in Figure 2.7. In this context the size and the color of the rectangles in the tree map are used as *information scent* (Pirolli et al., 2000), because they indicate how much additional information can be found by exploring a particular term. To distinguish leaves and inner nodes, inner nodes have a dark outline, whereas leaves do not have an outline.

The tree map provides several interactive features. When moving the mouse over the tree map visual feedback is provided by highlighting the rectangle under the cursor, i.e. the selected term, with a yellow outline, as shown in Figure 2.9. Single-clicking any rectangle in the tree map will retrieve details about the associated studies from the ArrayExpress Archive and list them in the query results table below the tree map. Selecting a study from this table takes the investigator to the main ArrayExpress Archive query results page to display the full description of the study as shown in Figure 2.2.

By double-clicking an inner node the investigator can descend to the next level of the ontology tree. The result of this operation depends on the selected exploration mode. In *branch mode*, the tree map is updated to show only the child terms of the term that was double-clicked, whereas in *level mode* the depth level of the whole tree map is increased. This difference between level and branch mode and the visibility of rectangles is illustrated in Figure 2.10 and further demonstrated by examples in Sections 2.3.3 and 2.3.4.

[4]The *hue* of a color is what colloquially is referred to as "color", e.g. "red", "blue" or "purple", while "light red" is a *tint* of red, i.e. red mixed with white, and "dark red" is a *shade* of red, i.e. red mixed with black. Tints and shades refer to the *lightness* of a color (Wong, 2010).

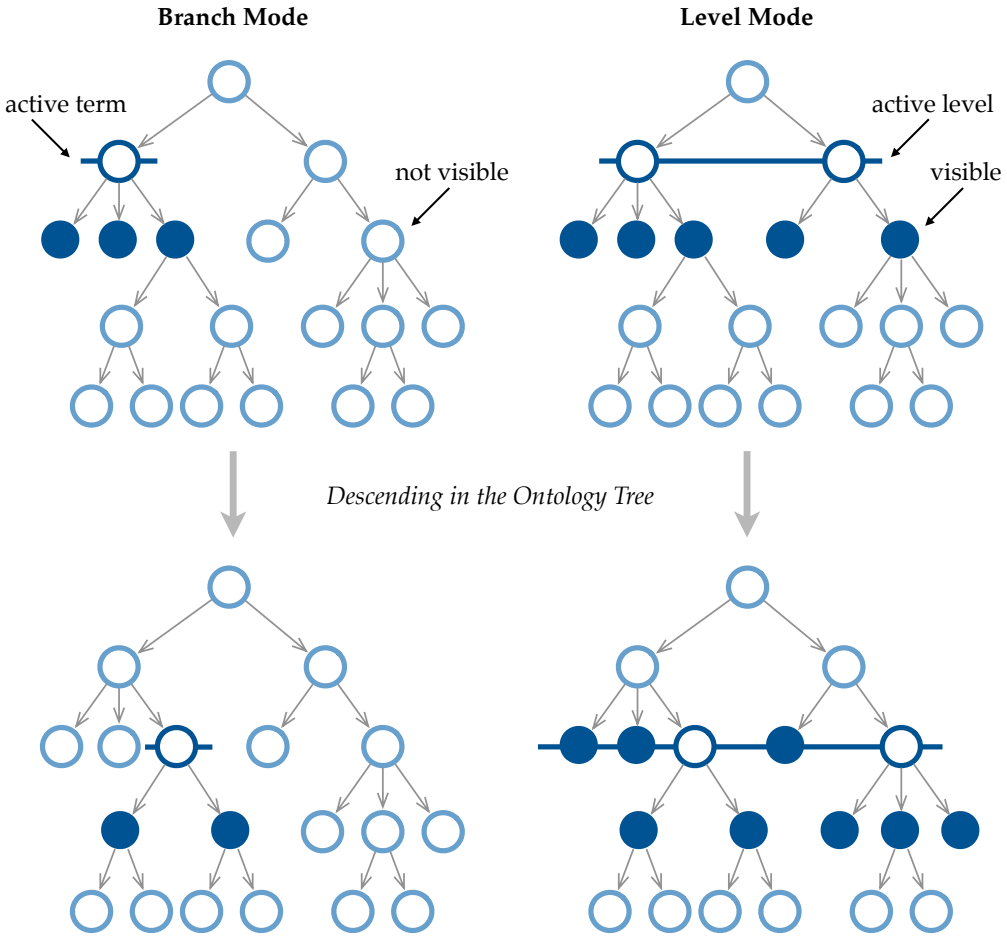

Figure 2.10: Differences between *branch mode* and *level mode*. In *branch mode* only the children of a single, active term are visible, whereas in *level mode* the children of all terms at the active depth level are visible. In *level mode* leaves are visible at all depths equal or lower than their own (see lower right tree).

2.3.2 Navigation Bar

In *branch mode* the navigation bar on the right shows small instances of the tree map corresponding to each of the ancestor terms of the active term shown in the main tree map. These tree maps also highlight the term that was explored on the corresponding level and essentially present the path through the ontology tree to the currently active term. Double-clicking any of tree maps in the navigation bar will make the corresponding term the active term shown in the main tree map and thus take the investigator

up in the ontology tree.

In *level mode* a slider bar is shown instead of the small tree maps and by moving the slider handle the investigator can select the depth at which the tree map should be shown. The slider handle also updates when the main tree map is used to descend down the ontology tree by double-clicking on any of the rectangles.

The navigation bar also contains a collapsible *child term list* that shows all child terms of the currently active term. In this list leaves are set in italics to distinguish them from inner nodes. Moving the mouse over a child term will highlight the term in yellow in the list, as well as in the tree map, and vice versa. Inner nodes can be double-clicked to the same effect as double-clicking an inner node in the tree map. Single-clicking a term will retrieve details of the associated studies. The list also shows the total number of studies associated with each child term and their corresponding subtrees. Unlike the tree map, the child term list also contains terms that are not associated with any studies in the ArrayExpress Archive.

2.3.3 Branch Mode

The branch mode allows the investigator to explore the content of the archive by selecting an ontology branch of interest, such as "disease", and to successively refine this choice by descending down the ontology tree. As only direct child terms of the active term as are shown, the space to place labels for the rectangles is maximized. At the same time, other branches can be accessed efficiently through the navigation bar.

Figure 2.11: Exploration of the ArrayExpress Archive in *branch mode*. The navigation bar child term list on the right is expanded. The yellow box indicates the branch that will be explored. *(Source: www.ebi.ac.uk/~nils/aex, retrieved on 6 July 2010)*

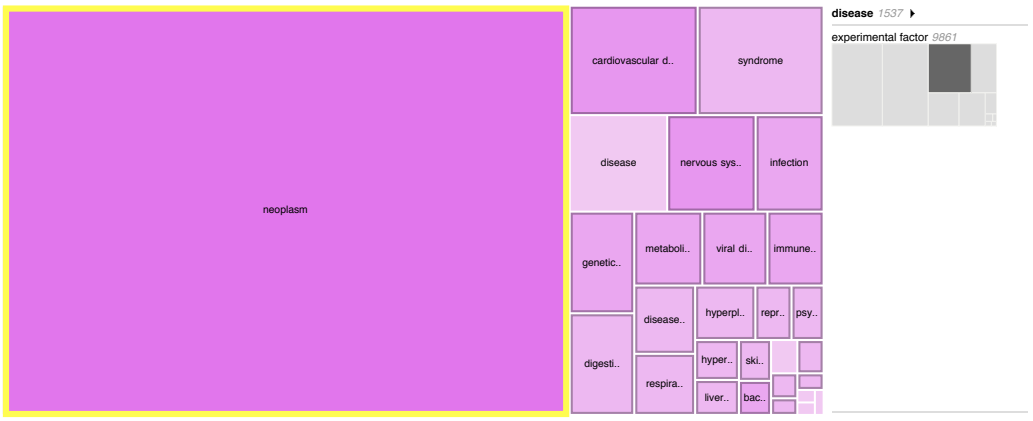

Figure 2.12: Exploration of the "disease" branch in *branch mode*. The navigation bar child term list on the right is collapsed. The yellow box indicates the branch that will be explored. *(Source: www.ebi.ac.uk/~nils/aex, retrieved on 6 July 2010)*

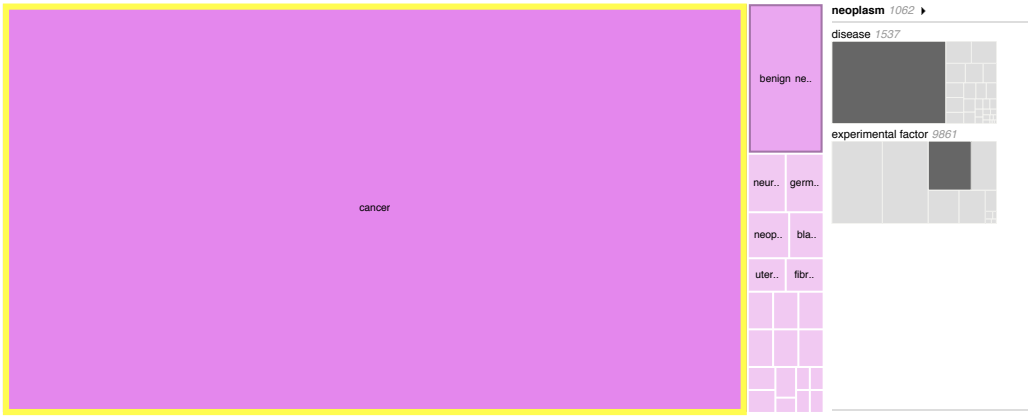

Figure 2.13: Exploration of the "neoplasm" branch in *branch mode*. The navigation bar child term list on the right is collapsed. The yellow box indicates the branch that will be explored. *(Source: www.ebi.ac.uk/~nils/aex, retrieved on 6 July 2010)*

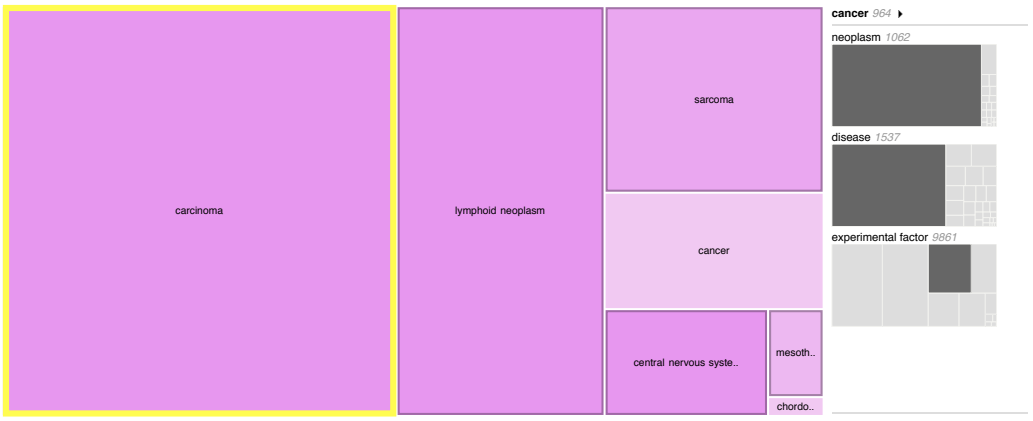

Figure 2.14: Exploration of the "cancer" branch in *branch mode*. The navigation bar child term list on the right is collapsed. The yellow box indicates the branch that will be explored. *(Source: www.ebi.ac.uk/~nils/aex, retrieved on 6 July 2010)*

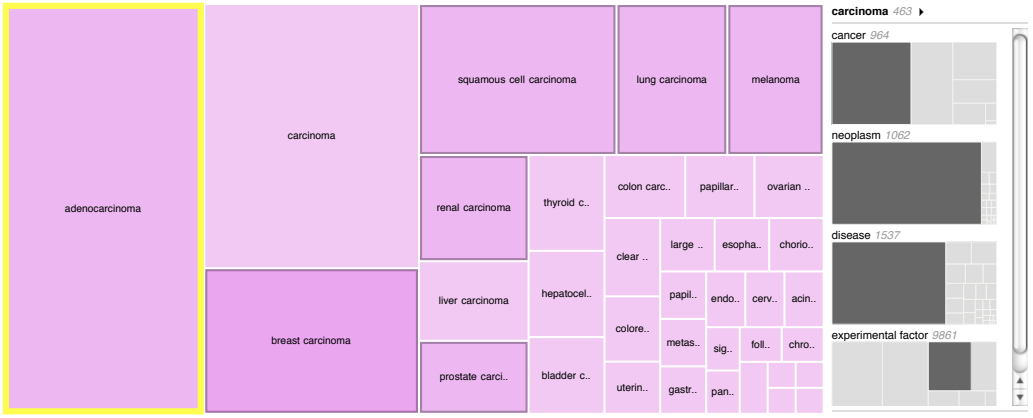

Figure 2.15: Exploration of the "carcinoma" branch in *branch mode*. The navigation bar child term list on the right is collapsed. The yellow box indicates the branch that will be explored. *(Source: www.ebi.ac.uk/~nils/aex, retrieved on 6 July 2010)*

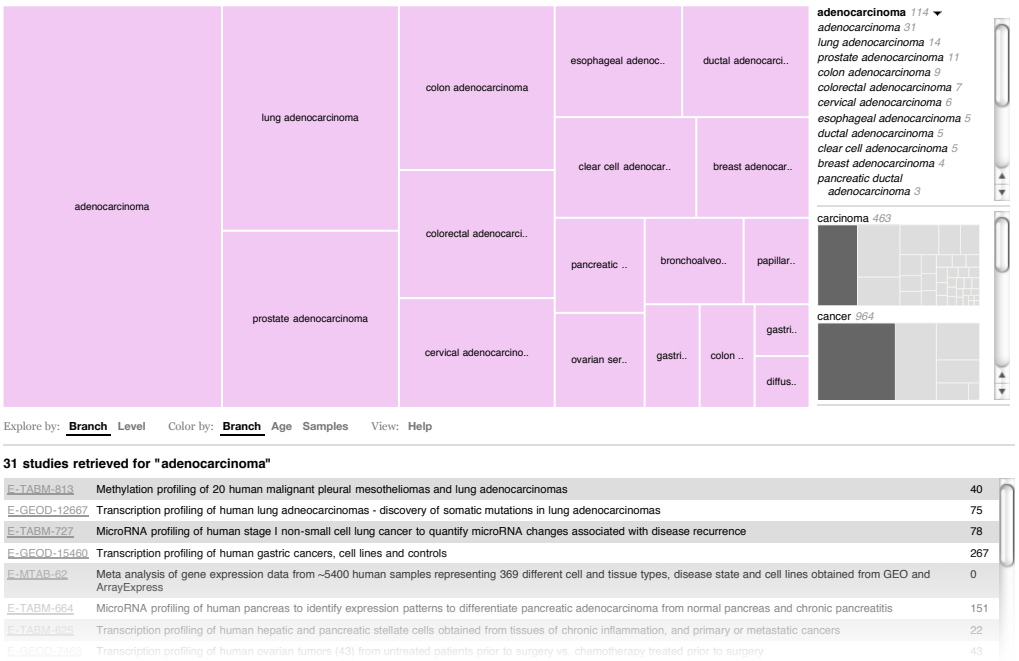

Figure 2.16: Exploration of the "adenocarcinoma" branch in *branch mode*. The navigation bar child term list on the right is expanded. The leftmost rectangle labelled "adenocarcinoma" is a pseudo leaf inserted to represent studies associated directly with the internal node "adenocarinoma". The query results table shows the studies associated with "adenocarcinoma". *(Source: www.ebi.ac.uk/~nils/aex, retrieved on 6 July 2010)*

2.3.4 Level Mode

The level mode provides a visualization of the overall ontology tree and available studies at a chosen depth level. Figures 2.17 through 2.20 show the tree at various depths. The investigator can visually compare the number of studies available for different branches of the ontology. It is also possible to determine which parts of the ontology are less structured (e.g. cell line) and which ones are more structured (e.g. organism part).

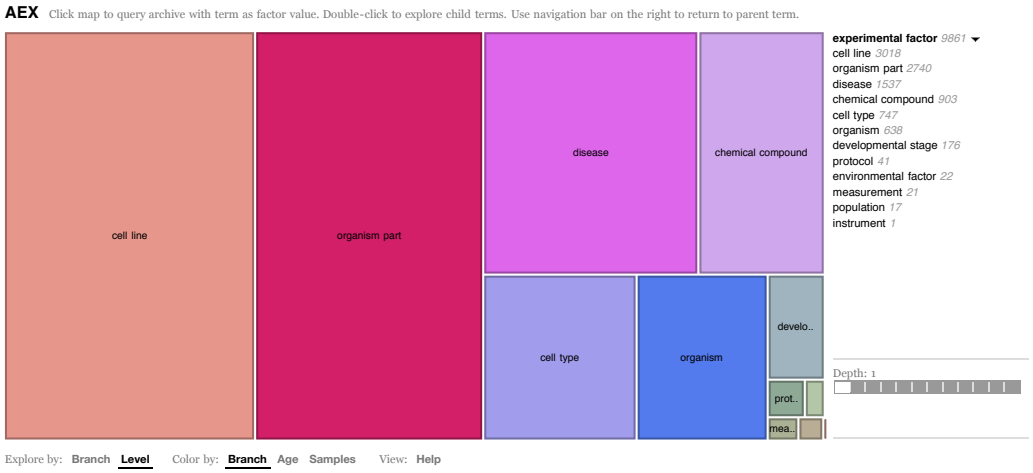

Figure 2.17: ArrayExpress Explorer showing the Experimental Factor Ontology at depth 1. *(Source: www.ebi.ac.uk/~nils/aex, retrieved on 6 July 2010)*

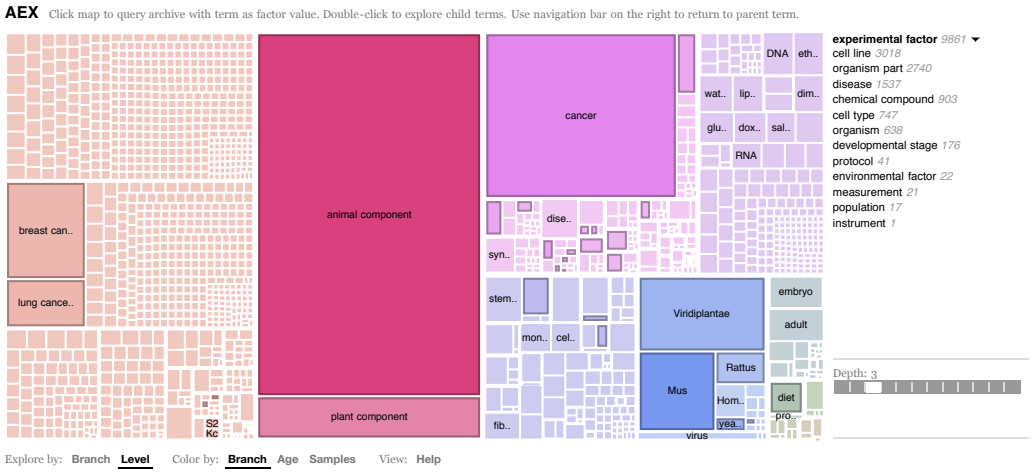

Figure 2.18: ArrayExpress Explorer showing the Experimental Factor Ontology at depth 3. *(Source: www.ebi.ac.uk/~nils/aex, retrieved on 6 July 2010)*

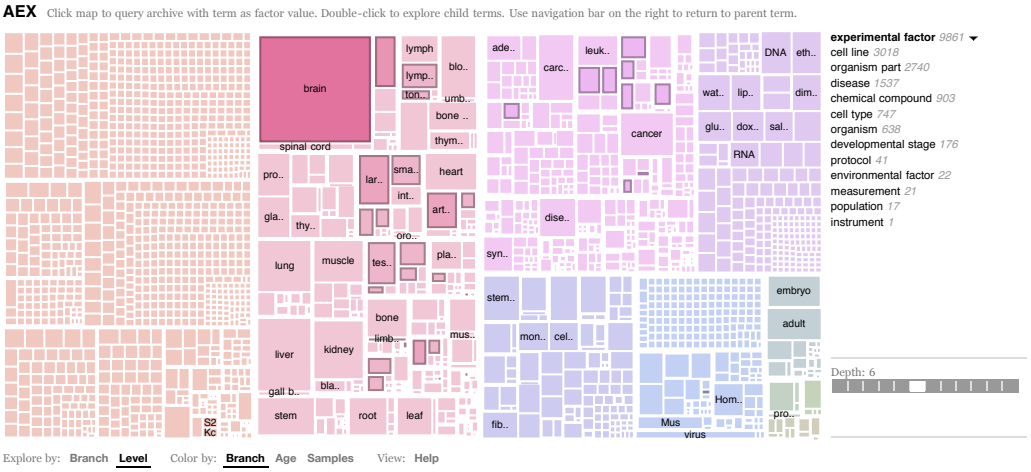

Figure 2.19: ArrayExpress Explorer showing the Experimental Factor Ontology at depth 6. *(Source: www.ebi.ac.uk/~nils/aex, retrieved on 6 July 2010)*

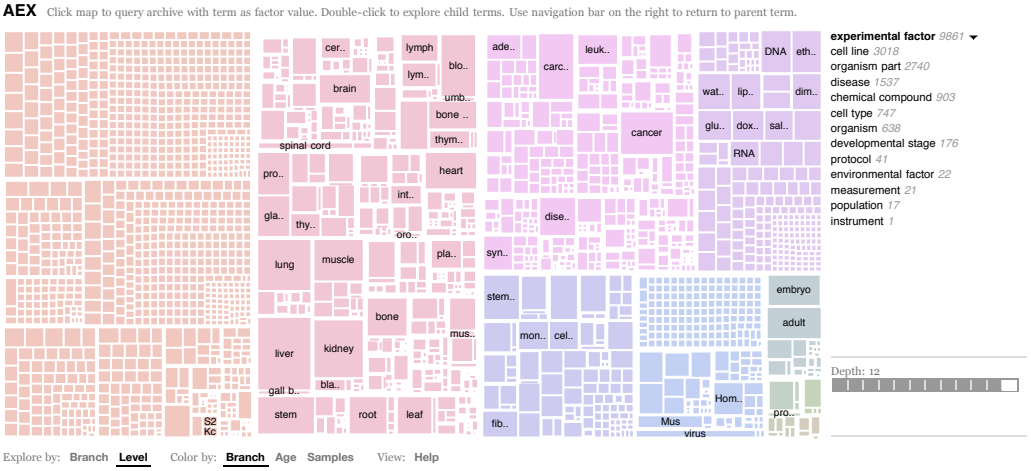

Figure 2.20: ArrayExpress Explorer showing the Experimental Factor Ontology at depth 12. All shown terms are leaves of the ontology. *(Source: www.ebi.ac.uk/~nils/aex, retrieved on 6 July 2010)*

2.3.5 Visualizing Quantitative Variables

In order to visualize a quantitative variable associated with the ontology terms in addition to the number of studies the color of the rectangles in the tree map can be used as discussed in Section 2.2.

Figure 2.21 shows a visualization of the median number of samples in the studies in the ArrayExpress Archive across the whole ontology. This visualization reveals an interesting pattern that is unlikely to be discovered with the keyword-based query interface: The many small dark green rectangles on the left side of the map and in the lower right corner indicate that many cell lines (left) and Arabidopsis strains (lower right corner) have only been studied in a small number of studies but that many samples were analyzed in these studies. Figure 2.18 on page 54 shows labels of the areas mentioned here. A similar pattern can be observed for some type of carcinomas (top part of the map, left of the middle), which are often investigated in studies that comprise a large number of samples.

A mapping of the age of the oldest study associated with the ontology terms is shown in Figure 2.22. This visualization shows, for in-

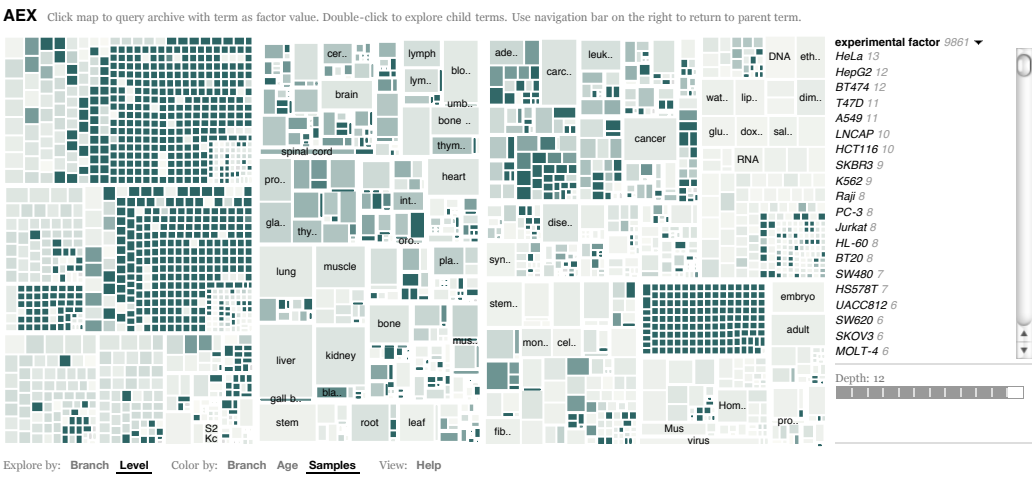

Figure 2.21: ArrayExpress Explorer showing median sample count of the studies associated with each term in the context of the overall ontology and archive content. The darker the color, the more samples are contained in the studies. Counts were cut off at 250 to improve the resolution for terms with lower sample counts. *(Source: www.ebi.ac.uk/~nils/aex, retrieved on 6 July 2010)*

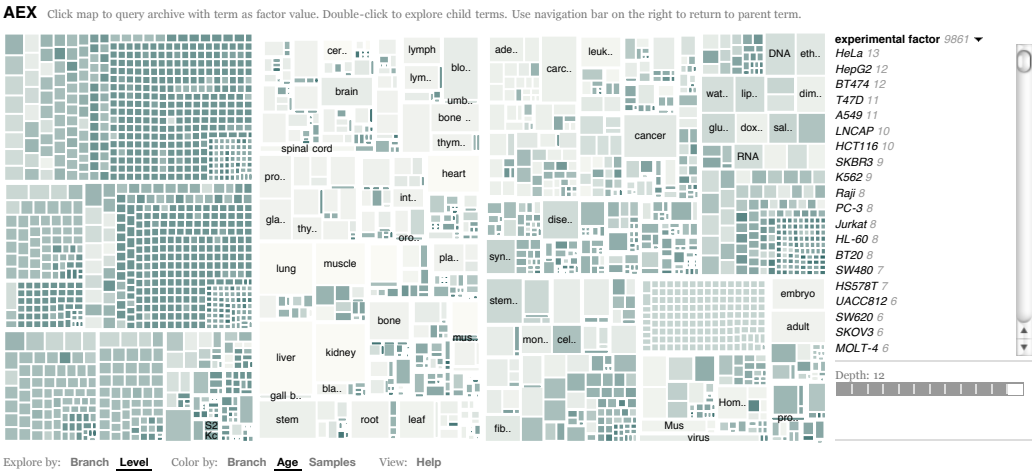

Figure 2.22: ArrayExpress Explorer visualizing time since release of the oldest study associated with each term in the context of the overall ontology and archive content. The darker the color, the more recent the oldest study associated with the term. *(Source: www.ebi.ac.uk/~nils/aex, retrieved on 6 July 2010)*

stance, that the various organism parts were among the first conditions studied with transcriptomics technologies, as indicated by the large block of lightly colored rectangles left to the middle and right to the cell line block.

2.4 Discussion

ArrayExpress Explorer, the exploratory user interface for the ArrayExpress Archive presented here, offers three main contributions. First, it links the Experimental Factor Ontology (EFO) to the content of the archive and uses the ontology structure to organize the content of the archive. Second, through a tree map visualization, the interface provides a graphical view on the ontology that can be used to browse the content of the archive and to obtain an overview of the content. Third, by mapping additional quantitative variables derived from the data sets to the ontology and visualizing these in the tree map, global patterns relating to the growth of the archive become visible.

Using the ArrayExpress Explorer to traverse the ontology tree and

data set associated with the ontology terms conveys the scope and depth of the content of the ArrayExpress archive. Investigators will become more familiar with the content of the archive using the ArrayExpress Explorer, than by using the existing keyword-based query interface. This helps them to put the retrieved data sets in context and to discover areas of the archive that they could have missed using keyword search.

However, there are also some disadvantages to the approach implemented in the ArrayExpress Explorer. For example, by using the EFO as the central structure, some studies may not be accessible through the interface because their experimental factor values are annotated with labels that do not map to the EFO. An inverse problem is that studies usually map to more than one ontology term and large meta data sets then can skew the distribution of study counts significantly. Related to this issue is the string matching of ontology terms to experimental factor labels described in Section 2.2, which leads to many false positive hits and should be refined in the future.

Several studies in the human-computer interaction field have examined the advantages and disadvantages of browsing versus keyword-based search. Hearst (2009, see Chapter 3) reports that the general consensus is that browsing and keyword-based search should be used in combination, because both approaches have strengths and weaknesses, which depend on the given tasks. It seems very reasonable to compare the existing ArrayExpress Archive user interface with ArrayExpress Explorer in a user study to identify which interface performs better for which tasks. The goal of this exercise would be to design an interface that integrates both keyword-based querying and visual browsing.

The current implementation of the ArrayExpress Explorer could be extended in several ways. A straightforward extension that would further support investigators in exploring the ArrayExpress Archive is a keyword-based search for ontology terms and direct access to the tree map view and data sets associated with the retrieved term. Furthermore, additional quantitative variables such as quality scores derived from the

studies could be mapped to the tree map. A related addition is a feature that would allow investigators to query the ArrayExpress Atlas (Kapushesky et al., 2010) with a gene name to identify conditions under which this gene is differentially expressed. These results could then be visualized on the tree map to provide a comprehensive overview of the expression pattern of the query gene. A similar visualization tool was developed by Baehrecke et al. (2004), who used a tree map of the Gene Ontology to visualize expression levels.

In principle, the approach implemented in the ArrayExpress Explorer can also be applied to support exploration of other collections of biomedical data sets. Instead of the EFO other appropriate hierarchical structures can be used to organize the content. For instance, the descriptor hierarchy provided by the Medical Subject Headings (MeSH; www.nlm.nih.gov/mesh) would be an appropriate choice in some cases.

Contributions and Publications

The web services in the ArrayExpress back-end used in the ontology-guided exploration approach described in this chapter were implemented by Nikolay Kolesnikov.

Chapter 3

Probabilistic Retrieval and Visualization of Data Sets

3.1 Introduction

In Section 2.1 the difference between knowledge- and data-driven exploration of repositories was discussed and a knowledge-driven approach to explore the content of the ArrayExpress Archive was introduced in Chapter 2. While such approaches are useful to help investigators to locate data sets that might be relevant for their work, the meta information associated with studies in the ArrayExpress and other repositories represents hardly any biology that is observed in these studies. To address the limitations of knowledge-driven methods, novel data-driven approaches are introduced in this chapter that perform retrieval based on patterns observed in the expression data.

The two approaches presented here have been developed to explore the content of the ArrayExpress Archive and go beyond basic lookup (Marchionini, 2006). Furthermore, the methods are designed to enable interpretation of retrieval results by placing retrieved studies in the overall context of the repository.

Simple content-based search, where the queries are expression pro-

files from one study and the set of most similar studies is retrieved have been described for instance by Fujibuchi et al. (2007) and by Hunter et al. (2001). The key problem that needs to be addressed is how to choose the distance measure based on which the similarity of the expression profiles will be assessed.

This retrieval problem is related to the suggestion that analysis of a new data set can benefit greatly from placing it in the context of all earlier data sets (Tanay et al., 2005). In the study by Tanay et al. (2005) the authors developed a method for extracting a set of biclusters from earlier studies and evaluated the activity of those biclusters in a new experiment. In another holistic analysis paper by Segal et al. (2004), a "module map" of gene modules versus clinical conditions was formed by first finding differentially expressed gene sets, then combining them into modules, and finally identifying modules differentially expressed over a set of arrays having the same annotation. A similar tool is the more recent *Connectivity Map* developed by Lamb et al. (2006), which relates diseases and chemicals via common gene expression profiles. These approaches can be extended by incorporating additional biological knowledge into the underlying model, for instance in the form of regulatory networks, partly assumed and partly learned from data. As a consequence, the computational complexity will increase accordingly.

The goal of the work presented in this chapter is to take the idea of extracting information about biological processes from a collection of transcriptomics data sets and to use it in the search process to focus the search on biologically relevant patterns. In this approach, two data sets are defined similar to each other when similar biological processes, for instance represented by biochemical pathways, are activated in both of them. Retrieval is based on this similarity and the system is queried with activation patterns of biological processes that are typically extracted from a query data set. Since the search is data-driven unexpected patterns can be discovered and the system complements knowledge-driven approaches that are based on meta information. Moreover, the focus of this work is on

patterns that arise from differentially expressed genes as a result of the experimental setup. Additionally, the models applied to the collection of data sets are reasonably simple in order to keep the searches scalable, but they still are able to extract relevant patterns.

The approach presented here consists of three key components. These components are designed to ensure that the retrieved experiments are relevant in the sense that related biological processes are affected in both the retrieved data sets and in the query data set and that these changes can be attributed to experimental factor values of the corresponding studies. The components are (1) a model for the activity of biological processes across the collection of data sets, which makes the different experiments and data types stored in the database commensurable, (2) a method for performing searches given the model with a data set as the query, and (3) techniques for visualization of the search results and the model.

The model is designed both to incorporate prior knowledge about biological processes and to derive patterns from the data, while keeping the computational load manageable. Prior knowledge about biological processes is provided by gene sets extracted from earlier analyses or databases, which have also been used in some of the earlier holistic analyses. The gene sets are incorporated into the approach by using Gene Set Enrichment Analysis (Subramanian et al., 2005) in a new way. Each data set, both the query and the all others contained in the collection, are encoded as a vector containing the number of differentially expressed genes in each gene set. These vectors can be considered to be a *bag-of-words* representation of the experiments, where the gene sets are the words and the number of differentially expressed genes in the gene sets are weights for the gene sets, similar to word counts in text documents. The analogy is described in more detail in Section 3.2.2. This step makes the different experiments commensurable. Moreover, when the differential expression is measured between an experimental variable and a control, the encoding focuses on the patterns caused by the experimental factors, which pro-

vides a starting point for the interpretation of the results.

The model used in the approach is known as a *topic model* (Blei et al., 2003b) or *discrete principal component analysis* (Buntine and Jakulin, 2004), which has been applied successfully in textual information retrieval. In bioinformatics, topic models have been used, for instance, to identify components in haploinsufficiency profiling data (Flaherty et al., 2005) and in discretized gene expression data (Gerber et al., 2007). They are suitable for finding latent components from count data, such as text documents represented as *bags-of-words*. Since they are probabilistic models, they can infer the underlying components while taking the uncertainty in the data into account. In order to deal with transcriptomics data, counts of words are changed to counts of differentially expressed genes in gene sets with words corresponding to gene sets. Each data sets thus corresponds to an activity profile over the latent components of the model, and each component corresponds to a distribution over the gene sets. The differences from previous applications of topic models to discretized genomic data (Flaherty et al., 2005; Gerber et al., 2007) are the use of gene sets to bring in biological knowledge, focusing to effects elicited by the experimental factors, and the application to retrieval and exploration.

Given a topic model, there are well-justified methods for doing information retrieval (Buntine and Jakulin, 2004; Griffiths and Steyvers, 2004) for texts, where the query is simply another document. The same principles apply here for querying with a new data set. To visualize both the components for interpretation of the biological findings and the retrieval results to browse the collection existing methods (Venna and Kaski, 2007a) are applied and new ones introduced.

3.2 Methods and Data

3.2.1 Collection of Data Sets

A total of 288 human transcriptomics data sets that contained a preprocessed and normalized gene expression matrix were obtained from the ArrayExpress Archive (Parkinson et al., 2009). These data sets are also part of the ArrayExpress Atlas and have been specifically curated for meta analysis and contain high quality data. As described in Section 1.1.4, each data set is associated with a collection of experimental factors describing the variables under study, e.g. "disease state" or "gender". Each microarray in a data set has a specific value for each of the experimental factors, e.g. "disease state = normal" and "gender = male".

From the collection, a subset of data sets was obtained that have the experimental factor "disease state". The data sets in this study were decomposed into so-called *comparisons*, of control samples against samples representing a particular pathology. This yielded a total of 105 comparisons that included a wide range of pathologies such as different cancer types, as well as neurological, respiratory, digestive, infectious, and muscular diseases. The only significantly frequent broad category was cancer, with 27 comparisons.

The remaining experiments in the data set were also systematically decomposed into binary comparisons. For each experimental factor in an experiment, either two values of that experimental factor (e.g. disease A vs. disease B), or one value versus all others (e.g. control against all treatments) were chosen for the comparisons. In experiments with more than one experimental factor, the factors whose values are not being compared provide a *context* for the comparison. For example, when comparing two values of "disease state", e.g. "normal" vs. "cancer", different comparisons for "gender = male" and for "gender = female" can be obtained.

For each comparison, all possible combinations of contextual factors were generated. Only comparisons that had at least 6 microarrays assigned to each phenotype were kept. Feature identifiers used for the

probes on the microarrays were mapped to HUGO gene symbols (Eyre et al., 2006) and multiple measurements for a single gene were collapsed using the median. The total number of obtained comparisons, including the 105 "control versus disease" comparisons mentioned above, was 768.

3.2.2 Gene Set Enrichment Analysis

Gene Set Enrichment Analysis (GSEA) tests if a set of genes is coordinately associated with the difference between two phenotypes in a microarray experiment. The method has been used extensively in recent years and only a brief description is provided here. Full details and a description of the software that was used for described here can be found in the original papers by Mootha et al. (2003) and Subramanian et al. (2005).

As illustrated in Figure 3.1, GSEA starts by computing a ranked list of the genes in the experiment, according to how well each gene discriminates between the two phenotypes. This can be achieved by using metrics such as fold change or signal-to-noise ratio. Then, a weighted Kolmogorov-Smirnov (KS) running statistic, called the *enrichment score* (ES), is computed over the list. To compute the ES, a ranked list of all genes is created based on their correlation with one of the phenotypes. For every position in this ranked list, a score representing the difference between the number of genes in the gene set, weighted by their rank, and the number of genes not in the gene set, up to the corresponding position is computed. The maximum of these scores is the ES for a given gene set. The detailed formula for the computation of the ES is provided in Subramanian et al. (2005).

After normalization the ES is used to compute significance measures such as the false discovery rate (FDR) q-value[1] (Storey, 2003). The computation of the statistic also produces a subset of the genes in the set.

[1]The False Discovery Rate (FDR) is a measure used to control for multiple comparisons in hypothesis testing. The FDR is the expected fraction of false positives among all rejected null hypotheses. The q-value measures the significance of particular feature in terms of the FDR and is analogous to a p-value, which measures significance in terms the false positive rate.

This subset, called the *Leading Edge Subset*, constitutes what could be considered the *core* of the gene set in the given comparison.

As mentioned before, GSEA is used in this approach to include biological knowledge in the form of the pre-defined gene sets. Furthermore, the differential expression within each set is quantified as a count. In brief, a *comparison* derived from a data sets essentially consists of a collection of samples that are divided into two groups defined by distinct phenotypes. These phenotypes are designated by A and B, respectively. In order to assess which gene sets are differentially expressed in either of the two phenotypes, GSEA is run for both the comparison $A \rightarrow B$ and the comparison $B \rightarrow A$. The gene sets used in this approach are taken from the *Molecular Signatures Database* (www.broadinstitute.org/gsea/msigdb/) (Subramanian et al., 2005) and are limited to the collection of canonical, manually compiled pathways (collection C2-CP). The results from both GSEA runs ($A \rightarrow B$, $B \rightarrow A$) for each data set are combined and gene sets sorted according to the magnitude of their *normalized enrichment score* (NES), which is the ES normalized to account for the size of the gene set. The 50 gene sets with the highest absolute NES are selected for further processing. This choice is motivated by previous observations that often gene sets, which do not reach a standard FDR q-value of 0.25, are still effectively relevant to the condition under study, and that these are generally consistent among laboratories conducting similar microarray experiments (Subramanian et al., 2005). Finally, the size of the leading edge subset of each of the 50 gene sets with the highest NES is obtained as a count.

By running the above procedure for every comparison derived from the collection of data sets generates a list of significant gene sets, each associated with an integer value (the size of the leading edge subset for that particular comparison). This representation can be considered analogous to the so-called *bag-of-words* representation for text documents. In textual information retrieval, it is common to represent a document by how many times each word in the vocabulary appears in that document. The order of the words, and thus all grammar, is therefore omitted, hence

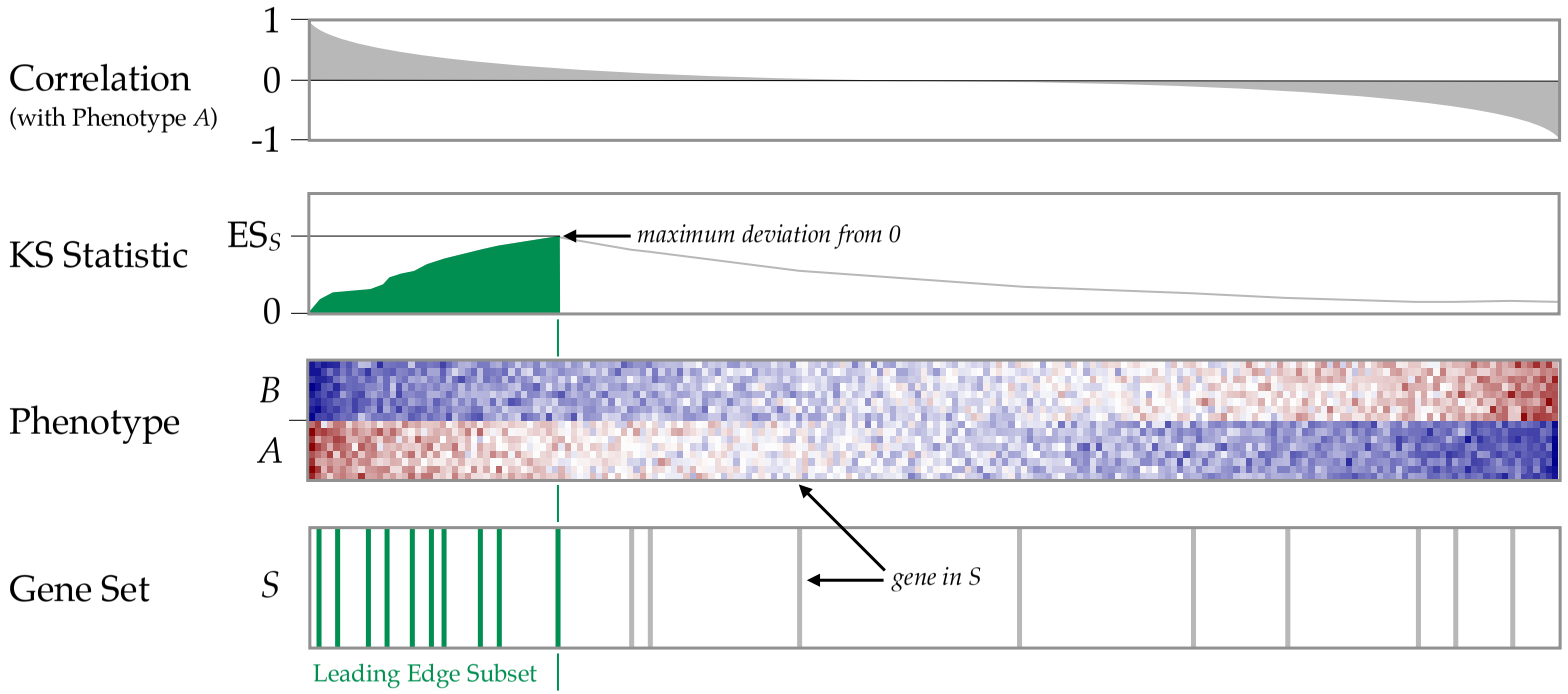

Figure 3.1: Schematic overview of Gene Set Enrichment Analysis after Subramanian et al. (2005). In the heatmap of the expression matrix the rows are samples associated with either phenotype A or B, and the columns are genes. The genes are sorted according to their correlation with phenotype A.

the name "bag-of-words". The procedure described above effectively generates a bag-of-words representation for each comparison derived from a data set. As a result, conceptually each comparison can be regarded as a document having several words from a vocabulary. In the context of transcriptomics data, the vocabulary is the collection of canonical pathways, and each gene set found to be significant is a word.

In essence, the above procedure generates a representation of differential expression that is amenable to probabilistic modeling with topic models, and for topic model-based information retrieval tools.

3.2.3 Topic Models

Topic models are probabilistic unsupervised models for finding latent components in document collections. Topic models are also known as Latent Dirichlet Allocation (LDA; Blei et al., 2003b) or discrete Principal Component Analysis (dPCA; Buntine and Jakulin, 2004). Provided a collection of documents in a bag-of-words representation, it models each document as a probability distribution over so-called *topics*. A topic, the central concept, is itself a probability distribution over words in the vocabulary of the collection. The model is a *generative hierarchical model*, which can be specified by formulating the generative process from which the data are assumed to arise. More formally, the generative process is defined by the distribution over topics for each document d, and the distribution over words for each topic t, which are specified, respectively, by the parameters of a hierarchical model, which are the random variables $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_t$,

$$\begin{aligned} \boldsymbol{\theta}_d &\sim \text{Dirichlet}(\alpha), \\ \boldsymbol{\phi}_t &\sim \text{Dirichlet}(\beta). \end{aligned}$$

The model is illustrated in Figure 3.2. Here α and β are scalar hyperparameters for symmetric Dirichlet probability distributions, which are conjugate priors of the multinomial distribution and regulate the sparsity

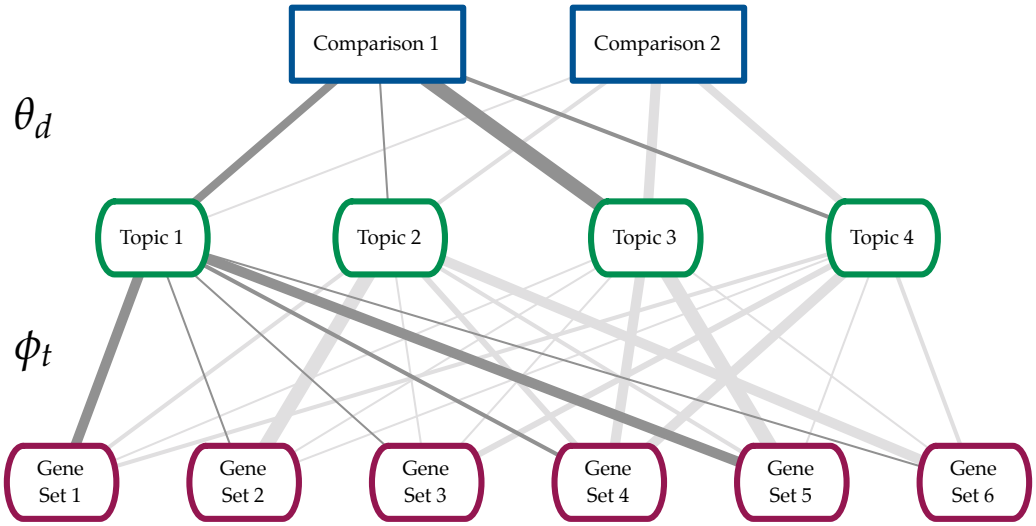

Figure 3.2: An illustration of the structure of the topic model used to model comparisons. Every comparison, or document, in the case of text data, is associated with probability distribution $\boldsymbol{\theta}_d$ over topics. Likewise, every topic is associated with a probability distribution $\boldsymbol{\phi}_t$ over gene sets, or words, in the case of text data. Here the probabilities are indicated by the thickness of the lines. The distributions for Comparison 1 and Topic 1 are highlighted.

of the model. Each word in a document is assumed to come from exactly one topic. For word i in document d, a topic is chosen using the topic probability distribution of the document. This amounts to sampling from a scalar variable $z_{d,i}$,

$$z_{d,i} \mid \boldsymbol{\theta}_d \sim \text{Multinomial}(\boldsymbol{\theta}_d).$$

After choosing a topic $z_{d,i}$, the corresponding word $w_{d,i}$ is sampled from the distribution of the topic over words,

$$w_{d,i} \mid z_{d,i}, \boldsymbol{\phi}_{z_{d,i}} \sim \text{Multinomial}(\boldsymbol{\phi}_{z_{d,i}}).$$

The above definitions correspond to the variant of LDA described by Griffiths and Steyvers (2004).

In the approach described here, topic models are used to model the comparisons that have been processed by GSEA. The relationship to text document modeling is that each comparison is conceptualized as a docu-

ment. In this conceptualization, each word is a gene set, and each topic is a probability distribution over gene sets. Thus topics can be considered to represent biological processes. They specify an ordering on gene sets, with the ordering indicating how likely it is that a gene set is differentially expressed. By considering the top gene sets in a topic, one can obtain a biological picture that is broader and more holistic than the one described by a single gene set. Finally, by having a probability distribution over topics, a comparison effectively assigns different weights to biological processes. In the remainder of this chapter the terms "experiment" and "document", as well as "gene set" and "word", are used interchangeably.

For the model used here the hyperparameters are set to $\alpha = 1$, as in the approach by (Griffiths and Steyvers, 2004), and $\beta = 0.01$. This value for β results in a sparser model and thus a better separation of the topics. The number of topics is fixed at $T = 50$ and was selected by computing the log-likelihood of the data after inferring the model for several values of T. The model is learned using a method described by Griffiths and Steyvers (2004). In that method, so-called *collapsed Gibbs sampling* is applied to find assignments of the words of each document to the topics, by first analytically integrating out the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ to obtain the joint probability of the document collection and the word-to-topic assignments,

$$P(\boldsymbol{w}, \boldsymbol{z}) = \int P(\boldsymbol{w}, \boldsymbol{z} \mid \boldsymbol{\theta}, \boldsymbol{\phi}) P(\boldsymbol{\theta}) P(\boldsymbol{\phi}) \, d\boldsymbol{\theta} \, d\boldsymbol{\phi}.$$

The values of the z are then sampled using a Gibbs sampler from the conditional probability distribution $P(z_{d,i} \mid \boldsymbol{z}_{-(d,i)}, \boldsymbol{w})$, where $\boldsymbol{z}_{-(d,i)}$ is obtained by discarding $z_{d,i}$ from $\boldsymbol{z}$. Sampling is performed iteratively for a total of 2000 scans. On an Intel 1.73GHz Core 2 Duo CPU, this takes approximately 23 minutes. Computations are performed using the Topic Modeling Toolbox for MatLab (psiexp.ss.uci.edu/research/programs_data/toolbox.htm).

The procedure is repeated for a total of eight samplers. Out of the samples, the sample having the highest probability is chosen for interpretation and the parameter values $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are estimated based on the as-

signments of words to the topics.

3.2.4 Probabilistic Search

The topic model represents each comparison as a distribution over topics. It is then natural to measure similarity of comparisons in terms of distances between their distributions over the topics. Suitable distance measures for distributions include the (symmetrized) Kullback-Leibler divergence, Jensen-Shannon divergence or Hellinger distance; unfortunately all of these have problems with sparsity, which necessarily results when the dimensionality is high, as in the case of the model presented here.

For the approach described here similarities were obtained by computing the probability that the gene sets in a query comparison were generated by any another comparison in the collection, which is a more natural way for the kind of probabilistic model used here (Buntine and Jakulin, 2004). In more precise terms, this amounts to computing

$$P(\boldsymbol{w}_q \mid \boldsymbol{\theta}_d) = \prod_{w \in \boldsymbol{w}_q} \sum_{t=1}^{T} \theta_{d,t} \phi_{t,w},$$

where $\boldsymbol{w}_q$ is the collection of gene sets in a query q and T is the number of topics in the model. The equation states that for each word in the query the overall probability that it was generated by any topic, given the topic proportions in the potentially relevant experiment, is computed. By repeating the same query for all comparisons derived from the collection, a ranked list is obtained that is ordered by the relevance of each comparison to that query, which is the most straightforward way to retrieve comparisons with a given comparison as a query. The computation of all possible queries takes less than 5 seconds on an Intel 1.73GHz Core 2 Duo CPU.

3.2.5 Visualization of the Relationship between Comparisons, Topics and Gene Sets

Visualization of the topic model is essential for interpretation of the biological findings of the analysis. One goal of the described approach is to gain insight into the structure of the collection of data sets and the biological processes recorded in it. In order to do so, it has to be possible to examine the topic composition of the comparisons, as well as the gene set composition of the topics.

The results obtained from GSEA and the topic model are essentially two matrices P_t and P_g containing the topic probabilities across the comparisons and the gene set probabilities across the topics. The connection between P_t and P_g are the topics. Accordingly, the matrices can be considered a disjoint union of two complete bipartite graphs, where the probabilities in the matrix represent edge weights. The resulting graph can be laid out by placing the nodes for experiments, topics and gene sets in three parallel rows, where the middle row contains the nodes for the topics and is shared by the two subgraphs. This is shown on a small scale in the illustration of the topic model in Figure 3.2.

For the visualization a subset of edges is selected, since the two bipartite graphs are complete. Rather than making a hard selection, a reduced line width and color opacity of the edges based on the corresponding weights are used. With this strategy, the edges representing a high probability are emphasized and those standing for lower probabilities are essentially removed.

Each topic is assigned a distinct color from a discrete rainbow color scheme and all edges connected to the topic are drawn in this color. This makes it easier for the investigator to follow the edges from the topic to the corresponding comparisons or gene sets. At the same time, the links having a particular color are easily distinguished and provide an overview for the interpretation of that particular topic, in terms of its distribution over both gene sets and comparisons where this topic plays a role.

Clutter is reduced by reordering gene sets and topics so that the number of intersecting edges is low. A suitable heuristic for achieving this is to compute a complete linkage hierarchical clustering of the gene sets and of the experiments to obtain a partial ordering for both. As a distance measure the symmetrized Kullback-Leibler divergence between the corresponding distributions is used. Furthermore, the topics are sorted by the index of the maximum value in the corresponding column of P_g. Additionally, Bézier curves instead of straight lines are used to connect topics with comparisons and gene sets. The Bézier curves form edge bundles, which further reduces clutter. In order to increase the space available to plot comparison and gene set names they are arranged along the radius of a circle instead of along a straight line.

Figure 3.3 shows the resulting visualization, which was created with a custom software tool written in *Processing* (www.processing.org) that generates PDF output. The complete visualization is readable on an interactive display that supports zooming and panning. However, to keep the visualization readable also on paper a subset of topics, for which the sum of probabilities given the documents is the highest, is selected. In detail, the top 10 topics in the subset of the 105 main comparisons of control versus a disease and the top 10 topics in the complete data set are selected and the union of these two sets is computed. This results in a set of 13 topics. Additionally, the number of gene sets is reduced in the visualization by choosing the 25 most probable gene sets for each topic, and taking the union of all these topics. This is justified by the observation that the probabilities for gene sets in the topics typically level off beyond the top 25 gene sets. This results in 211 gene sets for the visualization of the 13 selected topics.

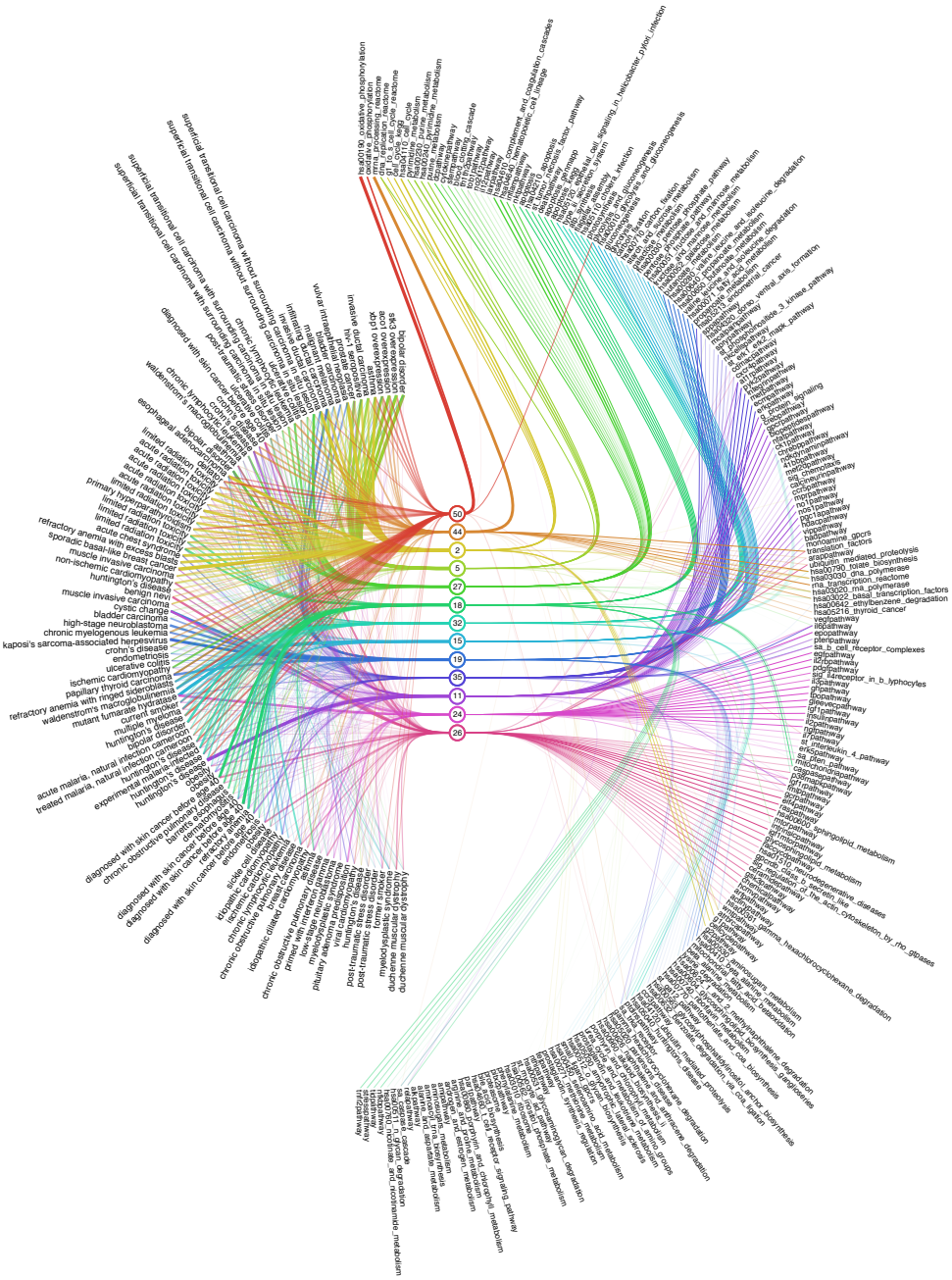

Figure 3.3: Visualization of the topic model. A subset of 13 topics (center column), 211 gene sets (right) and 105 comparisons (left) is shown. For details and a discussion see the text.

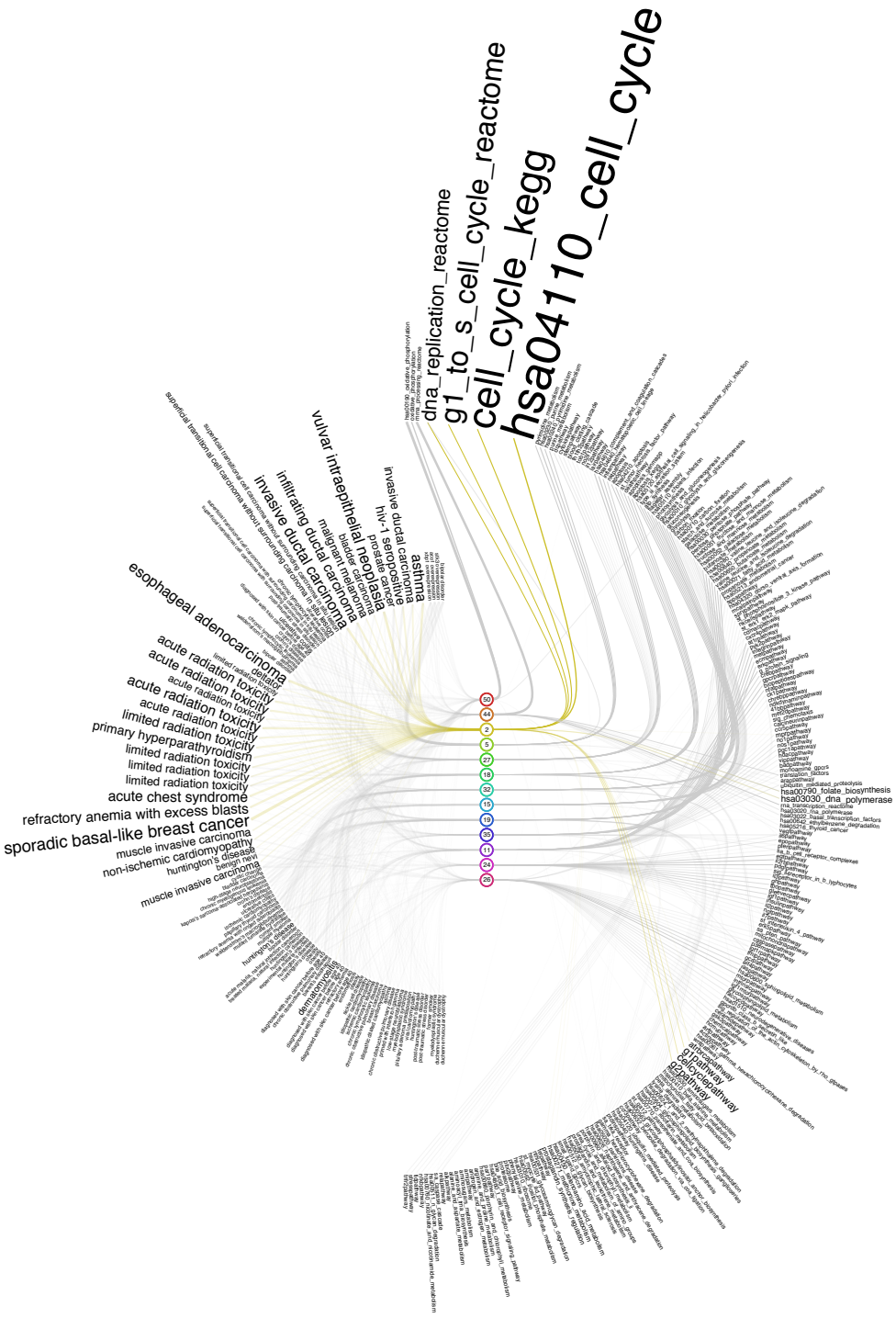

Figure 3.4: Topic 2 highlighted in the visualization of the topic model. Labels for comparisons are scaled according to their probability for the highlighted topic and gene set labels are scaled according to their probability in the topic probability distribution.

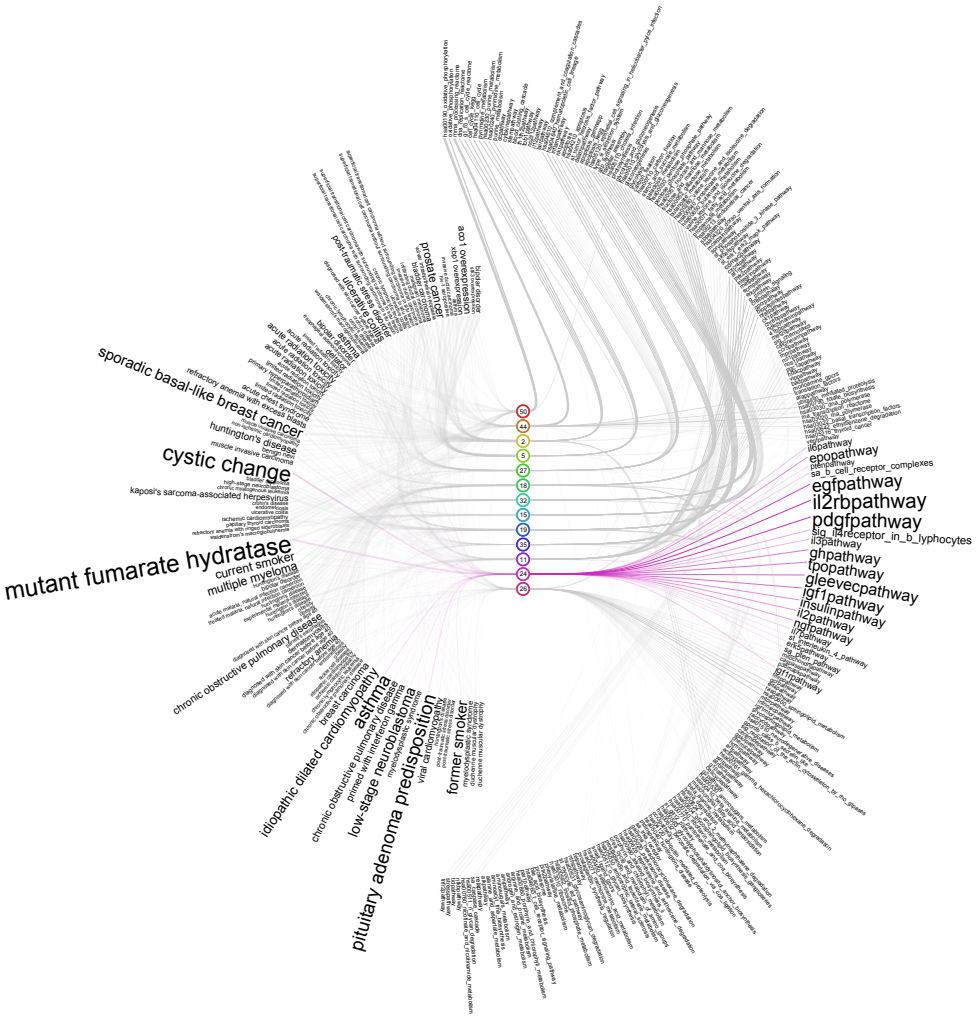

Figure 3.5: Topic 24 highlighted in the visualization of the topic model. Labels for comparisons are scaled according to their probability for the highlighted topic and gene set labels are scaled according to their probability in the topic probability distribution.

3.2.6 Visualizing Retrieval Results

To complement the standard ranked lists, retrieval results can be presented on a projection display showing all comparisons. Assuming the projection is good, the display is useful in putting the retrieval result into the context of the overall collection of comparisons. Clusters and outliers in the retrieval results become evident, results of different queries can be easily compared and the whole collection can be interactively browsed while simultaneously seeing the retrieval results.

To visualize retrieval results, all experiments are projected onto a 2-dimensional display using a recent projection method that has been shown to outperform alternative methods in the task of retrieving similar data points (here: comparisons) given the display. The method is called Neighbor Retrieval Visualizer (NeRV; Venna and Kaski, 2007a) and has been developed specifically for visualizing data in retrieval tasks and for explorative information visualization. NeRV needs to be given the relative cost of misses and false positives of the true similarities between the data points. In this application false positives are penalized, resulting in a display that is trustworthy in the sense that two points that are similar in the visualization can be trusted to have also been similar before the projection.

Like other multidimensional scaling methods, NeRV starts with a pairwise distance matrix between all experiments. Here the symmetrized Kullback-Leibler divergences between the topic distributions of the documents are used. The pure projection of the comparisons only shows their relative similarity, and for further interpretation the display needs to be coupled with the topic content of the comparisons. It is possible to include this important information by including glyphs in the projections to represent the distribution of topics (Yang et al., 2007). Including the glyphs has the additional advantage that since a nonlinear projection of a large data set in a two-dimensional space cannot preserve all similarities, the imperfections are detectable based on the glyphs.

The glyphs are designed to represent the probability distribution over the topics of a comparison by dividing a square into vertical slices

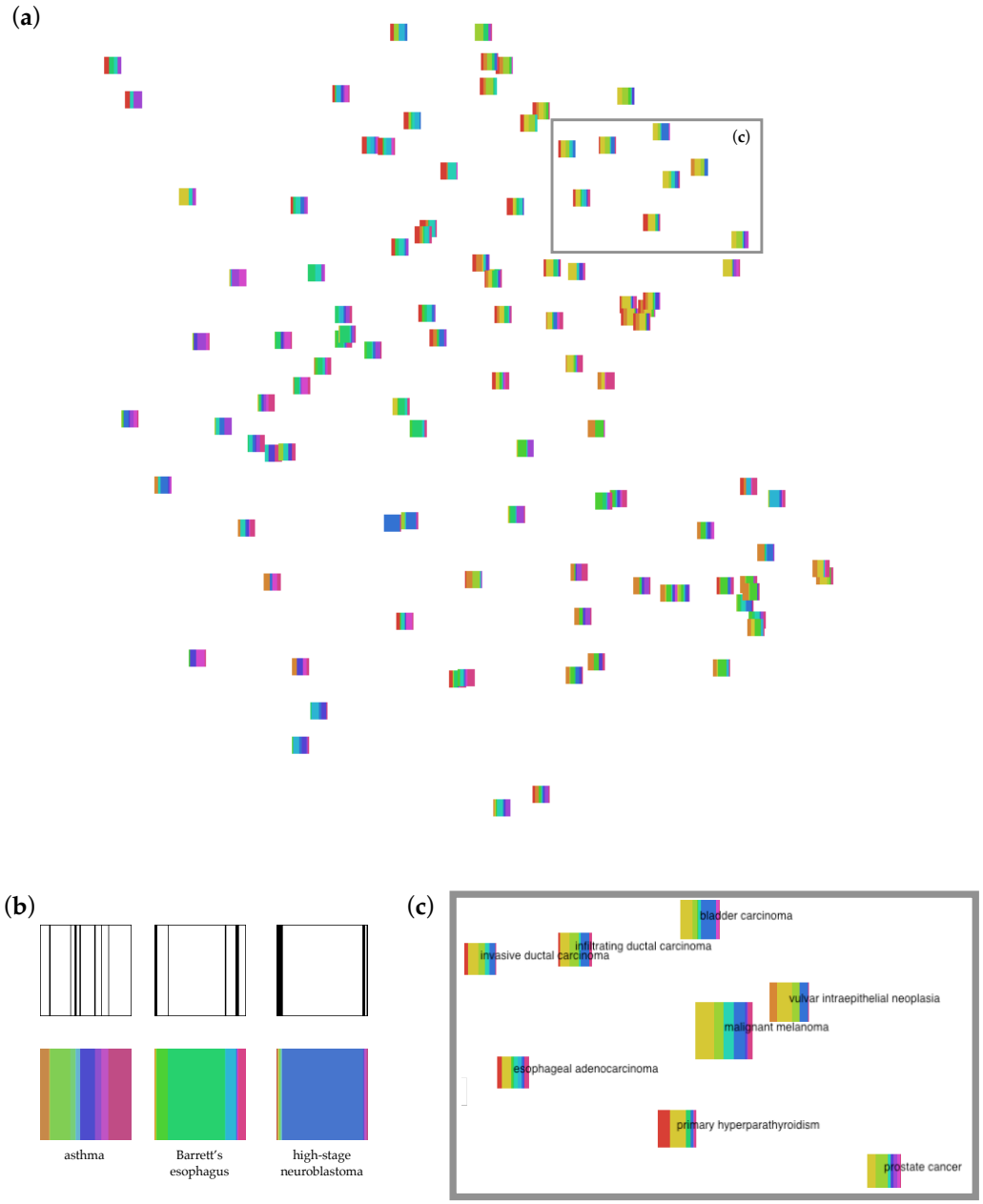

Figure 3.6: The collection of 105 comparisons visualized as glyphs on a plane. Topic colors in all glyphs match topic colors in Figure 3.3. (**a**) NeRV projection of the 105 experiments, each shown as a glyph. (**b**) The slices of each glyph show the distribution of topics in the experiment. (**c**) Enlarged region from (**a**) where glyphs have additionally been scaled according to their relevance to the query with the "malignant melanoma" comparison shown in the center. A detailed description of this experiment is included in Section 3.3.

that each represent a topic. The width of the slice represents the probability of the topic. This is illustrated in Figure 3.6 (**b**) in the top row. While this is sufficient for comparing the shape of the probability distributions of documents, the slices are also colored with a distinct color representing the topic, as shown in Figure 3.6 (**b**) in the bottom row. The coloring has the additional distinctive purpose that it connects the topics of the glyphs visually with the same topics in the display of Figure 3.3, which can then be used for further interpretation.

3.3 Results

3.3.1 Inferred Topics

By analyzing the most probable gene sets for each topic, its underlying biological theme can be inferred. The most probable gene sets in most of the topics learned by the model are coherent, and the topics taken together describe a wide range of processes. The analysis here is focused on the same most prominent topics shown in the visualizations, which were selected based on the sum of probabilities over all comparisons as described in Section 3.2.5. The top five gene sets for each of the 13 topics are shown in Table 3.1.

The topics are related to diverse themes such as cell cycle (Topic 2), DNA replication (Topics 2 and 5), organic compound metabolism (Topics 5 and 19), G protein signaling (Topic 11) glycolysis (Topic 15), apoptosis (Topic 18), cell growth and proliferation (Topics 24 and 26), cell differentiation (Topic 27), infection (Topic 32), cell communication (Topic 35), gene expression (Topic 44), and oxidative phosphorylation (Topic 50).

In some topics, some of the top gene sets are almost identical. This stems from the fact that those gene sets are highly overlapping and are therefore being put into similar topics with similar probabilities.

Although Table 3.1 is illustrative of the variety of topics found by the model, understanding each topic requires looking beyond the top five

Table 3.1: Top five gene sets for the 13 most probable topics. An acronym for the source of the gene set was included either to distinguish between gene sets with similar names, or when the name of the gene set already includes a mention of that source, i.e. KEGG (Kanehisa and Goto, 2000), GenMAPP (Salomonis et al., 2007), BioCarta (www.biocarta.com), or Reactome (Vastrik et al., 2007).

Topic 2	**Topic 5**	**Topic 11**
Cell Cycle (BIOCARTA)	Purine Metabolism (KEGG)	G Protein Signaling
Cell Cycle (KEGG)	Pyrimidine Metabolism (KEGG)	Biopeptides Pathway
G1 to S Cell Cycle (REACTOME)	Purine Metabolism (GENMAPP)	NFAT Pathway
DNA Replication (REACTOME)	Pyrimidine Metabolism (GENMAPP)	CREB Pathway
G2 Pathway	DNA Replication (REACTOME)	GPCR Pathway
Topic 15	**Topic 18**	**Topic 19**
Gluconeogenesis	Apoptosis (GENMAPP 1)	Valine Leucine and Isoleucine Degradation
Glycolysis	Apoptosis (KEGG)	Propanoate Metabolism (KEGG)
Glycolysis and Gluconeogenesis (KEGG)	Apoptosis (GENMAPP 2)	Fatty Acid Metabolism
Glycolysis and Gluconeogenesis (GENMAPP)	Apoptosis (GENMAPP 3)	Propanoate Metabolism (GENMAPP)
Fructose and Mannose Metabolism	Death Pathway	Valine Leucine and Isoleucine Degradation
Topic 24	**Topic 26**	**Topic 27**
IL2RB Pathway	mTOR Pathway	Hematopoietic Cell Lineage
PDGF Pathway	Sphingolipid Metabolism	Complement and Coagulation Cascades
EGF Pathway	eIF4 Pathway	Inflammation Pathway
Gleevec Pathway	RAS Pathway	NKT Pathway
IGF-1 Pathway	IGF-1 mTOR Pathway	Dendritic Cell Pathway
Topic 32	**Topic 35**	**Topic 44**
Epithelial Cell Signaling in *H. pylori* Infection	Integrin Pathway	mRNA Processing (REACTOME)
Cholera Infection (KEGG)	Met Pathway	RNA Transcription (REACTOME)
Photosynthesis	ERK Pathway	Translation Factors
ATP Synthesis	AT1R Pathway	Folate Biosynthesis
Flagellar Assembly	ECM Pathway	Basal Transcription Factors
	Topic 50	
	Oxidative Phosphorylation (KEGG)	
	Oxidative Phosphorylation (GENMAPP)	
	Glycolysis and Gluconeogenesis	
	IL-7 Pathway	
	Gamma Hexachlorocyclohexane Degradation	

gene sets. For instance, in Topic 2, gene sets until the eighth position[2] are not very informative about the process the topic is representing, beyond the fact that it is related to cell cycle and DNA replication. However, the gene set at the ninth position, "ATR BRCA Pathway", contains a signaling system that includes genes BRCA1 (breast cancer 1, early onset) and BRCA2 (breast cancer 2, early onset). These genes are involved in the cellular response to DNA damage, and their mutations have been found to increase breast cancer susceptibility (Tutt and Ashworth, 2002). Investigation of the comparisons with the highest probability for this topic reveals that the top four comparisons are cancer-related: normal tissue versus sporadic basal-like breast cancer, vulvar intraepithelial neoplasia, breast carcinoma, and esophageal carcinoma. As the only two breast cancer experiments in the data set appear among those four top experiments, these results indicate that Topic 2 has relevance not only for cell cycle and DNA replication, but also for breast cancer.

Another interesting example is found in the top gene sets of Topic 44. One of the gene sets corresponds to genes involved in folate biosynthesis. Folate has an important role in DNA and RNA synthesis, and low folate levels are known to promote a number of pathologies (Glynn and Albanes, 1994; Au et al., 2009; Hoffbrand et al., 1968). Computing the comparisons with the highest probability for this topic reveals that the top four results pertained to comparisons between normal tissue versus Crohn's disease, chronic lymphocytic leukemia, and chronic myelogenous leukemia, as well as a comparison between patients with normal tissue and cancer patients with acute radiation toxicity. Folate deficiency has been observed both in patients with Crohn's disease (Hoffbrand et al., 1968) and in patients with leukemia (Au et al., 2009). Once again, the model assigned comparisons to meaningful topics and, moreover, is able

[2]The ranking of the top ten gene sets for Topic 2 is: 1. Cell Cycle (BIOCARTA), 2. Cell Cycle (KEGG), 3. G1 to S Cell Cycle (REACTOME), 4. DNA Replication (REACTOME), 5. G2 Pathway, 6. Cell Cycle Pathway, 7. G1 Pathway, 8. DNA Polymerase (BIOCARTA), 9. ATR BRCA Pathway and 10. Folate Biosynthesis. This is also visible in Figure 3.4, where these gene sets are highlighted by a larger font size corresponding to their probability for Topic 2.

to relate experiments according to the mechanisms shared between them. The assignment between topics and comparisons is not disjoint as in clustering, but instead each comparisons can genuinely belong to several topics.

The previous two examples illustrate that the topic model is in fact finding topics that correspond to meaningful biological processes. By combining gene sets into topics, a holistic model of the differential activation of biological processes is created. The approach also appears to be robust, as the topic model was inferred from a collection of experiments from a wide range of different sources, and as the above examples show, similar comparisons from different laboratories and samples match to the same biological processes. The robustness observed by the approach describes here is at least partly due to the robustness of the methods that were combined. For instance, GSEA is known to be robust with respect to laboratory and sample variations, and topic models are robust with respect to noise in the input data.

3.3.2 Visualization of the Model

A major strength of the topic model visualization in Figure 3.3 is that it connects gene sets to experiments while making the connection by compressing the relationships through the topics. This enables the efficient interpretation of topic distributions of the comparisons, and thus the comparisons themselves. Furthermore, the visualization allows the investigator to begin the exploration of the model with a comparison, a topic or a gene set.

The larger structure of the model becomes evident immediately, namely that topics hardly ever share their top gene sets, while topics are shared across experiments with similar probabilities quite frequently. It is also possible to observe that some experiments have what could be called a "primary topic" that is indicated by a wider-than-average edge connecting the experiment to a topic. For example, in Figure 3.3 an instance of a "high-stage neuroblastoma" comparison can be identified where Topic

19 seems to be the primary topic. The glyph on the right in Figure 3.6 (**b**) confirms this.

The visualization also reveals how gene sets are distributed across topics and that there is a range of different distributions. As shown in Figure 3.4, it can for example be seen that Topic 2 has very high probabilities for four gene sets and very low probabilities for the remaining gene sets, where as Figure 3.5 illustrates that Topic 24 has rather uniformly distributed probabilities for a wide range of gene sets.

Figure 3.6 (**a**) shows a NeRV projection of the comparisons including glyphs describing the probability distribution over the top 13 topics. While the visualization of the topic model in Figure 3.3 provides some insight into the structure of the experiment space the projection immediately provides an overview of clusters and outliers. Only a few distinct clusters can be identified in the subset of 105 comparisons, but this is not surprising given the range of phenotypes that have been investigated in those comparisons.

The glyphs reveal how topic usage is changing across comparisons and explain which topics are shared by comparisons forming a cluster. The change in topic usage is gradual in most parts of the projection, but seems abrupt in others. This could indicate imperfectness in the projection where not all similarities have been preserved by the dimensionality reduction.

3.3.3 Evaluation of the Retrieval Performance

The performance of the method can be evaluated quantitatively by retrieving relevant experiments given a query experiment. For this purpose, queries were performed with individual cancer comparisons where all other cancer comparisons were considered to be relevant, and all non-cancer comparisons to be irrelevant. Average Precision (AP) and recall (R) were used to measure the retrieval performance. For retrieval methods that return a set of items better precision (P) means the number of possibly relevant items in that set is maximized, which can be expressed as

$$P = \frac{\text{number of relevant items that are retrieved}}{\text{total number of retrieved items}}.$$

Since the model described here returns a ranked list, rather than just a set of items, the average precision is used. Average precision is defined as follows (Manning et al., 2009, see Chapter 8):

$$AP = \frac{1}{n} \sum_{k=1}^{n} P_k,$$

where n is a cut-off rank for the list of returned items and P_k is the "precision at k", i.e. the precision for the first k returned items. Improved recall means that the number of possibly irrelevant objects that are retrieved is minimized. Recall is defined as

$$R = \frac{\text{number of relevant items that are retrieved}}{\text{total number of relevant items}}.$$

The cancer category was chosen for this evaluation because it has the largest number of comparisons in the collection and, more importantly, the associated experiments are from several laboratories and different cancer types. For the other diseases the number of experiments is either too small or they come from a single larger experiment, making retrieval potentially too easy.

Overall, the system was queried with each of the 27 comparisons comparing normal versus cancerous tissue. As a result, a ranked list of comparisons was obtained, sorted by the probability of the query, given the comparisons and the model, as discussed in Section 3.2. The average precision was computed over the top 10 retrieved experiments, i.e. $AP = \frac{1}{10} \sum_{k=1}^{10} P_k$. As a baseline the average precision over randomly ranked results was computed as well. By randomizing 1000 times, an estimate of the confidence intervals was obtained. The average of the precision-recall curves for all queries were computed as well, for both the topic model and a random baseline.

As shown in Figure 3.7(**a**), in more than half the queries, the average

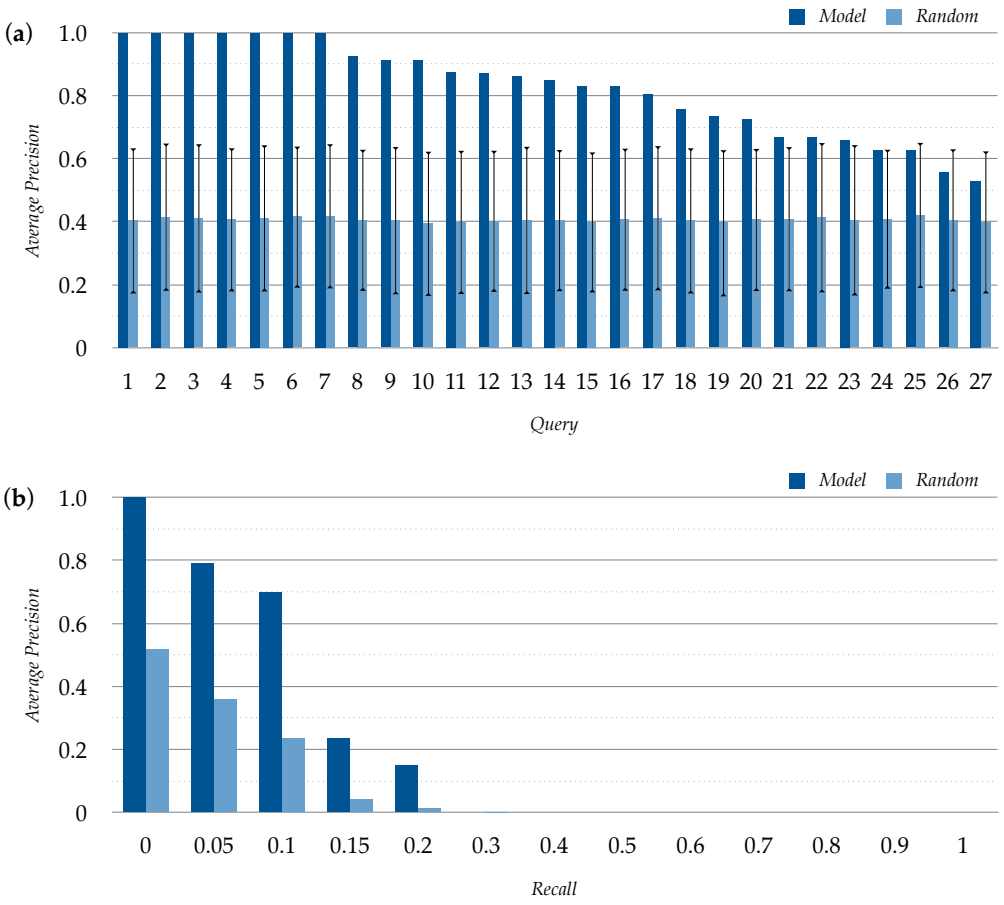

Figure 3.7: (**a**) Average Precision for cancer queries for the top 10 results. Queries are sorted by the average precision given by the topic model. Error bars represent the 99% confidence interval of the random permutation results. (**b**) Interpolated average precision at eleven standard recall levels. The dark blue bars corresponds to the model; the light blue bars corresponds to the random baseline.

precision is above 0.8, and in 20 of the 27 queries the topic model-based retrieval is above the confidence interval of the random baseline. As seen in Figure 3.7(**b**), the precision-recall values show that the trade-off between precision and recall in the method is reasonable and well above the random baseline. The false positives in the top 10 ranked experiments were studied as well, and on average 20% of those were found to be cancer-related (e.g. benign tumor), showing an advantage of the topic model over standard annotation-based searches and suggesting that the actual

retrieval performance is actually better than the quantitative evaluation suggests.

3.3.4 Searching for Experiments

The potential of the probabilistic relevance search is illustrated by two case studies in which different comparisons are used a queries.

3.3.4.1 Case Study: Malignant Melanoma

The collection is queried with a comparison of normal tissue against malignant melanoma. The top two results are comparisons of normal tissue against bladder carcinoma and vulvar intraepithelial neoplasia. The next two results are hyperparathyroidism and a study of intra-pulmonary airway epithelial cells from non-smokers versus current smokers. The remaining top ten results are from comparisons of normal tissue against bladder carcinoma (twice), infiltrating ductal carcinoma, prostate cancer, breast carcinoma, and esophageal adenocarcinoma. It is clear that cancer experiments have a high prevalence in the top results, given the melanoma query. Interestingly, a study of intra-pulmonary airway epithelial cells from smokers is among the top results. Although the annotation is not completely clear as to what the actual pathology is in this study, it is plausible that it might be a cancer-related one. This highlights the capability of the described approach for hypothesis generation in an explorative context. Finally, it is known that hyperparathyroidism is associated with a higher cancer incidence (Nilsson et al., 2007); a relationship that is highlighted by the melanoma query.

Figure 3.6 (**c**) visualizes the topic distributions for experiments found relevant to the melanoma query. The visualization not only highlights the most relevant experiments, but also the relationship between them. In particular, a subset of the carcinoma experiments appears to become separate from the glandular-related pathologies (primary hyperparathyroidism, and prostate cancer). Alternatively, Figure 3.8, which is

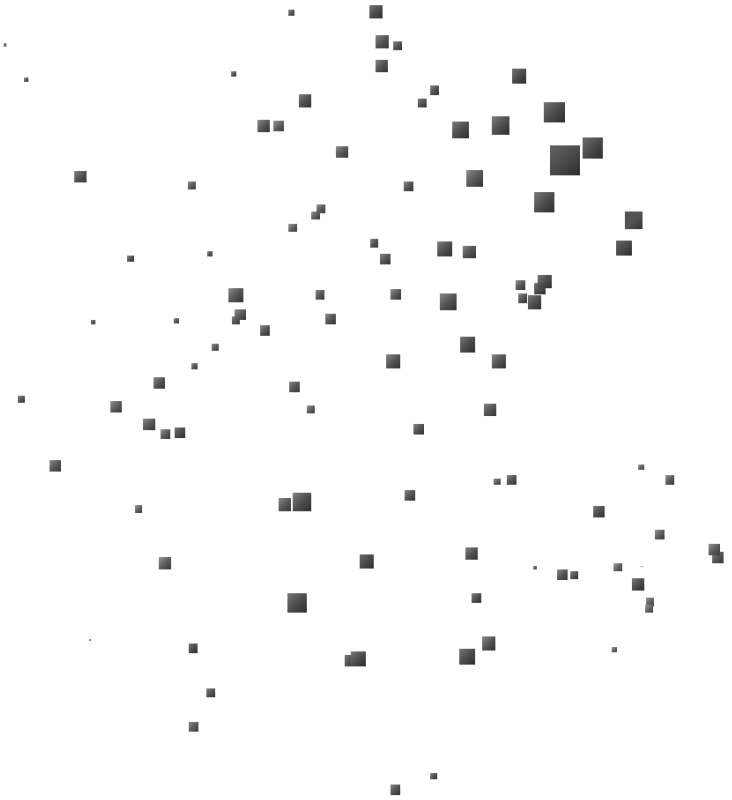

Figure 3.8: NeRV projection of the subset of 105 comparisons, illustrating the outcome of querying the model with a melanoma experiment. The glyph size encodes the relevance of each experiment to the query. The bigger the glyph, the higher the relevance of the experiment to the query. The query itself is represented by the biggest glyph.

also a NeRV projection with glyphs, distinguishes the relevance of each experiment by changing the glyph size accordingly.

3.3.4.2 Case Study: Myelogenous Leukemia

In this case study, a query is performed with a comparison of myelogenous leukemia against a control. Surprisingly, the top result is Crohn's disease. Although Crohn's disease is a digestive system disease, it is linked to the query through involvement of folate biosynthesis, as described in the previous section. The second result is chronic lymphocytic leukemia and the remaining results are ischemic cardiomyopathy, post-traumatic stress dis-

order, multiple invasive and transitional cell carcinomas, and chronic obstructive pulmonary disease, all against normal controls. Although the top ten results span a large class of diseases, some of which are hard to connect to the query pathology, this case study highlights the fact that the method is capable of extracting meaningful results among the top retrieved comparisons. This is evident both by the fact that a disease that is very similar to the query is ranked higher than other diseases that are broadly similar (chronic lymphocytic leukemia), and by the fact that Crohn's disease is ranked highly, which, although not immediately identifiable as similar to the query, has been reported to share properties with it.

3.4 Discussion

The methods introduced in Section 3.2 allow the retrieval of comparisons relevant to a given query comparisons. Retrieval and interpretation are based on methods for modeling and visualizing differential gene set expression in a collection of transcriptomics experiments that have been decomposed into binary comparisons. The probabilistic model combines two approaches that independently have been shown to be effective. The model is able to cluster gene sets into components, called topics, that exhibit a significant biological coherence and that are meaningfully related to particular experimental variables. As shown in Section 3.2.4, the probabilistic nature of the model allows for a precise formulation of retrieval, in which the model is queried with the differential expression in gene sets of a comparisons and it returns a ranked list of relevant comparisons. By querying the model with cancer experiments, it was shown that performance significantly better than random can be obtained, measured by average precision. More importantly, the mean average precisions on average is at the good value of about 0.82, whereas the random baseline is at about 0.40.

The quantitative analysis was complemented by two case studies. The model was able to associate melanoma with several cancer types. Also

demonstrated was how the model finds hypothetical connections between experiments by selecting an experiment of epithelial tissue in non-smokers versus current smokers as being highly relevant to cancer experiments, which naturally makes sense *a posteriori*. The model finds relations between Crohn's disease and leukemia, which was shown in a second case study, and also between hyperparathyroidism and cancer. Both findings can be confirmed in the literature. Finally, given a query comparison of leukemia versus a normal control, the model was able to extract, from a set of cancer comparisons, precisely another leukemia experiment as being the most relevant. The result indicates that the model not only manages to partition the comparisons into general classes but also provides fine-grained distinctions. These results are supplemented with a concise visual description of one of those case studies, highlighting the consistence in topic distributions between similar comparisons.

The visualization methods developed to complement the probabilistic approach were used to gain insight into the structure of the model. For instance, with the circular visualizations shown in Figures 3.3, 3.4 and 3.5 the topic and gene set distributions of comparisons and topics, respectively, can be studied. This is an important aspect in understanding how the biological processes in the compendium are represented by the model. The visualizations shown in Figures 3.6 and 3.8, which are based on projections of comparisons according to their topic distributions, enable the identification of sets of related comparisons and outliers. They also provide an overview of the overall distribution of comparisons in the space defined by the topics.

The described probabilistic approach can be further developed into two complementary directions. In the current approach, the complexity of the model was intentionally designed to be low by using a simple way of bringing in prior biological knowledge, and a reasonably simple probabilistic model. Since already such a simple approach proved to be very useful in retrieving relevant experiments, a next step could be to scale up to larger and more diverse experiment collections, for example by includ-

ing data from other species. An alternative direction is to include more detailed models, making the retrieval results and analyses more accurate and informative, at the cost of increased computational complexity. Straightforward extensions of the described approach are available in form of recent methods that attempt to improve on GSEA (e.g. Oron et al., 2008). Additionally, there have been a wide variety of extensions of topic models in recent years, for instance, allowing topics to be correlated (Blei and Lafferty, 2007) or to form a hierarchical structure (Blei et al., 2003a).

3.5 Refined Model and Multi-Species Data

To address some of the issues identified in the discussion of the original method described in the first part of this chapter, the method was extended with a refined model and applied to an expanded collection of data sets.

Like the original method, the extended method still consists of two main steps: the first step is a gene set enrichment analysis (GSEA) and based on the results of this analysis, a latent component model is derived in a second step. The model designed for the extended method is a probabilistic generative hierarchical model[3] that introduces a second layer of latent components. Since a lot of diseases and drug treatments are studied in animal models, the extended method described here was designed to handle data from vertebrate model organisms. The data sets to which the method was applied included studies performed with microarrays designed for human, mouse and rat transcriptomes.

A detailed overview of the data processing pipeline for the extended method is shown in Figure 3.9. Since the approach is similar to the original method described above, only the key differences between the methods are discussed here. The different parts of the method are described in the order in which they appear in the processing pipeline. As a

[3]The original model is also a *generative hierarchical model*, and only to distinguish between the two models in this chapter, the new model is referred to as the "hierarchical model" because it has an extra layer, while the previous model is referred to as the "original model".

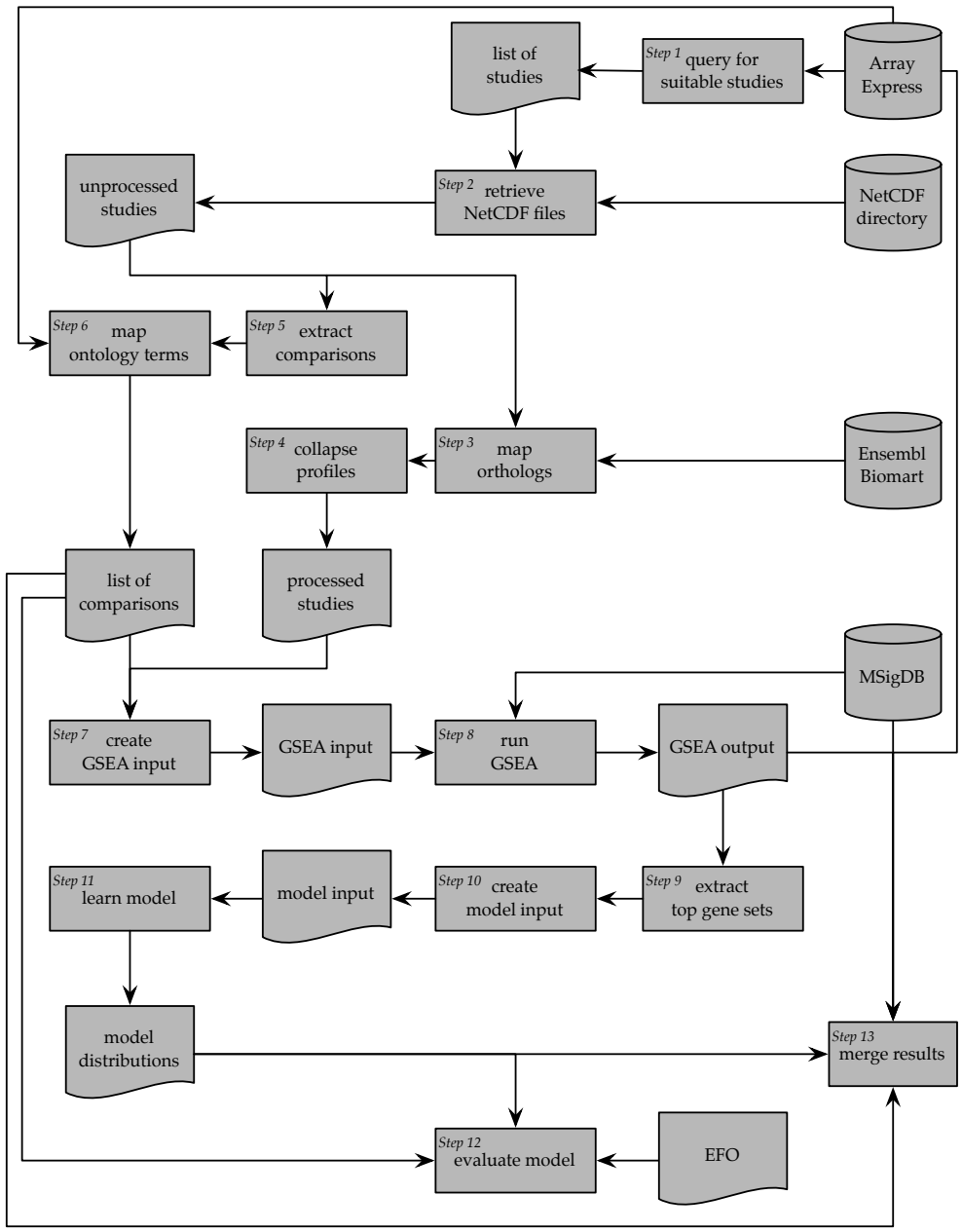

Figure 3.9: Data processing pipeline for the extended method. Steps 1 through 6 and 11 through 13 are discussed in the main text. All other steps are essentially the same as in the original method.

consequence, less important steps, such as data collection and preprocessing, are described before the key contributions.

3.5.1 Collection of Data Sets

As in the original study, microarray data sets were obtained from the ArrayExpress Archive by selecting all data sets for human (*Homo sapiens*), mouse (*Mus musculus*) and rat (*Rattus norvegicus*) that include a preprocessed expression matrix and sufficiently curated annotation (Figure 3.9, Step 1). The data sets fulfilling these criteria are employed to construct the ArrayExpress Atlas database (Kapushesky et al., 2010) and the same underlying data files in the Network Common Data Format (NetCDF) were used to import the data sets into R data structures for further processing (Step 2).

A total of 1,082 microarray data sets were retrieved on 26 October 2009. Out of those 474 were from human, 441 were from mouse, and 167 were from rat studies.

3.5.2 Mapping to Human Genes

The C2-CP collection of canonical pathway gene sets from the Molecular Signature Database (MSigDB) was used for GSEA as in the original study. Since these gene sets are only available as lists of human gene symbols, mouse and rat genes have to be mapped to human genes before GSEA can be applied (Figure 3.9, Step 3). An internal table of the ArrayExpress data base was used to access a mapping from array features to Ensembl Gene identifiers, which is provided by Ensembl (www.ensembl.org; Hubbard et al., 2009). Based on the the information in that table, human orthologs of mouse and rat genes were identified by querying the ortholog mapping provided by the Ensembl BioMart (Ensembl Release 56; Vilella et al., 2009; Kasprzyk et al., 2004). In a final step, all human Ensembl Gene identifiers were mapped to human gene symbols for use with the MSigDB gene sets. In data sets where multiple array features map to the same gene, and therefore multiple expression profiles exist for a gene, they are collapsed into a single profile by computing the median expression profile across the corresponding features (Figure 3.9, Step 4).

The validity of using such an approach to deal with cross-species transcriptomics data in gene set enrichment analyses is supported by several previous publications (e.g. Sweet-Cordero et al., 2005; Bourquin et al., 2006), which successfully applied similar ortholog mapping approaches.

3.5.3 Decomposition into Comparisons

The data sets were again decomposed into binary comparisons between two conditions, A and B, to be able to process them with GSEA (Step 5). For the extended method the following criteria were applied:

1. All samples for conditions A and B are annotated with exactly one of two different factor values that belong to the same experimental factor.

2. If there are additional experimental factors used in the study, the factor values of each of those must be the same for all samples associated with conditions A and B. These factor values form the *context* of the comparison.

3. For each condition there must be at least three samples.

4. *Neutral factors* are ignored. Neutral factors are factors that would not result in meaningful comparisons and have a very large number of associated factor values within a study. The factors "age" or "individual" are examples for such cases. The full list of neutral factors is shown in Appendix B.1.

All comparisons that are possible according to these rules were extracted, resulting in a total of 6,925 comparisons. Of those, 1,976 came from human studies, 2,137 came from mouse studies and 2,812 came from rat studies.

The extracted comparisons were further classified into whether they are *interpretable* or not. An interpretable comparison is defined as having exactly one condition that can be considered as a "control" or "normal" state in the context of the experiment. Such conditions are, for example,

wild type strains where different genotypes are being compared, a mock treatment where the effects of drugs are analyzed, or healthy tissues where cancers are studied. The assumption is that the effects observed in an interpretable comparison can be attributed to the non-control condition.

In order to identify interpretable comparisons, a list of *control factor values* was assembled by manually classifying all factor values used in the collection of data sets. The full list of control factor values is shown in Appendix B.2. A total of 908 comparisons were classified as interpretable, with 325 coming from human, 429 coming from mouse and 154 coming from rat studies. The number of interpretable comparisons is almost nine times higher than in the original study, where only 105 interpretable comparisons were used. Furthermore, the original study considered only comparisons of disease against some control as interpretable, whereas here interpretable comparisons derived from a wide range of different experimental factors were considered.

3.5.4 Hierarchical Probabilistic Model

One important observation made in the analysis of the topics inferred by the original approach, is that the Latent Dirichlet Allocation (LDA) model often inferred topics with notable overlap in the highest probability gene sets, as described in Section 3.3.1.

To address this problem, a generative hierarchical probabilistic model was developed for the extended method. The key difference between the models, that is most relevant for the discussion here is that the hierarchical model has an additional layer of latent components. Furthermore, it models binary data, rather than count data, and also describes activation probabilities for individual genes in addition to gene set activation probabilities (J. Caldas, personal communication).

The additional layer of latent components included in the hierarchical model represents so-called *modules*, which are essentially *super topics* that are intended to represent higher level processes observed in the

transcriptomics data. Every comparison is modeled as a multinomial distribution over modules and every module is a multinomial distribution over topics. Unlike in the original formulation, where topics are modeled as multinomial distributions over gene sets, the topics here are modeled as Bernoulli distributions to represent the binary activation of gene sets. Furthermore, each topic also models the binary activation of genes in each gene set as a Bernoulli distribution.

As before, the model is learned using *collapsed Gibbs sampling* (Griffiths and Steyvers, 2004) based on the results obtained by applying GSEA to all 6,925 comparisons (Figure 3.9, Step 11). For each comparison, the top 50 gene sets, based on their normalized enrichment score (NES), were considered to be activated. Likewise, genes were considered to be activated when they are included in the leading edge subset of an activated gene set. Several combinations for the number of modules and topics were tested. Based on the posterior probability of the model, which is the probability of the model given the observed data, the number of modules was fixed at 45 and the number of topics was fixed at 60.

For retrieval, the hierarchical model is applied in the same way as the original model. The relevance of a retrieved comparison to query comparison is computed as the probability of the data observed in the query comparison, given the model parameters for the retrieved comparison. For a given query this probability is computed for all other comparisons in the collection and the returned list of results is ranked according to the probabilities. Retrieved comparisons with higher probabilities are considered to be more relevant to the query and are ranked higher.

The topics identified by the hierarchical model can be interpreted through analysis of the gene sets with the highest probabilities, as in the original model. However, due to the fact that modules are combinations of topics, it is less straightforward to find clear interpretations of modules.

3.5.5 Evaluation of Retrieval Results

In the original model, the evaluation of retrieval results relied on a manual classification of comparisons into "cancer-related" and "not cancer-related". This was possible because the number of comparisons was rather small. For the extended method described here, a more sophisticated and scalable approach was devised for the evaluation of retrieval results (Figure 3.9, Step 12).

3.5.5.1 Ontology-based Relevance Score

The directed, acyclic graph structure of the Experimental Factor Ontology (EFO, Malone et al., 2010) is a representation of the relationships between experimental factor values used in the studies in ArrayExpress. For evaluation purposes and to compare the extended method described here to other information retrieval methods, the EFO was used as an external "gold standard", based on which the relevance of a retrieved comparison given a query is measured. Similar evaluation methodologies have been described, for instance by Hibbs et al. (2007), who employed the Gene Ontology (Ashburner et al., 2000) to cross-validate a gene-centric search engine for expression data.

For this purpose, the experimental factor values defining the interpretable comparisons were mapped to terms in the EFO (Release 1.7) if possible (Figure 3.9, Step 6). The mapping used here is a curated mapping that is also used for the ArrayExpress Atlas and was obtained from an internal table of the ArrayExpress database. When the non-control condition of an interpretable comparison can be mapped to the EFO, the comparison is called an *evaluable* comparison. A total of 219 evaluable comparisons were identified based on the mapping from the ArrayExpress Atlas, with 137 coming from human studies, 39 coming from mouse studies and 43 coming from rat studies.

To compute the similarity between terms in the EFO, a modified version of the *Jaccard coefficient* (Jaccard, 1908) was employed. The classic

Jaccard coefficient J is a distance measure used to determine the similarity between two sets Q and R as:

$$J(Q,R) = \frac{|Q \cap R|}{|Q \cup R|}.$$

Applied to the EFO, the sets Q and R are defined as the ontology terms on the shortest paths between the root and term q and the root and term r. The modified Jaccard coefficient J' used here is defined as:

$$J'(Q,R) = \begin{cases} 1, & \text{if } r \text{ is a child of } q; \\ J(Q,R) & \text{otherwise.} \end{cases}$$

In the context of the retrieval method described above, the modified Jaccard coefficient is used to determine the graded relevance $rel(q,r) = J'(Q,R)$ of a retrieved comparison mapped to term r, when the query is mapped to term q. The relevance will be at its maximum of 1 when both the retrieved comparison and the query map to the same ontology term or when the retrieved comparison maps to a child of the query term. Accordingly, this relevance measure is not symmetric and can yield different results when q and r are exchanged.

3.5.5.2 Query Result Scoring

The precision and recall measures used to evaluate the retrieval performance of the original model require that a retrieved comparison is judged to be either relevant or irrelevant. However, since the modified Jaccard coefficient provides a graded relevance between 0 and 1, precision and recall cannot be applied to evaluate the performance of the extended method.

Järvelin and Kekäläinen (2002) describe a family of "cumulative gain" evaluation methods that are based on graded relevance judgements. Applied to the retrieval of comparisons, these methods measure how much the investigator gains when a comparison with a particular relevance is found at a particular rank in the result list for a query.

The *Discounted Cumulative Gain* for a ranked list of graded relevance judgements $rel(q, r_i)$, with $i = 1, \ldots, C$ where C is the number of interpretable comparisons, is defined as

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel(q,r_i)} - 1}{log_2(i+1)},$$

with p being the position in where ranked list is cut. To evaluate the extended method p was set to C, which means that the complete ranked list was taken into account. The interpretation of the DCG is that retrieved comparisons of equal relevance become less valuable, or provide less gain, the farther away from the top of the list they occur.

In order to compare the DCG across retrieval methods, it has to be normalized. The *Normalized Discounted Cumulative Gain* (NDCG) is the DCG relative to the best possible or *Ideal Discounted Cumulative Gain* (IDCG) and is defined as

$$NDCG_p = \frac{DCG_p}{IDCG_p}.$$

The IDCG is obtained by computing the DCG for the list of comparisons ranked by their relevance according to the gold standard, here expressed by $rel(q, r)$.

Applied to all 219 evaluable comparisons, the *median NDCG* for the extended method with 45 modules and 60 topics is 0.8790. In order to interpret this value, the NDCG was also computed for the original method and two common information retrieval methods (J. Caldas, personal communication). The median NDCG for the 219 comparisons using the original method with 130 topics applied to the data set is 0.8781. The median NDCG for *Term Frequency - Inverse Document Frequency* (TF-IDF; Spärck Jones, 1972) with cosine similarities based on the count representation used in the original method is 0.8767. Finally, the median NDCG obtained by using *Spearman rank correlation* (Spearman, 1904) as the similarity measure based on the fold change ratios of the expression data for

the 219 evaluable comparisons is 0.8856.

Even though the Spearman rank correlation performs slightly better than the extended method, the four evaluated retrieval methods have very similar performance according to the gold standard derived from the Experimental Factor Ontology. The advantage of the methods described in this chapter is that they readily provide the investigator with relevant information for the interpretation of the results, unlike TF-IDF and Spearman rank correlation. This is illustrated by the case studies in Section 3.6.

3.5.6 Query Interface

In order to query the collection of 908 interpretable comparisons, and to interpret query results, a basic web interface was developed (Figure 3.9, Step 13). Through the interface the collection of comparisons can be queried either with a comparison or with a gene set to retrieve the corresponding top 25 related comparisons. A screenshot of the query results page is shown in Figure 3.10.

For the initial analysis of the results presented here, only the top 25 retrieved comparisons for every query were considered. For the sake of simplicity, all queries were precomputed and the interface is a collection of static Hypertext Markup Language (HTML) pages containing internal and external links that allow the investigator to explore the collection of comparisons and retrieve additional information about studies as well as gene sets.

In order to help the investigator with the interpretation of the results, the top 25 gene sets are shown for both the query and the top 25 retrieved comparisons. The top 25 gene sets for a comparison are obtained by multiplying the module probabilities of the comparison with the topic probabilities and the gene set probabilities and by ranking the gene sets according to the result. Additionally, if the gene set was found to be differentially upregulated or downregulated by GSEA, this information is also available in the web interface.

Figure 3.10: Web-based query interface for the extended method. Details about the query are shown at the top in the light blue box and query results are listed below. (1) Name of the non-control condition of the comparison, also used as name for the comparison. (2) Name of the control condition and species. (3) Context of the comparison. (4) Accession number and description of the study from which the comparison was extracted. Clicking the name of the study opens the corresponding entry in the ArrayExpress Repository. (5) List of top 25 gene sets for the comparison. Can be expanded and collapsed. Gene sets written in red are upregulated and gene sets written in blue are downregulated according to GSEA. Gene sets written in black were not found to be differentially expressed. (6) Internal and external gene set information. Clicking on "R" opens the gene set page with the top 25 studies associated with the gene set. Clicking on "M" opens the corresponding page on the MSigDB website. (7) Internal information. (8) Top 25 comparisons for the query results. (9) Internal link that queries the collection with the selected comparison.

Table 3.2: Query results for a benign nevi comparison. Text in parentheses after the name of the species is the context of the corresponding comparisons.

Q	**benign nevi** vs normal in *Homo sapiens*	E-GEOD-3189
1	**viral cardiomyopathy** vs normal in *Homo sapiens*	E-GEOD-1145
2	**polycystic ovary syndrome** vs normal in *Homo sapiens* (none)	E-GEOD-1615
3	**myelodysplastic syndrome** vs normal in *Homo sapiens* ("" & female)	E-GEOD-2779
4	**idiopathic cardiomyopathy** vs normal in *Homo sapiens*	E-GEOD-1145
5	**familial cardiomyopathy** vs normal in *Homo sapiens*	E-GEOD-1145
6	**EWS-FL1 transfected** vs mock transfected in *Homo sapiens*	E-GEOD-1822
7	**ischemic cardiomyopathy** vs normal in *Homo sapiens*	E-GEOD-1145
8	**Parkinson's disease** vs normal in *Homo sapiens* (female)	E-GEOD-7621
9	**RAD001** vs placebo in *Mus musculus* (wild_type & 48 h)	E-GEOD-1413
10	**SR-A mutant** vs wild_type in *Mus musculus* (bilateral olfactory bulbectomy & 8 h)	E-GEOD-3455
11	**diabetes mellitus** vs normal in *Rattus norvegicus* (4 weeks)	E-MEXP-515
12	**post-partum cardiomyopathy** vs normal in *Homo sapiens*	E-GEOD-1145
13	**Rs1h null** vs wild_type in *Mus musculus*	E-GEOD-5581
14	**hypertrophic cardiomyopathy** vs normal in *Homo sapiens*	E-GEOD-1145
15	**carcinoma in situ lesion** vs normal in *Homo sapiens* (bladder)	E-GEOD-3167
16	**1 d** vs 0 in *Homo sapiens* (female)	E-GEOD-1295
17	**ochratoxin** vs none in *Rattus norvegicus* (kidney & 7 m)	E-GEOD-2852
18	**malignant melanoma** vs normal in *Homo sapiens*	E-GEOD-3189
19	**non steroidal anti-inflammatory drugs** vs none in *Homo sapiens* (osteoarthritis)	E-GEOD-7669
20	**coxsackievirus B3** vs uninfected in *Homo sapiens* (none & 3 h)	E-GEOD-697
21	**valproic acid** vs none in *Homo sapiens* (normal)	E-GEOD-1615
22	**12 h** vs 0 h in *Homo sapiens* (control)	E-GEOD-3183
23	**Beta4 nAChR subunit null** vs wild_type in *Mus musculus* (none)	E-GEOD-6614
24	**dexamethasone** vs none in *Homo sapiens* (24 h)	E-GEOD-1815
25	**Dysf-/-** vs wild_type in *Mus musculus* (left ventricular myocardium)	E-GEOD-2507

3.6 Case Studies

The web interface described in Section 3.5.6 was used to explore the collection of 908 interpretable data sets and to interpret retrieval results computed with the hierarchical model. To demonstrate the utility of the method, several examples for interesting query results are presented and discussed in the following sections. All gene symbols are human gene symbols unless otherwise noted.

3.6.1 Benign Nevi, Malignant Melanoma and Cardiomyopathies

When querying the database with the comparisons *benign nevi vs. normal*, shown in Table 3.2, and *malignant melanoma vs. normal*, shown in Table 3.3, the top 25 results in both cases contain a range of different cardiomy-

Table 3.3: Query results for a malignant melanoma comparison. Text in parentheses after the name of the species is the context of the corresponding comparisons.

Q	**malignant melanoma** vs normal in *Homo sapiens*	E-GEOD-3189
1	**ochratoxin** vs none in *Rattus norvegicus* (kidney & 7 m)	E-GEOD-2852
2	**hypertrophic cardiomyopathy** vs normal in *Homo sapiens*	E-GEOD-1145
3	**benign nevi** vs normal in *Homo sapiens*	E-GEOD-3189
4	**ischemic cardiomyopathy** vs normal in *Homo sapiens*	E-GEOD-1145
5	**idiopathic cardiomyopathy** vs normal in *Homo sapiens*	E-GEOD-1145
6	**post-partum cardiomyopathy** vs normal in *Homo sapiens*	E-GEOD-1145
7	**24 h** vs 0 in *Homo sapiens* (none)	E-GEOD-2803
8	**EWS-FL1 transfected** vs mock transfected in *Homo sapiens*	E-GEOD-1822
9	**familial cardiomyopathy** vs normal in *Homo sapiens*	E-GEOD-1145
10	**viral cardiomyopathy** vs normal in *Homo sapiens*	E-GEOD-1145
11	**Parkinson's disease** vs normal in *Homo sapiens* (female)	E-GEOD-7621
12	**diabetes mellitus** vs normal in *Rattus norvegicus* (4 weeks)	E-MEXP-515
13	**breast cancer** vs normal in *Homo sapiens*	E-MEXP-882
14	**hypoxic-ischemic injury** vs none in *Mus musculus* (none)	E-GEOD-1999
15	**soluble tumor necrosis factor alpha** vs control in *Mus musculus*	E-GEOD-4518
16	**polycystic ovary syndrome** vs normal in *Homo sapiens* (none)	E-GEOD-1615
17	**myelodysplastic syndrome** vs normal in *Homo sapiens* ("" & female)	E-GEOD-2779
18	**12 h** vs 0 h in *Homo sapiens* (control)	E-GEOD-3183
19	**RAD001** vs placebo in *Mus musculus* (wild_type & 48 h)	E-GEOD-1413
20	**valproic acid** vs none in *Homo sapiens* (normal)	E-GEOD-1615
21	**SR-A mutant** vs wild_type in *Mus musculus* (bilateral olfactory bulbectomy & 8 h)	E-GEOD-3455
22	**carcinoma in situ lesion** vs normal in *Homo sapiens* (bladder)	E-GEOD-3167
23	**Hoxc13 overexpressing transgenic** vs wild_type in *Mus musculus*	E-GEOD-2374
24	**dermatomyositis** vs normal in *Homo sapiens*	E-GEOD-5370
25	**dexamethasone** vs none in *Homo sapiens* (24 h)	E-GEOD-1815

opathies (viral, idiopathic, familial, ischemic, post-partum, hypertrophic). Furthermore, the two comparisons retrieve each other, indicating a link between the two conditions, and also several cancer-related conditions, such as transfection with Ewing sarcoma family fusion gene, breast cancer, and carcinoma in situ lesion.

The link between benign nevi and malignant melanoma is well established and has been studied extensively (Talantov et al., 2005). However, the link between these conditions and (cardio)myopathies has been reported only once before in a recent study by Hu and Agarwal (2009), where the authors used gene expression data in an approach conceptually similar to the Connectivity Map (Lamb et al., 2006) to identify links between human diseases. In their paper, Hu and Agarwal suggest that the link between benign nevi/malignant melanoma and cardiomyopathies is an inverse relationship that was found due to the cell growth properties

of the benign nevi/malignant melanoma and the muscular weakness or wasting properties of the cardiomyopathies.

The extended method provides further evidence that the relationship is indeed inverse, as many of the top 25 gene sets that are affected by the conditions are upregulated in benign nevi/malignant melanoma and downregulated in cardiomyopathies. For instance, the most relevant gene set for benign nevi is the *phosphoinositide 3-kinase (PI 3-K) pathway* (www.broadinstitute.org/gsea/msigdb/cards/PTDINSPATHWAY.html), which, among other things, is involved in cell survival and cell proliferation (Engelman, 2009). The second most relevant gene set is the Ras pathway (www.broadinstitute.org/gsea/msigdb/cards/RASPATHWAY.html), which is upregulated in this comparison. The Ras pathway activates the PI 3-K pathway, thus inhibiting apoptosis. In the malignant melanoma comparison, the Ras pathway is also among the top 25 gene sets and upregulated. In contrast to the benign nevi and malignant melanoma comparisons, the Ras pathway is among the top 25 gene sets and downregulated in almost all cardiomyopathies.

This case study is an example of how the information provided by the extended method can be used to both identify interesting links between seemingly unrelated conditions and also perform a high-level analysis of these links.

3.6.2 Pancreatic Ductal Adenocarcinoma, Insulin and Inflammation

The model found a relationship between *pancreatic cancer vs. normal* and insulin-related conditions as well as obesity. The top 25 comparisons are shown in Table 3.4. The pancreatic cancer in this comparison is a pancreatic ductal adenocarcinoma (PDAC).

The most relevant result when querying with PDAC is a preadipocyte cell line from mouse, in which IRS4 (insulin receptor substrate 4) has been knocked out. This is expected to have an effect on insulin

Table 3.4: Query results for a pancreatic cancer comparison. Text in parentheses after the name of the species is the context of the corresponding comparisons.

Q	**pancreatic cancer** vs normal in *Homo sapiens*	E-MEXP-950
1	**IRS-4** vs wild_type in *Mus musculus*	E-GEOD-2556
2	**growth hormone** vs control in *Mus musculus* (48 h)	E-GEOD-2120
3	**insulin** vs none in *Homo sapiens*	E-GEOD-7146
4	**D374Y-PCSK9** vs wild_type in *Homo sapiens*	E-MEXP-1235
5	**obesity** vs normal in *Homo sapiens* (male)	E-GEOD-2508
6	**severe malarial anaemia** vs normal in *Homo sapiens*	E-GEOD-1124
7	**partial paw denervation** vs sham denervation in *Rattus norvegicus* (3)	E-GEOD-2874
8	**Gyk knockout** vs wild_type in *Mus musculus*	E-GEOD-4656
9	**quercetin** vs none in *Mus musculus* (POR null & jejunum)	E-GEOD-4262
10	**U0126** vs none in *Homo sapiens* (coxsackievirus B3 & 9 h)	E-GEOD-697
11	**E2F2-/-** vs wild_type in *Mus musculus* (48 h)	E-MEXP-1413
12	**Trembler** vs wild_type in *Mus musculus* (B cell & P4)	E-GEOD-1947
13	**IRS-1** vs wild_type in *Mus musculus*	E-GEOD-2556
14	**24 h** vs 0 h in *Homo sapiens*	E-MEXP-1194
15	**Cnr1 -/- /Cnr2 -/-** vs wild_type in *Mus musculus* (dinitrofluorobenzene)	E-GEOD-7694
16	**non-progressive HIV infection** vs uninfected in *Homo sapiens* (CD8+ T cell)	E-GEOD-6740
17	**bic-deficient** vs wild_type in *Mus musculus* (Th1)	E-TABM-232
18	**Brg1 null** vs wild_type in *Mus musculus*	E-GEOD-5371
19	**SOD1 mutant** vs control in *Mus musculus* (6 weeks & spinal cord)	E-GEOD-3343
20	**IL-22** vs control - untreated in *Homo sapiens*	E-GEOD-7216
21	**alpha-tocopherol + gamma-tocopherol** vs none in *Mus musculus* (5 months)	E-GEOD-8150
22	**ulcerative colitis** vs normal in *Homo sapiens* (female)	E-GEOD-3365
23	**IMP(1,3)A** vs mock in *Homo sapiens*	E-MEXP-548
24	**lipopolysaccharide** vs none in *Homo sapiens* (low response)	E-GEOD-3491
25	**chimpanzee diet** vs control diet in *Mus musculus*	E-GEOD-6297

signaling. Also highly ranked is a knock-out of IRS1 (insulin receptor substrate 1) from the same original study.

The second most relevant result is a comparison of mouse adipocytes treated with growth hormone, which has been found to stimulate the expression of Activating Transcription Factor 3 (ATF3) (Huo et al., 2006). ATF3 is known to have a role in glucose homeostasis (Allen-Jennings et al., 2001). Treatment with growth hormone for 48 hours (as in the retrieved comparison) has been found by Huo et al. (2006) to regulate an immune response that potentially affects insulin signaling.

The third most relevant comparison found by the method is a comparison between normal and insulin-injected human muscle tissue, creating hyperinsulinemic conditions. It is well known that a link exists between pancreatic cancer and diabetes mellitus (Wang et al., 2003), which leads to the suspicion that the hyperinsulinemia studied here is related to

pancreatic cancer through its link to diabetes mellitus.

Further highly relevant results reveal, for instance, a link to a human HepG2 cell line, which overexpresses D374Y-PCSK9, a mutated allele of proprotein convertase subtilisin/kexin type 9 (PCSK9). D374Y-PCSK9 is a known key regulator of serum cholesterol, but also suspected to down-regulate certain stress-response genes and some inflammation pathways (Ranheim et al., 2008). A further link is provided by a recent study in PCSK9 knock-out mice, which showed that the mice were hypoinsulinemic, hyperglycemic and glucose-intolerant and the study suggests that pancreatic islet cells require PCSK9 for normal functioning (Mbikay et al. (2010), see Langhi et al. (2009) also for potentially contradictory results).

A comparison between wild-type and glycol kinase (GK) knock-out mice is also among the top most relevant results. The authors of the study from which this comparison originates found that, among other things, the lack of GK affects the expression of several genes that are involved in insulin signaling and insulin resistance (Rahib et al., 2007). A comparison from a study investigating the infection of HeLa cells with Coxsackie B3 virus was found as another highly relevant result related to insulin and the PDAC query. Coxsackie B viruses have been suspected to be an environmental trigger for insulin dependent diabetes mellitus type 1 (T1D) since the early 1980s (Peng and Hagopian, 2006).

Another result that has been found to have high relevance is a comparison between adipocytes from obese and non-obese human subjects. The authors of the corresponding study found that a large number of genes associated with inflammation and immune response are upregulated in obese subjects (Lee et al., 2005). This link could be due to similar processes as the ones found in the growth hormone study described above. Furthermore, this link could explain why another highly ranked result is a comparison in which the effects of the anti-inflammatory agent "Querectin" (Stewart et al., 2008) was studied.

A highly relevant, but unexpected, result is a comparison of wild-type B cells and B cells from mice carrying a point mutation ("trembler")

in the peripheral myelin protein 22 (PMP22) gene. PMP22 is involved in demyelination and dysmyelinating peripheral neuropathies (Giambonini-Brugnoli et al., 2005; Gabriel et al., 2000), which are associated with diseases such as Charcot-Marie-Tooth disease Type 1A(CMT1A) or diabetes mellitus (Chahin et al., 2007). CMT1A is usually caused by a partial duplication of the PMP22 gene (Meyer zu Hörste et al., 2006); but recently it has also been found that the PMP22 region is amplified in PDACs (Funel et al., 2009), which establishes a potential link to the query comparison. In previous studies it has been shown that the gene is actually expressed in these cancers (Li et al., 2005).

In this case study 11 of the top 14 retrieved comparisons were linked either directly to the query comparison (pancreatic cancer) or to related conditions (insulin signaling, diabetes mellitus, inflammation). The 11 comparisons came from 10 different studies in our collection (E-MEXP-950, E-GEOD-2556, E-GEOD-2120, E-GEOD-7146, E-MEXP-1235, E-GEOD-2508, E-GEOD-4656, E-GEOD-4262, E-GEOD-697, E-GEOD-1947), which indicates that the method indeed can identify links across studies. Interestingly, two of out three comparisons for which links could not be found are from studies that appear to have never been published. This may indicate problems with the data or the experimental setup, which gives reason to believe that these comparisons might be false positive hits.

3.6.3 Glioblastoma

When querying the collection of comparisons with *glioblastoma vs. normal*, the twelve most relevant results all involve samples from nervous tissue, either from the brain or the central nervous system. Among the top 25 most relevant results, which are shown in Table 3.5, a total of 19 comparisons are found that involve nervous tissue. The comparisons do not show a clear pattern, as they include cancers, induced brain and spinal cord injuries, genetic modifications, treatments with various chemicals and brain disorders such as bipolar disorder and Alzheimer's disease.

Table 3.5: Query results for a glioblastoma comparison. Text in parentheses after the name of the species is the context of the corresponding comparisons.

Q	**glioblastoma** vs normal in *Homo sapiens*	E-MEXP-567
1	**experimental autoimmune encephalomyelitis (recovery)** vs normal in *Rattus norvegicus*	E-MEXP-1025
2	**experimental autoimmune encephalomyelitis (relapsing)** vs normal in *Rattus norvegicus*	E-MEXP-1025
3	**astrocytic tumor** vs normal in *Homo sapiens*	E-MEXP-567
4	**kainate** vs control in *Rattus norvegicus* (24 h)	E-GEOD-1156
5	**neurofibriallary tangle** vs normal in *Homo sapiens*	E-GEOD-4757
6	**0.5 h** vs 0 in *Rattus norvegicus* (sham & Rattus norvegicus)	E-GEOD-2392
7	**experimental autoimmune encephalomyelitis (acute)** vs normal in *Rattus norvegicus*	E-MEXP-1025
8	**8 h** vs 0 in *Rattus norvegicus* (sham & Rattus norvegicus)	E-GEOD-2392
9	**spinal cord contusion** vs none in *Rattus norvegicus*	E-GEOD-2599
10	**lateral fluid percussion-induced injury** vs sham in *Rattus norvegicus* (Rattus norvegicus & 8 h)	E-GEOD-2392
11	**R6/1 transgenic** vs wild_type in *Mus musculus* (27 weeks)	E-GEOD-3621
12	**R6/1 transgenic** vs wild_type in *Mus musculus* (18 weeks)	E-GEOD-3621
13	**9 d** vs 0 d in *Mus musculus* (embryoid body)	E-GEOD-2972
14	**3H-1,2-dithiole-3-thione** vs none in *Rattus norvegicus*	E-GEOD-3173
15	**kainate** vs control in *Rattus norvegicus* (240 h)	E-GEOD-1156
16	**diabetes mellitus** vs normal in *Rattus norvegicus* (vanadyl sulfate)	E-GEOD-3068
17	**1.5 d** vs 0 d in *Mus musculus* (embryoid body)	E-GEOD-2972
18	**adenoviral vector** vs none in *Mus musculus*	E-GEOD-3172
19	**FrCasE** vs mock infected in *Mus musculus*	E-MEXP-459
20	**spinal nerve transection** vs sham surgery in *Rattus norvegicus*	E-MEXP-976
21	**bipolar disorder** vs normal in *Homo sapiens* (male)	E-GEOD-5389
22	**severe spinal cord injury** vs normal in *Rattus norvegicus* (spinal cord (T10) & 2 d)	E-GEOD-464
23	**creatine** vs control in *Mus musculus*	E-GEOD-5140
24	**moderate spinal cord injury** vs normal in *Rattus norvegicus* (spinal cord (T10) & 2 d)	E-GEOD-464
25	**monocular deprivation right eyelid sutured** vs control in *Mus musculus*	E-GEOD-4537

This case study illustrates that the retrieval of comparisons can be based on tissue specificity, rather than on conditions such as diseases or treatments. Analysis of the most relevant gene sets for these comparisons would yield further insight into the specific processes that are affected in these comparisons. It is important to point out that general tissue specific expression patterns most likely are not the cause for the similarity observed between these comparisons, as the differential analysis is designed to remove these effects.

3.6.4 Lung Adenocarcinoma

Lung adenocarcinoma vs normal in human yields a high number of comparisons (14 out of 24) from three studies that investigated adipogenesis (E-GEOD-2192, E-GEOD-1123) or adipose differentiation (E-GEOD-2746), as

Table 3.6: Query results for a lung adenocarcinoma comparison. Text in parentheses after the name of the species is the context of the corresponding comparisons.

Q	**lung adenocarcinoma** vs normal in *Homo sapiens* (Homo sapiens)	E-GEOD-2514
1	**4 d** vs 0 in *Mus musculus* (early B-cell factor 1 transfection)	E-GEOD-2192
2	**48 h** vs 0 in *Mus musculus* (wild_type)	E-GEOD-2746
3	**24 h** vs 0 in *Mus musculus* (PKB alpha null)	E-GEOD-2746
4	**48 h** vs 0 in *Mus musculus* (PKB alpha null)	E-GEOD-2746
5	**2 d** vs 0 in *Mus musculus* (empty vector)	E-GEOD-2192
6	**RAD001** vs placebo in *Mus musculus* (AKT transgenic & 48 h)	E-GEOD-1413
7	**interferon-gamma** vs none in *Mus musculus* (Yersinia enterocolitica WA(pTTS, pP60) (control) & BALB/c)	E-GEOD-2973
8	**early B-cell factor 1 transfection** vs empty vector in *Mus musculus* (4 d) & E-GEOD-2192	
9	**4 d** vs 0 in *Mus musculus* (PPARgamma2 transfection)	E-GEOD-2192
10	**2 h** vs 0 in *Mus musculus* (PKB alpha null)	E-GEOD-2746
11	**Nrl null** vs wild_type in *Mus musculus* (2 days)	E-GEOD-8972
12	**adenocarcinoma** vs normal in *Homo sapiens* ("" & "" & prostate)	E-MEXP-1331
13	**sporadic basal-like breast cancer** vs normal in *Homo sapiens*	E-GEOD-3744
14	**Nrl null** vs wild_type in *Mus musculus* (60 days)	E-GEOD-8972
15	**4OH-tamoxifen** vs none in *Homo sapiens* (ERbeta transfected)	E-GEOD-2292
16	**retinoic acid** vs none in *Mus musculus* (12 h)	E-GEOD-1588
17	**10 d** vs 0 in *Mus musculus* (empty vector)	E-GEOD-2192
18	**24 h** vs 0 in *Mus musculus* (wild_type)	E-GEOD-2746
19	**4 d** vs 0 in *Mus musculus* (empty vector)	E-GEOD-2192
20	**IDMB** vs none in *Mus musculus*	E-GEOD-1123
21	**interferon-gamma** vs none in *Mus musculus* (uninfected & C57BL/6J)	E-GEOD-2973
22	**PKB alpha null** vs wild_type in *Mus musculus* (0)	E-GEOD-2746
23	**2 h** vs 0 in *Mus musculus* (wild_type)	E-GEOD-2746
24	**tristetrapolin-deficient** vs wild_type in *Mus musculus* (actinomycin D & 60 m)	E-GEOD-5324
25	**tristetrapolin-deficient** vs wild_type in *Mus musculus* (actinomycin D & 120 m)	E-GEOD-5324

shown in Table 3.6. Furthermore, a prostate adenocarcinoma, a sporadic breast cancer, and a osteosarcoma cell line treated with the cancer drug tamoxifen were found to be among the highly relevant results. Examples for gene sets that were found to be important for many of these comparisons are the ECM receptor interaction pathway, the focal adhesion pathway, which is downregulated almost in all comparisons, and the P53 signaling pathway.

The analysis of the results found in this case study did not provide any clear answer why the adipogenesis and adipose differentiation-related comparisons are highly relevant for the lung adenocarcinoma query. However, given that comparisons related to these processes were retrieved from three different studies, there is evidence that indeed there is a connection that an in-depth analysis could reveal.

Table 3.7: Query results for a fast food diet comparison. Text in parentheses after the name of the species is the context of the corresponding comparisons.

Q	**human fast food diet** vs control diet in *Mus musculus*	E-GEOD-6297
1	**human cafe diet** vs control diet in *Mus musculus*	E-GEOD-6297
2	**Wilms' tumor** vs normal in *Homo sapiens*	E-GEOD-2712
3	**influenza** vs uninfected in *Mus musculus* ("" & wild_type)	E-GEOD-3203
4	**POR null** vs wild_type in *Mus musculus* (none & colon)	E-GEOD-4262
5	**Lmna$^{H222P/+}$** vs wild_type in *Mus musculus*	E-GEOD-8000
6	**glipizide** vs none in *Mus musculus* (93 kcal per wk for 8 wk)	E-GEOD-2431
7	**very low fat, high carbohydrate diet** vs control diet in *Mus musculus* (wild_type)	E-GEOD-3889
8	**IL-26** vs control - untreated in *Homo sapiens*	E-GEOD-7216
9	**clear cell sarcoma of the kidney** vs normal in *Homo sapiens*	E-GEOD-2712
10	**chimpanzee diet** vs control diet in *Mus musculus*	E-GEOD-6297
11	**48 h** vs 0 h in *Mus musculus* (E2F2-/-)	E-MEXP-1413
12	**glycogen synthase -/-** vs wild_type in *Mus musculus* (liver)	E-GEOD-2198
13	**pkd1 -/-** vs wild_type in *Mus musculus* (flow)	E-TABM-411
14	**24 h** vs 0 in *Homo sapiens*	E-GEOD-620
15	**48 h** vs 0 in *Mus musculus* (gastrocnemius muscle)	E-GEOD-4463
16	**AKT transgenic** vs wild_type in *Mus musculus* (placebo & 48 h)	E-GEOD-1413
17	**2 ppm** vs 0 mM in *Rattus norvegicus* (inhalation & formaldehyde & 19 d)	E-GEOD-7002
18	**POR null** vs wild_type in *Mus musculus* (quercetin & jejunum)	E-GEOD-4262
19	**24 h** vs 0 in *Mus musculus* (wild_type)	E-GEOD-2746
20	**POR null** vs wild_type in *Mus musculus* (none & liver)	E-GEOD-4262
21	**growth hormone** vs control in *Mus musculus* (48 h)	E-GEOD-2120
22	**quercetin** vs none in *Mus musculus* (wild_type & jejunum)	E-GEOD-4262
23	**RP1 knockout** vs wild_type in *Mus musculus* (10 d)	E-GEOD-128
24	**c-MYC knockdown** vs control in *Homo sapiens* (HeLa & cervical cancer)	E-GEOD-5823
25	**glaucoma** vs normal in *Homo sapiens*	E-GEOD-2378

3.6.5 Fast Food Diet

An interesting result was found when the collection of comparisons was queried with *human fast food diet vs. control diet* in mouse. The top 25 retrieved comparisons are shown in Table 3.7. This query retrieved comparisons from two other studies that investigated the effect of different high calorie diets on livers (E-GEOD-2431 and E-GEOD-3889). Additionally, the query retrieved two other diet comparisons (cafe diet, chimpanzee diet) from the same study as the query (E-GEOD-6297). A commonality of these comparisons is that they were all performed on mouse livers. Overall there were 9 comparisons performed on mouse livers among the top 25 retrieved results.

Highly relevant gene sets for many of the comparisons retrieved by the fast food diet query are part of the carbohydrate metabolism and lipid

metabolism. The gene sets represent, for instance, butanoate metabolism, propanoate metabolism, pyruvate metabolism and fatty acid metabolism. This most likely reflects the fact that the effect of high calorie diets was analyzed in the corresponding studies. Many other top 25 gene sets belong to the amino acid metabolism category, for instance the glutathione, glycine, serine and threonine metabolism.

3.7 Discussion of the Extended Method

The original retrieval method described in 3.2 was extended in Section 3.5 to include a hierarchical latent component model and made applicable to larger and more complex collections of transcriptomics studies. The refined model can be applied to collections that span multiple species and a wide range of experimental factors, such as diseases, genotypes and drug treatments. A new evaluation methodology for the model, based on the Experimental Factor Ontology as a gold standard, was devised to handle the wide range of conditions included in the expanded collection of studies.

As with the original method, the extended method is applied to perform queries across the comparisons derived from the data sets in the collection. By providing a comparison as query, the investigator retrieves a list of comparisons that are ranked by their relevance to the query. Additionally, the investigator can examine the most relevant gene sets and the direction of their differential expression for every comparison. The utility of this approach was demonstrated in a series of case studies. For example, in one case study a link between cardiomyopathies and benign nevi/malignant melanomas was identified and determined to be an inverse relationship due to the direction of change in highly relevant gene sets. Further examination of the most relevant gene sets for the retrieved comparisons provided insight into the pathways that are affected in both disease categories, which confirmed a hypothesis reported in the literature (Hu and Agarwal, 2009). Another detailed case study showed how

an extensive literature analysis provided explanations for how conditions and processes such as insulin signaling, diabetes mellitus and inflammation are linked to the pancreatic cancer query. Several other case studies showed that the method is capable of retrieving related tissues (nervous tissues in glioblastoma case study), processes (adipogenesis in lung adenocarcinoma case study) and treatments (high calorie diets in fast food diet case study) from independent studies. It also has to be pointed out that the analyses presented in the case studies were not performed by domain experts. It can be expected that investigators with more knowledge about the underlying biology of the query will be able to provide better and more in-depth interpretations of the retrieval results.

The evaluation of the extended method against the EFO gold standard revealed that retrieval performance of the hierarchical model is comparable to the original model and retrieval based on TF-IDF as well as Spearman rank correlation, and that there is no significant difference between the methods. The major selling point of the method presented here is clearly the interpretability of the the results in conjunction with its competitive retrieval performance.

Some challenges were observed in the analysis and evaluation of the method that provide room for future improvements. For instance, as mentioned in Section 3.5.4, since the modules in the hierarchical model capture the correlation of the topics, it is not possible to directly assign specific processes or functions to them. However, this is still possible for the topics, and both topics and modules can be used as a starting point to explore the collection of comparisons.

A general weakness of the gene set-based approach is its low resolution. While it is possible to identify which pathways are affected in a comparison, it is not possible to determine directly which of the underlying genes are involved and which ones are not. Another potential problem that could occur is that the same pathway is affected in two comparisons but the affected genes in those comparisons are mutually exclusive. To address this problem, the gene level information incorporated into the hi-

erarchical model can potentially be used. However, this option has not yet been explored.

Applications of the method beyond heterogeneous collections of data sets from public repositories are a promising direction of research. For instance, running the method on large, but more homogeneous collections of data sets that contain less noise would be worth pursuing. One example of such a collection are the data sets of the Connectivity Map (Lamb et al., 2006). The Connectivity Map currently contains expression profiles from over 6,000 microarray studies performed on Affymetrix Human Genome U133 family microarrays, in which the effects of more than 1,300 drugs on five different human cell lines were investigated. Application of the retrieval method to this collection would provide insight into the effects of drugs and reveal, for example, which drugs cause similar reactions in the cell lines. Another option is an approach similar to the one reported by Hu and Agarwal (2009), who merged the Connectivity Map collection with public data sets to link drugs to diseases.

Since both the original and the extended method were shown to be useful tools to query large collections of heterogeneous data sets, an important next step is to investigate how these methods can be combined with the query engines of existing repositories. This would allow investigators to perform data-driven queries for comparisons extracted from suitable data sets directly on the content of the repository. Additionally, the data-driven query functionality can be extended to allow investigators to use their own data sets as queries. As only the output of gene set enrichment analysis is required for the query, lightweight and secure solutions could be implemented that do not require investigators to upload their original data to a web server. Such a solution could be applied in addition to the traditional "single data set analysis", to gain further insight into the data by putting it in the context of previous studies and their corresponding publications.

Contributions and Publications

The work presented in Sections 3.1 through 3.4 was published in similar form in the proceedings of ISMB/ECCB 2009 as: J. Caldas, N. Gehlenborg, A. Faisal, A. Brazma, and S. Kaski. Probabilistic retrieval and visualization of biologically relevant microarray experiments. *Bioinformatics* 25(12):i145-i153, 2009. J. Caldas contributed the design and implementation of the probabilistic model and A. Faisal wrote the software to evaluate the model. The extended method presented in Section 3.5 is the result of a collaboration with J. Caldas, who designed and implemented the hierarchical model and helped with the calculation of the retrieval performances.

Chapter 4

Visualizing Large Data Sets

4.1 Introduction

This chapter deals with the visual exploration of expression profiles from large, individual transcriptomics data sets. The approach described here complements the work presented in Chapters 2 and 3, where the goal was to explore large collections of transcriptomics studies and identify data sets that are relevant to a particular research question. During the last decade, high-throughput analysis of transcriptomes with microarrays and other technologies has become increasingly mature and affordable, which has led to an increasing number of transcriptomics data sets that include hundreds or thousands of samples. Large numbers of samples create problems for the visualization of expression profiles from those studies. These problems are discussed in this chapter and a solution is developed based on insights from an evaluation of existing visualization methods.

Measuring the abundance of transcripts allows investigators to infer which biological processes are active in a cell. By comparing these activity patterns across different conditions, phenotypes or time points, investigators can gain insight into gene function, regulation and interactions (Quackenbush, 2001). High-throughput technologies, along with increasing automation of laboratory procedures, have enabled measurement of transcript abundance for many thousands of genes simultaneously and

under a wide range of conditions, in a reasonable amount of time.

As discussed in Chapter 1, microarrays are the most commonly used technique to measure transcript abundance. They measure the expression levels of thousands of genes in an organism in a single experimental assay. The technique has been applied in a wide range of different areas, such as disease classification (Quackenbush, 2006) and biomarker identification (Nevins and Potti, 2007), drug discovery (Parissenti et al., 2007) and developmental biology (White, 2001).

The data is typically represented in the form of an expression matrix, which was introduced in Section 1.1.4. The expression matrix contains the expression levels as real numbers. Rows represent expression profiles across the samples, which are represented by the columns. Genes are associated with a range of attributes representing prior biological knowledge or information derived from the expression data themselves. The samples are annotated with information about their role in the experimental design, for instance, whether they are representing a control or a treatment.

The questions that investigators are trying to answer when analyzing the expression profiles in the matrix fit into two broad categories (Katagiri and Glazebrook, 2009). The first category is the identification of genes whose expression levels change with the labels of the conditions, for example when the organism is treated with a drug. In the second category, the objective is the identification of genes whose expression levels change in a similar pattern across many different conditions. This provides evidence for co-regulation of genes, which is an indicator for their involvement in the same biological process (Quackenbush, 2001). It is important to point out that the analysis of transcriptomics data is an open-ended, exploratory process that is primarily a means to formulate hypotheses, which are then tested either experimentally or computationally.

Visualization of gene expression profiles is an important component of the analysis, given its exploratory nature. State-of-the-art visualization techniques are heat maps and profile plots and their various extensions for expression profiles, as well as scatter plots for projections of

expression profiles into 2- or 3-dimensional space. These techniques have been implemented in a vast array of data analysis tools and software packages, which have recently been reviewed by Gehlenborg et al. (2010). The visualizations are applied in the context of data analysis techniques, such as statistical hypothesis tests and machine learning methods for data exploration that support the identification of patterns, like clustering. Interpretation of patterns is aided by the integration of class information and other previous knowledge into the visualizations.

Heat maps and profile plots are effective techniques to visualize expression profiles of hundreds of genes across a few dozen samples. However, these techniques do not scale to data sets with expression profiles that have been measured across hundreds or even thousands of samples.

The motivation to analyze and address this scaling problem is based on the observation that with increasingly mature and affordable microarray platforms, the number of studies that include hundreds of samples has been increasing over the years. At the end of July 2010, ArrayExpress (Parkinson et al., 2009) held over 12,900 studies. Out of those, 559 studies contained 100 or more samples. Figure 4.1 shows how the number of newly submitted studies with 100 or more samples has developed over the last seven years.

The first contribution of the work presented in this chapter is the identification of two scalability problems that occur when existing visualization methods for gene expression profiles are used to visualize hundreds or thousands of genes that have been measured across hundreds or thousands of samples. Based on a thorough evaluation of heat maps, profile plots and scatter plots for the visualization of gene expression profiles, a *graphical scaling problem* and a *perceptual scaling problem* were found. The graphical scaling problem is inherent to the design of heat maps and profile plots, whereas the perceptual scaling problem is independent of the methods used for visualization and created by the large number of samples that are being studied.

The second contribution of this chapter is the *Space Maps* visual-

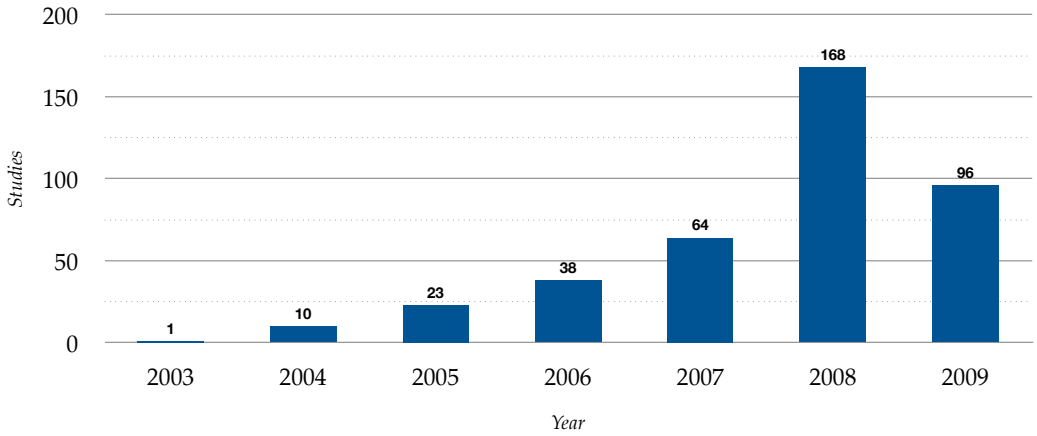

Figure 4.1: Newly submitted large data sets in the ArrayExpress Archive. The height of the bars represents the number of studies submitted in the corresponding year that comprised 100 or more samples. The drop in 2009 can, at least partly, be explained by the implementation of a data import pipeline from GEO into the ArrayExpress Archive that lead to a notably higher overall number of new studies in 2008 compared to previous years. *(Source: www.ebi.ac.uk/microarray-as/ae/browse.html, numbers reflect counts at the end of each year)*

ization technique, which addresses the scalability issues that were identified. The technique is based on the Value and Relation display (Yang et al., 2007) and employs tree map-like (Johnson and Shneiderman, 1991) hierarchical, multi-resolution glyphs based on *space-filling Hilbert curves*. In the visualization field a *glyph* is a compact, graphical representation of a multi-dimensional data point. See Section 4.3 for more details. Representation of expression profiles as 2-dimensional glyphs addresses the graphical scaling problem and the introduction of multi-resolution glyphs is a remedy for the perceptual scaling issue. The *Space Maps* technique can be used to visualize and explore hundreds of gene expression profiles with several thousand samples.

The remainder of this chapter is organized as follows: In Section 4.2 the data, tasks and state-of-the-art visualization methods in gene expression data analysis are reviewed; in Section 4.3 the Space Maps technique is introduced; in Section 4.5 several case studies of the Space Maps technique applied to a real world data set are presented that illustrate how the technique can be applied to large transcriptomics data sets. Finally, in Sec-

tion 4.6 the findings are discussed and directions for further research are suggested.

4.2 Visualization in Data Analysis

4.2.1 Data Structure and Content

The number of profiles in the expression matrix depends on the organism that is being studied. Whereas the yeast *S. cerevisiae* has 6,532 known protein-coding genes (aug2010.archive.ensembl.org/Saccharomyces_cerevisiae/Info/StatsTable), many higher organisms have over 20,000 known protein-coding genes, for instance humans, who have 21,257 genes (aug2010.archive.ensembl.org/Homo_sapiens/Info/StatsTable), and a much larger number of different transcripts.

Before any analysis begins, missing values in expression profiles are either imputed or the corresponding genes are removed, as described in Section 1.1.4. Furthermore, the expression matrix is normalized to remove any systematic variation that occurred during the measurement of the individual conditions. The columns of the expression matrix are the dimension axes of the "gene expression space" and each expression profile corresponds to a point in this space. With this interpretation, it is straightforward to define similarity measures between expression profiles and to apply methods such as clustering, outlier detection and other pattern discovery methods.

The analysis of expression profiles is usually performed within the context of what is already known about genes and the studied conditions. This knowledge is typically provided as meta information associated with the rows and columns of the expression matrix, as illustrated in Figure 4.2.

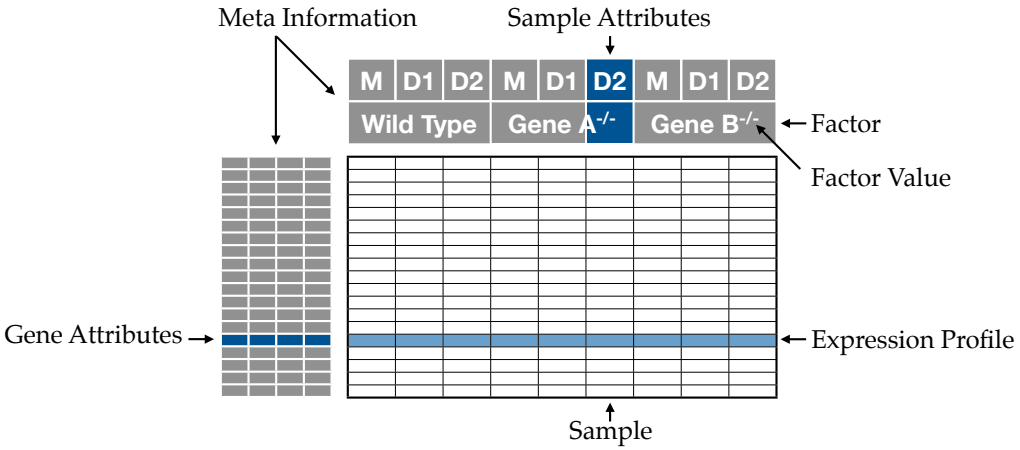

Figure 4.2: Expression matrix with associated gene and sample attributes. See Figure 1.1 for the meaning of factor values shown here.

4.2.1.1 Gene Attributes

Every gene can be associated with what can be called *knowledge-derived attributes*, for example membership in a particular functional class, regulation through a particular regulator protein or association with a particular disease. Gene names and textual descriptions are further types of knowledge-derived gene attributes.

In contrast to knowledge-derived gene attributes, *data-derived attributes* are obtained from the expression profiles themselves. These can be simple summary statistics such as mean or variance of the expression profile, but can also be more sophisticated statistics such as p-values for differential expression between two sets of conditions. If clustering has been applied, cluster membership can be assigned as an attribute. The rank of a gene according to some summary statistic is another example of a data-derived attribute.

Gene attributes may induce a total or partial ordering on the expression profiles. Summary statistics or p-values can be converted into ranks that define a total ordering, while a partial ordering can be derived from class attributes. Some knowledge- and data-derived class attributes can be mapped to a hierarchy, which reduces the degrees of freedom for

conversion into partial ordering. For instance, class attributes corresponding to cluster memberships obtained from hierarchical clustering have a natural hierarchical relationship. Class labels obtained from the Gene Ontology (Ashburner et al., 2000) can be treated similarly.

In general, the selection of gene attributes used in the analysis depends heavily on the biological question one is trying to answer. The key observation is that every gene may be associated with a whole array of textual, categorical, ordinal and ratio attributes that the investigator may wish to include in the analysis, as well as in the visualization.

4.2.1.2 Sample Attributes

Samples also have a number of associated attributes. As described in Section 1.1.4, the experimental design assigns every sample at least one experimental factor and one of its associated factor values. For example, if the experiment is designed to detect differences between a knock-out genotypes and a wild type strain, one factor would be "Genotype". Factor values are instances of a factor; in this example, the factor values could be "Wild Type", "Gene $A^{-/-}$" and "Gene $B^{-/-}$".

Often, more than one factor value is associated with a sample. Building on the previous example, the investigator could, for instance, be interested in the effect of particular drugs on the expression levels in the genotypes. The second factor is then "Drug" and the associated factor values may be "Mock", i.e. a control treatment, "Drug 1" and "Drug 2". Each sample is then assigned a combination of two factor values, one from "Genotype" and one from "Drug", yielding nine possible combinations. However, it is common to find only a subset of the theoretically possible combinations in the data.

In the work presented here, the case of experimental factors that impose a total ordering on the samples, such as time or concentration levels, is not considered. Studies that include these factors typically do not result in a large number of samples.

In the definition used here, a condition is synonymous with a bio-

logical sample - such as a tissue sample from a biopsy or a cell culture - in which the expression levels of transcripts are measured. As discussed in Section 1.1.3, expression levels under a given condition are almost always measured in biological and technical replicates in order to be able to control for variability in the biological samples and in the measurement technology (Allison et al., 2006). Replicates are aggregated by computing statistics such as mean or median, and for most aspects of the analysis the aggregate is used rather than the individual measurements. However, when expression profiles are visualized, it is desirable to show the individual measurements as well, because they provide insight into the amount of variance contained in the data.

Aggregation of replicates creates a flat hierarchy on the samples where the aggregate is parent to its constituent samples. It is very often possible to construct a deeper, knowledge-derived hierarchy based on samples where aggregates are biologically meaningful summaries of the samples. Either a problem-specific hierarchy is constructed by the investigator or a hierarchy can be generated by mapping the samples to ontologies, such as the Experimental Factor Ontology (Malone et al., 2008, 2010). More specialized ontologies such as the Disease Ontology (Dyck and Chisholm, 2003) or Cell Type Ontology (Bard et al., 2005) may be used, depending on the samples that are being studied.

If it is not possible, or not beneficial, to construct a knowledge-derived hierarchy, a data-derived hierarchy can be created, for example, by hierarchical clustering. In this case, aggregates may not have an immediate biological interpretation, but this may lead to discovery of groups of related conditions.

4.2.2 Tasks in Gene Expression Data Analysis

It is important to point out that the expression matrix is a symmetric object[1] that can be analyzed either in a gene-centric or a sample-centric way.

[1]However, the expression matrix is usually not a symmetric matrix in the mathematical sense.

On the one hand, an investigator might be interested in finding sets of genes that exhibit similar expression levels across a range of samples, e.g. to discover a biological pathways. On the other hand, another investigator might be interested in finding sets of samples in which genes are exhibiting similar expression levels, e.g. to discover subtypes of a disease or to classify samples. In both the gene-centric and the sample-centric way the same computational methods would be used, for instance clustering. Since most visualization methods are typically applied for the study of gene expression profiles rather than samples expression profiles, and also for simplicity, the work presented in this chapter is discussed from the gene-centric perspective.

The goals of gene expression data analysis, the methods that are employed and the prior knowledge that is integrated vary greatly depending on the area of application. In one study, the goal could be the prediction of gene function for genes that have not yet been annotated, by finding genes with similar expression profiles and following a "guilt by association" approach (Quackenbush, 2003). The investigator in another study might be interested in identifying genes whose expression levels are affected by a particular disease in order to devise a test that can used to diagnose the disease prior to any clinical signs (Hwang et al., 2009). Another investigator might want to use gene expression data to build a predictive model for the regulation of transcription (Bonneau, 2008).

Based on experience and literature (Slonim, 2002; Quackenbush, 2001) on transcriptomics data analysis, a core set of analytical tasks that are part of almost every analysis were identified. These tasks are essentially pattern identification and interpretation tasks that are performed on the annotated expression matrix. Table 4.1 provides an overview of these tasks, which represent combinations of the low-level components of analytic activity that Amar et al. (2005) have identified. They are employed here to discuss the requirements for visualization techniques that are applied in gene expression data analysis.

The tasks presented in Table 4.1 are not independent, but are usu-

Table 4.1: Low-level analytical tasks in gene expression analysis and how visualization can be used to support these tasks. Activities defined by Amar et al. (2005) are set in italics.

Task	Visualization
T1 (*Retrieve Value*): Look up expression level of a given gene in a given sample.	Provide the whole profile as context for the particular measurement.
T2 (*Correlate*): Find out under which samples or factor values a given gene has a particularly high or low expression level.	Present profile in a way that allows efficient identification of relatively high or low expression levels and the corresponding samples.
T3 (*Extrema, Range, Characterize Distribution*): Determine the range of expression levels in a given profile and how much they vary across the profile.	Present profile so that the range and distribution of expression levels can be evaluated efficiently.
T4 (*Sort*): Find a given number of profiles that are most similar to a given profile.	Present similar profiles in the context of all other profiles to allow evaluation of similarity in relation to similarities in the overall data set.
T5 (*Cluster*): Identify clusters of genes whose expression profiles are similar across all samples.	Present clusters in the context of the overall set of profiles for evaluation of the goodness of the result. Integrate class information and other gene attributes into representation for interpretation of results.
T6 (*Correlate*): Determine which genes are differentially expressed in one or more samples.	Show profiles of differentially expressed genes in context of other profiles.
T7 (*Find Anomalies*): Identify single or small sets of atypical expression profiles (outliers).	Provide an overview of the expression space that helps to identify why profiles are outliers.
T8 (*Characterize Distribution*): Determine how profiles are distributed in expression space.	Provide overview of distribution of expression profiles in the expression space that allows efficient identification of clusters and outliers.
T9 (*Filter, Characterize Distribution, Correlate*): Determine how profiles of genes with a given attribute or attribute range are distributed in the expression space.	Provide overview of location of the selected profiles in the expression space and remaining profiles as context. Present expression profiles so that their location can be interpreted through patterns in profile.
T10 (*Characterize Distribution*): Given a set of profiles determine how the values of a given gene attribute are distributed.	Provide overview of distribution of attributes within the set as well as in the overall expression space as context.

ally tightly integrated in an exploratory analysis process. For instance, the identification of differentially expressed genes (T6) and examination of their expression profiles (T1 - T3) are naturally connected. Another example is the combination of clustering (T5) and analysis of how the expression profiles of the cluster are distributed (T8) or how gene attributes are distributed in the cluster (T10). In many of these tasks, one or more computational methods are applied and the results of these methods are then visualized. The role of the visualization is to provide an overview so that the results can be assessed efficiently and judged for their reliability. Furthermore, the visualization should reveal unexpected properties of the data and generate hypotheses that can serve as the basis for further analysis.

Low-level analysis tasks can be further divided into identification and interpretation of local and global patterns. Local patterns are patterns within an expression profile, and global patterns are patterns across the profiles in the expression matrix. Global patterns arise directly from the local patterns. T1 through T3 are primarily concerned with local patterns while T4 through T10 are concerned with global patterns. The distinction is made because the local and global patterns pose different challenges for the visualization techniques that are discussed in Section 4.2.3.

4.2.3 Evaluation of State-of-the-Art Visualization Techniques

It is common in transcriptomics data analysis to reduce the number of expression profiles before visualizations are applied. Subsets can be created based on the variance of expression profiles, on p-values from a statistical test or similar measures. No matter which method is chosen to select the subset, in the following discussion it will be assumed that the investigator has selected a set of a few hundreds of genes. Such reductions are justified, for instance, because typically only a small fraction of the genes in an organism are involved in a disease or in the response to a drug, and the

majority of the genes will not be affected.

Removal of samples from the data set for visualization or any other purpose is hardly ever performed and usually this is only the case when the quality of the data for the sample is very low. The reason for this is that expression levels are measured under a particular condition because the investigator deemed it relevant to understanding the disease or biological system that is being studied. There is also a financial aspect to be considered, because for every condition one or more biological samples have to be taken and the gene expression levels have to be measured on a separate microarray, which generates additional costs for personnel and materials.

Before heat maps, profile plots and scatter plots are discussed in detail, it is important to emphasize that many sophisticated and well-designed visualization techniques are available to explore and analyze transcriptomics data in the context of other data types, such as networks or sequences. One alternative approach is *correlation networks* based on gene expression profiles, which are highlighted in the discussion about scatter plots. There are also many advanced implementations and extensions of the techniques that are discussed here, but they all have the same limitations when it comes to visualizing expression profiles that have been measured across several hundred samples or more.

4.2.3.1 Heat Map

Wilkinson and Friendly (2009) provide an overview of the history of the heat map visualization and similar visualizations had been used in statistics for a long time. The heat map was first introduced to the field of gene expression data analysis by Eisen et al. (1998) and in similar form to pharmacology by Weinstein (1997). It is the most commonly used visualization technique for gene expression data and has been referred to as "a post genomic visual icon" due to its appearance in thousands of biomedical publications (Weinstein, 2008).

The heat map is essentially an expression matrix in which the ex-

pression levels have been color-coded by mapping them to a color ramp. If the expression levels have been *log*-transformed and are both positive and negative, a diverging color ramp is used and if the values are only positive, a sequential color ramp is used. If the matrix is sufficiently small, gene and sample names or other textual attributes can be placed on the rows and columns, respectively. Categorical and quantitative attributes can be color-coded and added as additional columns to the matrix, as shown in Figure 1.6 (**c**). Furthermore, quantitative variables can be used to scale the height of individual rows in the heat map or be integrated into the colors encoding the expression levels (Gehlenborg et al., 2005).

The most important feature of the heat map is that rows and columns can be reordered independently, like in Bertin's reorderable matrix (Bertin, 1983). Row reordering is always applicable to expression matrices because typically there is no inherent ordering of the genes. The reordering of the columns of the expression matrix depends on the experimental setup, which sometimes implies an ordering of the columns, for instance in time-series experiments, that prevents the reordering of columns in the visualization.

Orderings obtained through hierarchical clustering of the rows in the matrix will bring together similar expression profiles. This allows the investigator to identify clusters (T5) efficiently. If the rows are reordered according to the similarity to a query profile, it is straightforward to examine the similarity to the query in the context of the other profiles (T4). If samples are grouped according to their factor value, attribute genes with significant changes in their expression levels across factor values are easy to identify (T6).

A problem of the heat map is that its matrix structure prevents effective encoding of expression values as locations in the visualization. When the rows of the matrix have been reordered, the position of a profile may be considered to be a location-based encoding, however, it is only 1-dimensional and based on ranks, rather than distances. In such a setup, it is hard to detect outliers or to determine how close two profiles really

are to each other, with respect to the other profiles in the expression matrix (T7, T8, T10). However, if the heat map is combined with a visualization of a dendrogram as shown in Figure 1.6 (**c**), the additional information provided by the dendrogram can be used to identify outliers and to provide a better overview of the relative distances between expression profiles. It is also possible to modify the color mapping so that extreme values are emphasized.

If the expression profiles in the matrix only have several dozen samples, local patterns can be analyzed efficiently (T1 - T3). However, if expression profiles have hundreds, or even thousands, of samples this has a severe impact on the performance of the heat map in local pattern discovery tasks. Since the heat map grows horizontally with every sample, heat maps with hundreds of samples do not fit on a single screen page unless the number of pixels per sample is reduced drastically. However, even if every sample is represented only by a single pixel, expression profiles from data sets with several thousand samples will not fit on a single screen page, so this is not an effective solution.

The main problem caused by profiles not fitting on a single screen page is that important context information that is required for various local pattern discovery tasks is not accessible. In an interactive setting the information may be accessed through scrolling and panning, but only at the cost of other parts of the profile becoming invisible. This requires the investigator to remember those parts of the expression profile, because visual queries are not possible. However, given the limited capacity of human visual working memory, this is bound to fail (Ware, 2004, Chapter 11). This is a graphical scaling problem inherent in the design of the heat map, which renders it ineffective for visualization of expression profiles with hundreds or thousands of samples. Figure 4.3 illustrates why even huge displays are not suitable to visualize very large heat maps.

Besides this graphical scaling problem, a perceptual scaling problem (Jost and North, 2006) can be observed, which is caused by the large number of samples and is not a problem of the heat map *per se*. For in-

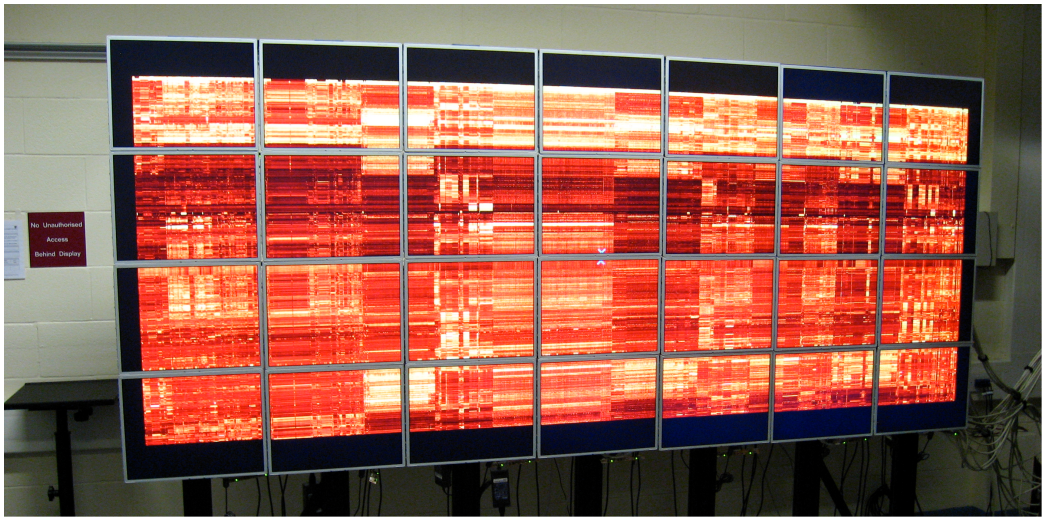

Figure 4.3: Large heat map visualization on a powerwall display consisting of 28 20″ panels with a total resolution of 11,400 x 4,800 pixels. The expression matrix shown here contains 1,000 genes measured across 5,372 samples (see Section 4.5 for a description of the data set) and in the resulting heat map each gene is 4 pixels high and each sample is 2 pixels wide. For comparison, the powerwall display is approximately 3 meters (10 feet) wide. (*Photo taken in the facility of the Visualization and Virtual Reality Group, School of Computing, University of Leeds.*)

stance, in T2 the investigator has to associate samples with expression levels to form a mental model of how a transcript behaves under a range of samples. With hundreds of samples, this task is almost impossible to achieve, unless the information is presented in a more structured way that provides an overview first, before presentation of the details.

4.2.3.2 Profile Plot

Profile plots, also known as *parallel coordinate plots* (Inselberg, 1985) in the visualization field, are another very common technique used to visualize expression profiles. A sample profile plot is shown in Figure 1.6 (**b**). In profile plots the dimension axes of the expression space are arranged in parallel and orthogonal to a horizontal axis. The parallel axes are equidistant and have the same scale. Since the dimension axes of the expression space are equivalent with the samples, an expression profile is visualized as a line intersecting dimension axes at the point corresponding to its ex-

pression level under the sample represented by the axis. An inherent disadvantage of this construction is that expression profiles are plotted on top of each other, which can be problematic even when the number of profiles is relatively small. This can be observed in the example shown in Figure 1.6 (**b**), where the profiles from the yellow cluster are hidden behind the other profiles. Kincaid and Lam (2006) address this issue in their Line Graph Explorer tool, but assume a fixed linear ordering of the samples. In the discussion here the assumption is that there is no fixed linear ordering of the samples.

Gene attributes can be incorporated into the plot by encoding them as attributes of the line representing the expression profile. Color can be used to encode categorical or quantitative variables as shown in Figure 1.6 (**b**), where cluster membership is encoded in the color of the expression profiles. Quantitative variables relating to the relevance of a gene can be encoded in the alpha channel of the line color or, to a very limited degree, also in line thickness. If the same variable is used to determine the plotting order of the profiles, this can be used to highlight the most relevant genes on top of the less relevant ones.

Profile plots are efficient for the exploration of local patterns (T1 - T3) if the number of profiles that provide the global context is small. Interpretation of clusters and their relationship to each other (T5, T8, T9) can be facilitated by encoding cluster membership in line color. However, parts of, or even a whole cluster of profiles, may be hidden behind the expression profiles of another cluster, which also applies to individual profiles that potentially would be classified as outliers (T7). This makes the exploration of cluster structure and outliers rather tedious. Similarity to a given profile (T4) can be encoded in plotting order and line color to support interpretation. This is quite efficient despite over-plotting, which in this case primarily affects less relevant, dissimilar profiles. The characterization of a set of expression profiles in terms of a gene attribute is well supported in profile plots (T10) as is the comparison of differentially expressed genes (T6) with other profiles. The latter can be achieved through

column-reordering, color-coding and selection of a sensible plotting order.

Apart from the inherent over-plotting issues, profile plots extend horizontally with every additional sample and eventually grow beyond the width of the screen. Attempts to counter the growth along the horizontal axis by decreasing the space between adjacent samples will make the plot difficult to interpret. The angle of line segments with respect to the horizontal axis will become close to 90 degrees unless vertical scaling is reduced to maintain an angle closer to 45 degrees, which is preferred (Heer and Agrawala, 2006). However, vertical scaling can only be reduced to a certain degree before the profiles become too flat to convey any information about the expression levels. Generally, the problem of horizontal expansion is worse in profile plots than in heat maps.

4.2.3.3 Scatter Plot

Scatter plots are not a technique for visualization of expression profiles *per se*, but represent an effective method to visualize the expression space in combination with dimensionality reduction techniques. For the purpose of visualization, dimensionality reduction methods are used to obtain projections from a high-dimensional space to a 2- or 3-dimensional space, which means that a gene expression profile as shown in Figure 1.6 (**b**) is represented by a mark at a 2-dimensional coordinate in a scatter plot, which is illustrated in Figure 1.6 (**a**). The interpretation of the dimensions in the projection depends on the method used to obtain the projection.

Common dimensionality reduction methods are the linear *Principal Components Analysis* (PCA) (Hotelling, 1933), as shown in Figure 1.6 (**a**), or variants of *Multi-Dimensional Scaling* (MDS) (Kruskal, 1964). Even though PCA is probably the most frequently applied method for this purpose in gene expression data analysis due to the rather straightforward interpretation of the dimensions in the projection, it is outperformed by many other methods (Venna and Kaski, 2007b). More recent methods that were found to perform well are, for example, the nonlinear methods *Locally Linear Embedding* (Roweis and Saul, 2000), *Isomap* (Tenenbaum et al.,

2000) and NeRV (Venna and Kaski, 2007a). Other approaches that have been applied for dimensionality reduction are based on *Non-Negative Matrix Factorization* (Lee and Seung, 1999).

By definition, dimensionality reduction leads to a loss of information and the goal of the aforementioned methods is to maintain as much information as possible in the lower dimensional space. The optimization criteria used for this purpose vary widely. The optimization process can also be supervised, i.e. guided by incorporating knowledge about the data points, in order to find a better representation. Methods in this category include *Supervised Local Linear Embedding* (DeRidder et al., 2003), *Supervised NeRV* (Peltonen et al., 2009) and a related method developed specifically for dimensionality reduction of expression profiles that uses the Gene Ontology to guide the optimization process (Peltonen et al., 2010). One disadvantage of these methods is that they are computationally more complex and thus require more resources in order to be applied to data sets with many thousands of genes.

Given a suitable dimensionality reduction method, expression profiles are projected from the high-dimensional expression space to a 2- or 3-dimensional projection space. For such projections scatter plots and their interactive extension to starfield displays (Ahlberg and Shneiderman, 1994) are the most suitable technique to visualize and explore the distribution of profiles in the expression space (T8) and to identify outliers (T7). They are also highly efficient in for the identification of clusters (T5) and genes with related expression profiles (T4).

A method related to scatter plots, that often performs similarly well on these tasks is *correlation networks* derived from transcriptomics data (Lee et al., 2004; Freeman et al., 2007). To generate a correlation network, the pairwise correlation between all expression profiles is computed, and with the help of a minimum correlation threshold a weighted adjacency matrix of all genes is constructed. This adjacency matrix can then be visualized as a graph by applying appropriate graph layout algorithms. In the graph, nodes represent genes and an edge between two nodes represents

a notable correlation between the expression profiles of the corresponding genes. When force-directed graph layout algorithms are applied, highly connected nodes representing groups of genes with correlated expression profiles are placed in close proximity and form clusters (Freeman et al., 2007). This allows the investigator to identify clusters (T5) and genes with related expression profiles (T4). An advantage of this approach is that once a network has been constructed, a wide range of network analysis algorithms can be applied to analyze the expression data.

A range of gene attributes can be encoded in scatter plots and correlation networks, which support tasks that require correlation of gene attributes with the expression profiles (T9, T10). Data points, or nodes in the case of graphs, can be colored according to class labels or, if the number of classes is low, a distinct shape such a circle, square or cross, can be assigned to each class. Quantitative variables can be encoded by mapping them to an appropriate color ramp and using the color to draw the data points. Another useful encoding of quantitative variables is in the alpha channel. For instance, if the variable is a measure for the relevance of a gene such as a p-value or a rank less relevant genes will fade into the background color, which creates an effect similar to semantic depth of field (Kosara et al., 2001). Quantitative variables can also be encoded in size or converted into a plotting order if a large number of data points are visualized.

While scatter plots and correlation networks are an effective technique for exploration of global patterns in the expression space, they fall short of providing insight into the local patterns that gave rise to the observed structure because the original expression profiles are not visible. A solution to this problem is to use brushing and linking to connect an interactive scatter plot or network visualization to a profile plot or a heat map where the actual expression profile can be examined.

The interpretability of scatter plots and correlation networks is not directly affected by the number of samples in the expression matrix since by definition they only display data in a 2- or 3-dimensional projection

space. This also makes them suitable for the visualization of sample profiles. However, due to the dependency of scatter plots on other methods for exploration of local patterns (T1 - T3), they are affected by the shortcomings of these methods when expression profiles have a large number of samples.

4.3 Space Maps: Extension of the Value and Relation Display with Hierarchical Glyphs

It is evident that the graphical scaling problem that arises when heat maps or profile plots are applied to visualize expression profiles with hundreds of samples is due to their construction. Since both techniques grow expression profiles only in the horizontal dimension, they use the available screen space very inefficiently and already the visualization of a single profile fails. In order to fit an expression profile with hundreds or even thousands of samples on a single screen page, a compact representation is required that keeps the aspect ratio[2] of the profile closer to the aspect ratio of the screen, for instance by using both the horizontal and the vertical dimension.

4.3.1 Pixel-Oriented Visualizations Techniques

As discussed earlier, gene expression profiles are essentially multi-dimensional vectors. Multi-dimensional vectors can be represented in form of glyphs, which are compact, graphical representations of multiple values. To represent an n-dimensional vector a glyph with n features is required. Figures 4.4 (**a**) and (**b**) show *Chernoff faces* (Chernoff, 1973) and *star glyphs*, which are classic examples of glyph visualizations (Siegel et al., 1972). However, these visualizations are not suitable for vectors with hundreds or even thousands of dimensions. They also have been found to be

[2]The aspect ratio of a rectangle is the length of its longer side divided by the length of its shorter side.

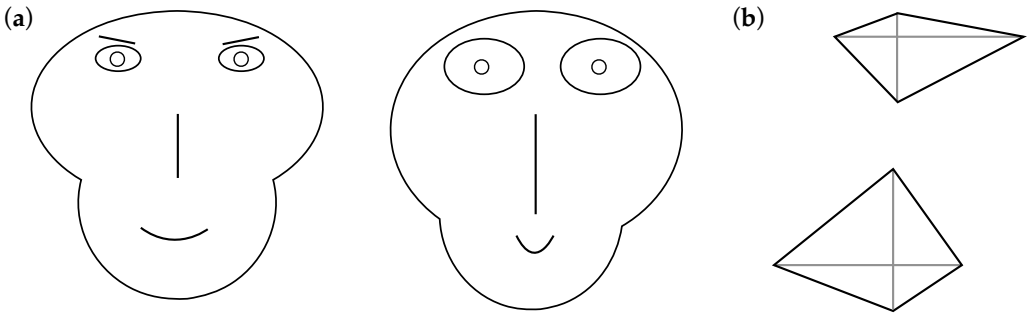

Figure 4.4: Glyphs for multi-dimensional data. (**a**) A classic example is so-called *Chernoff faces* (Chernoff, 1973). Up to 18 values can be encoded in the features of the face, including the length of the nose, curvature of the mouth, size of the eyes and shape of the head. (**b**) *Star glyphs* (Siegel et al., 1972) encode values in the length of evenly spaced rays drawn from a center point. Often the tips of neighboring rays are connected. The examples shown here represent 4-dimensional data points.

difficult to interpret, because, among other reasons, they are not making use of the human preattentive processing capabilities (Morris et al., 2000; Lee et al., 2003).

Another technique used to visualize multi-dimensional vectors is to map every data point to a single pixel of the display. In these so-called *pixel-oriented visualizations* (Keim, 1996) values are encoded in the color of the corresponding pixel as illustrated in the example in Figure 4.5 (**a**). Since every value is represented only by a single pixel, the main problem is the arrangement of the pixels in the display. Commonly, the display is divided into rectangular *subwindows*[3] and each vector is assigned a subwindow. The pixels belonging to a vector are then arranged within the area of corresponding subwindow by following a path that passes exactly once through every pixel of the subwindow. The mapping from dimensions to positions on the path, i.e. pixels in the subwindow, is based on either an implicit or an explicit ordering of the dimensions in the vector. The paths used for this purpose are typically *discrete space-filling curves*, which are discussed in detail in Section 4.3.2, or other highly structured recursive patterns.

[3] As discussed by Keim (2000), other shapes, such as wedges of circles, are also possible.

Pixel-oriented visualizations were originally designed to deal with a small number of very high-dimensional vectors and no appropriate solutions existed for cases when there is a large number of such vectors. The *Value and Relation* (VaR) display proposed by Yang et al. (2004, 2007) addresses this problem. Figure 4.5 (**b**) illustrates the construction of a VaR display. In the VaR display pixel-oriented glyphs are used to convey information about patterns in high-dimensional vectors. Furthermore, the glyphs are placed in 2-dimensional space so that the relationships between the vectors are preserved in the spatial layout of the glyphs. An example, in which MDS was used to project the high-dimensional vectors to determine the locations of glyphs, is shown in Figure 4.5 (**c**). Essentially, a VaR display is a scatter plot in which every data point is represented by a glyph. Besides dimensionality reduction methods, a regular grid can be used to arrange the glyphs in the VaR display, which is discussed in detail in Section 4.3.3.

The Space Maps visualization technique described in this chapter is an extension of the VaR display[4] and is based on the task analysis presented in Section 4.2.2. Like the VaR display, the method is a two-step approach: First glyphs are generated that convey information about local patterns within expression profiles and then these glyphs are placed on a plane so that global patterns in the expression space can be identified and interpreted efficiently.

4.3.2 Pixel-Oriented Glyphs for Expression Profiles

Local pattern exploration tasks (T1 - T3) require that all expression levels of a profile are visible on the screen. As discussed earlier, the essentially 1-

[4]It has to be pointed out that in the literature about pixel-oriented visualizations and the VaR display each glyph represents a dimension of some data space, i.e. typically a column of the data matrix. However, in the in the definition of the Space Maps technique and throughout this chapter each glyph represents a gene expression profile, i.e. typically a row of a data matrix. Since a matrix can easily be transposed, this is purely a difference in nomenclature, but potentially confusing for those familiar with the literature on pixel-oriented visualizations.

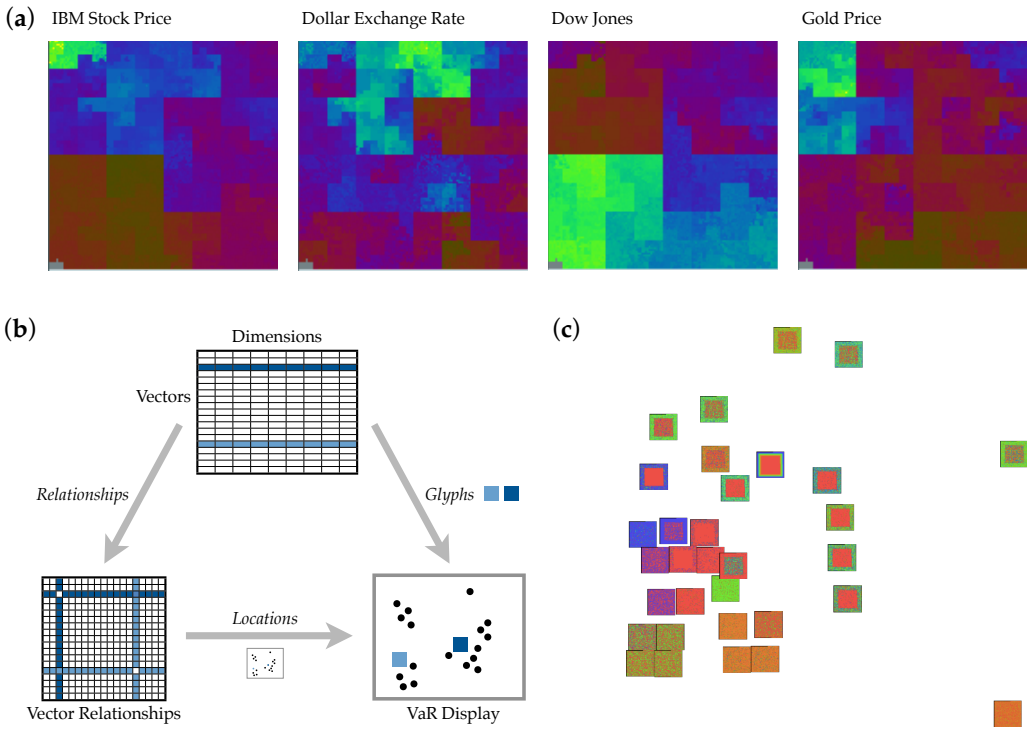

Figure 4.5: Pixel-oriented visualizations of multi-dimensional data. (**a**) A pixel-oriented visualization of IBM stock price, United States Dollar exchange rate, Dow Jones and gold price from January 1987 to March 1993. Each day in this interval is represented by nine values, resulting in a total of 16,350 values for each item. *(Image modified from Keim, 1996)* (**b**) Construction of the VaR display after Yang et al. (2007). Relationships between vectors determine the location of the glyphs in the VaR display and are usually distance or correlation measures. This step essentially results in a scatter plot visualization. The glyphs represent the values of the vector and are typically based on pixel-oriented visualizations. (**c**) A VaR display showing pixel-oriented glyphs arranged by multi-dimensional scaling (MDS) (Yang et al., 2004). Each glyphs represents 20,000 values. (*Image modified from www.cs.uncc.edu/~jyang13/graphhics21.htm with permission from the author.*)

dimensional representation of expression profiles in heat maps and profile plots causes these methods to fail on the these tasks when the profiles have hundreds or thousands of samples. In order to fit the complete expression profile on a single screen page, the samples are mapped to the cells of a rectangular grid that has an aspect ratio close to 1. If each cell of the grid has the size of one pixel[5], this approach is effective even for profiles with many thousand samples. This is due to the number of rows and columns

[5]In practice, this can often be scaled up to squares of 2×2, 3×3 or even more pixels.

of the grid being proportional to the square root of the number of samples in the profile.

Using a glyph representation with an aspect ratio close to 1 has several further advantages besides making better use of the available screen space: In cases where the investigator needs to interact with the visualization, they are easier to select with the mouse (Fitts, 1954) and they make much better use of the human visual field. *Acuity* - and with it the ability to see detail - drops rapidly when the distance to the fovea increases (Ware, 2004, see Chapter 2). This indicates that a representation with an aspect ratio close to 1 is more efficient because it requires fewer eye movements and less re-focusing to perceive it.

The challenge is to find a mapping of the samples to locations on the grid that supports the investigator in identifying and interpreting patterns in the expression profile. Wattenberg (Wattenberg, 2005) suggests the application of *discrete space-filling curves*[6] for this problem. Space-filling curves are bijections, which map every point on a 1-dimensional sequence to exactly one point on a 2-dimensional grid, and vice versa. Here space-filling curves are of interest that map points that are close together in the 1-dimensional sequence so that they end up close together in the 2-dimensional grid representation because they keep related samples in close proximity to each other. In computer science this is also known as the *locality* property of grid-indexings (Gotsman and Lindenbaum, 1996). Many different space-filling curves are known, but their ability to preserve locality varies greatly. Details are beyond the scope of this dissertation, but Figure 4.6 shows examples of the locality preservation achieved by three different space-filling curves.

Essentially, in space-filling curves that are better at preserving locality, the radius of the circle around a randomly chosen subsequence will be smaller. Figure 4.6 illustrates that the Hilbert curve (Hilbert, 1891) performs much better than the zig-zag and the spiral curve. In fact, the Hilbert

[6]Throughout this thesis "space-filling curve" and "discrete space-filling curve" are used interchangeably.

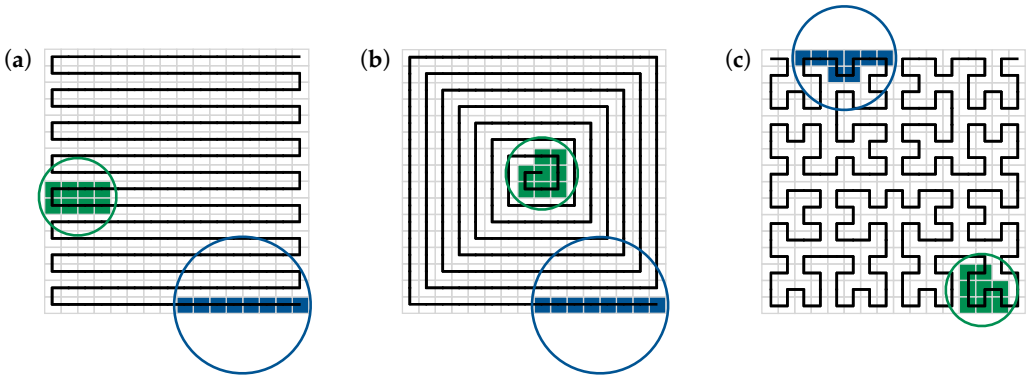

Figure 4.6: Locality of space-filling curves. An array of eight consecutive positions in the 1-dimensional sequence were chosen that mapped to the best locality preserving region (green) and the worst locality preserving region (blue) of the curve. (**a**) Zig-zag curve. (**a**) Spiral curve. (**c**) Hilbert curve (Hilbert, 1891). The smaller the radius of the circle around the colored area, the better the locality preservation.

curve is known to be one of the best locality preserving space-filling curves (Gotsman and Lindenbaum, 1996). However, Niedermeier et al. (2002) showed that with the so-called *H curve*, an even better space-filling curve exists and that this curve likely represents an optimal solution.

In an influential paper about pixel-oriented visualization techniques Keim (2000) studied the performance of several space-filling curves like the Hilbert curve, the *Morton curve*, the zig-zag curve and the spiral curve, as well as other highly structure recursive patterns. Keim found that the Hilbert curve is most suitable space-filling curve for the purposes discussed here[7]. A further reason to chose the Hilbert curve is that it can be generated with little effort based on an *L-system* (Lindenmayer, 1968) that uses string-rewriting to generate recursive structures. The L-system used to generate the Hilbert curve here is the following (after Peitgen and Saupe, 1988):

- Initialize with "L"
- "L" $\rightarrow$ "+RF-LFL-FR+" and "R" $\rightarrow$ "-LF+RFR+FL-"

[7]However, the Hilbert curve was found to be less suitable for cases where there is an intrinsic ordering of the dimensions, such as in time-series data. The reason for this problem is that the Hilbert curve is difficult to follow due to its complicated shape.

- "+" stands for a 90 degree rotation clockwise; "-" stands for a 90 degree rotation counter-clockwise and "F" stands for one step forward.

The L-System is used to generate the path of the Hilbert curve by first repeatedly replacing the "L" and "R" with the corresponding strings. Once the string has reached the desired length (measured by the number of "F" contained in the string), the "L" and "R" are removed and by following the rotation and stepping rules described above, the remaining sequence "+", "-" and "F" is converted into a series of coordinates that determine the path of the curve.

Mapping of the samples to the grid with a space-filling curve assumes that they are ordered according to some criterion. As discussed earlier, an ordering of the samples can be derived by mapping them to the leaves of a hierarchical structure such as an ontology, or by applying a hierarchical clustering algorithm. The space-filling curve then maps related samples to the same region of the grid, as illustrated in Figure 4.7.

The glyphs that are obtained using this approach present an effective solution to the graphical scaling problem that was identified in heat maps and profile plots, namely the growth along the horizontal axis with every additional sample. However, the perceptual scaling problem that is caused by the large number of samples in the expression profile persists. Eick and Karr (2002) propose the use of multi-resolution techniques to make visualizations more scalable, which follows the visual information seeking mantra of "overview first, zoom and filter, then details-on-demand" (Shneiderman, 1996).

These considerations led to the design of hierarchical glyphs that allow the investigator to view the expression profile at various levels of resolution, which is illustrated in Figure 4.8. For this purpose, expression levels are aggregated across samples that have the same parent node in the hierarchy, for instance, by computing their average expression level. Since the ordering of the samples in the expression profile is derived from the sample hierarchy, every node in it corresponds to a continuous, tightly packed region of grid cells in the glyph.

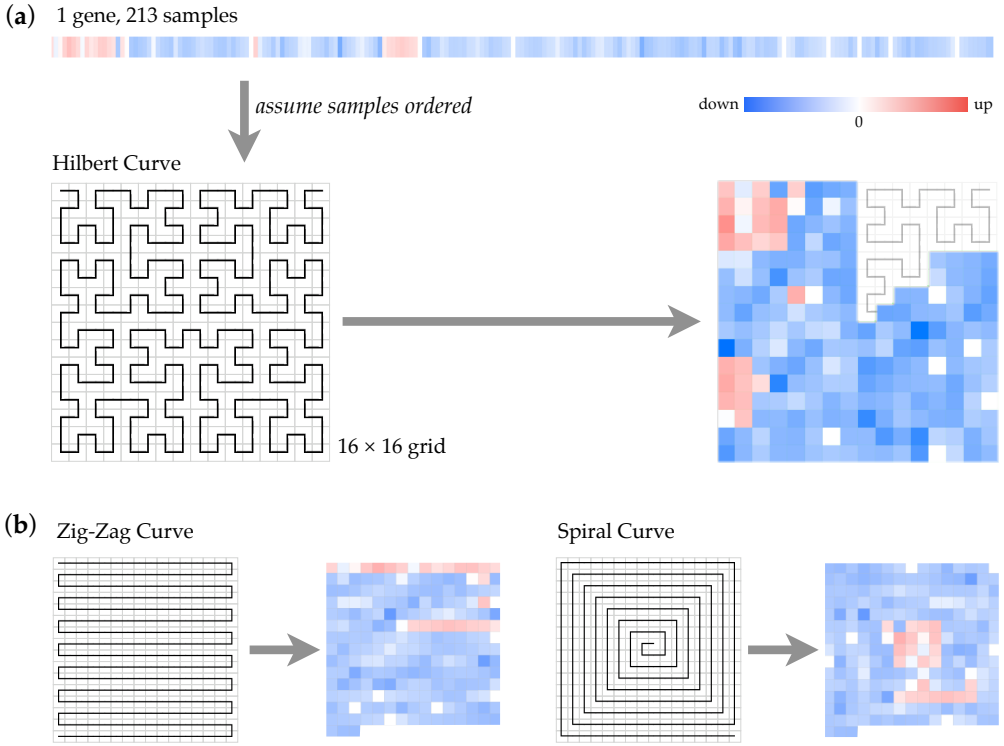

Figure 4.7: Mapping of an expression profile on a space-filling curve. (**a**) An expression profile in which an ordering has been defined on the samples is mapped to a square 16×16 grid following the path prescribed by a Hilbert curve. Since the expression profile contains only 213 samples the upper right corner of the grid remains unused. (**b**) Mapping of the same expression profiles using different types of space-filling curves.

The resolution of the glyph is reduced by coloring grid cells with the color corresponding to the aggregate value of a parent node in the sample hierarchy. This causes the grid cells associated with the corresponding node to be perceived as a single shape that corresponds to their union.

The identification of local patterns is more efficient with hierarchical glyphs because they provide a structure that supports the construction of a mental model that can be successively refined as necessary through interaction with the hierarchy. For instance, if a study measures gene expression levels in different tissues of the body, and among those the expression levels in various tissues of the nervous system, a first step would be to check the expression levels of a given gene in the overall nervous system. In further steps, the resolution can be increased to examine if there

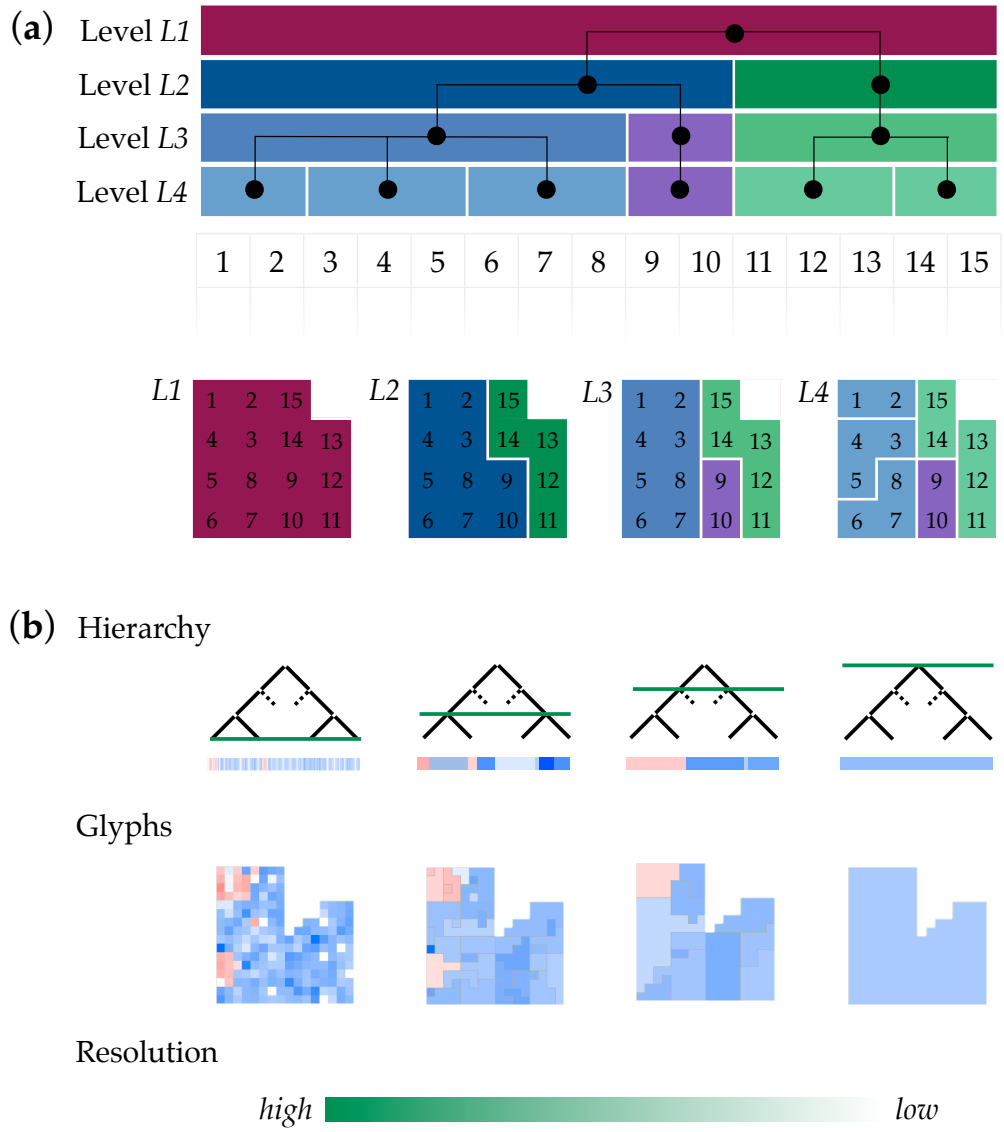

Figure 4.8: (**a**) Construction of the arrangement of the samples in the glyph. A hierarchy is associated with the samples of the expression matrix and used to derived an ordering. A Hilbert curve is used to map the 15 samples of the matrix to a 4×4 grid. One grid cell remains unused. The colors of the cells indicate the areas that are created through aggregation of samples by nodes in the hierarchy. (**b**) Illustration of glyphs with different levels of resolution created by applying a hierarchy to an expression profile.

are differences between, say, the central and the peripheral nervous system.

Experiments were undertaken with several different space-filling

curves in the context of this setup, some of which are shown in Figure 4.7 (**b**). It was observed that besides its good performance on the optimization of the mapping (Keim, 2000) the Hilbert curve has two properties that make it a good candidate for the purposes of the Space Maps method. First, it maps two samples that are adjacent in the 1-dimensional ordering onto the grid so that their grid cells share an edge. This ensures that all areas created by aggregation of samples are continuous. Second, when samples are aggregated, the shape of the area that is generated is very distinct due to the path of the Hilbert curve through the grid. It can be assumed that this helps investigators to build a mental model of the glyph arrangement because biological conditions or groups of samples can be associated with particular shapes. This is similar to recognizing the outlines of countries or other geographical regions on a familiar map.

4.3.3 Layout of Glyphs in 2-Dimensional Space

As discussed in Section 4.3.2, the glyphs representing the gene expression profiles are generated by arranging the samples in a 2-dimensional grid so that related samples are placed in close spacial proximity. In the next step, described below, the glyphs are arranged in 2-dimensional space so that related genes are placed close to each other. Due to the symmetry of gene and sample expression profiles in the expression matrix discussed in Section 4.2.2, the same approach can also be applied in a sample-centric way. In that case the genes are arranged on a grid to generate the glyphs and these are then be arranged in 2-dimensional space to represent relationships between samples. However, here only the gene-centric perspective is discussed.

To lay out the gene glyphs in 2-dimensional space, a grid-based approach can be used that works in the same way as the arrangement of samples in the glyphs. An ordering is defined on the genes, based on which they are mapped to the grid using a space-filling curve. If an ordering of the genes is derived from a *data-driven* hierarchical structure, such as a tree obtained from a hierarchical clustering of the expression profiles, this

is similar to Wattenberg's jigsaw layout (Wattenberg, 2005). If an ordering of the genes based on the similarity of their expression profiles is used, the grid layout provides a comprehensive, structured overview of the expression patterns in the expression matrix. Alternatively, the ordering of expression profiles can be based on some *knowledge-derived* classification of the genes, such as Gene Ontology classes or participation in a particular pathway, which allows the investigator to evaluate if the classes are reflected in the expression patterns.

Besides the grid-based layout, the Space Maps visualization supports a second approach based on dimensionality reduction techniques to arrange glyphs in 2-dimensional space, which is also used in the VaR display. Layouts based on dimensionality reduction share the same advantages as scatter plots that were discussed in Section 4.2.3.3 and are very efficient for global pattern discovery. This is due to the spatial encoding of relationships between the expression profiles. Since expression profiles are represented by the glyphs and are readily available, it is possible to form hypotheses about the global patterns through analysis of local patterns. For instance, if the investigator identifies a cluster of genes, the examination of the expression profiles can reveal what these genes have in common and how they are expressed across the samples.

A disadvantage of these layouts is that often glyphs overlap, which makes interpretation of expression profiles more difficult. In order to help investigators distinguish overlapping profiles, some basic remedies are supported by the Space Maps technique, as shown in Figure 4.9. Ward (2002) discusses a range of additional techniques for placing glyphs and reducing their overlap.

4.3.4 Integration of Gene Attributes

Some of the tasks that were identified in Table 4.1 require examination of both expression profiles and gene attributes (T9, T10). The Space Maps technique can integrate almost any kind of attribute directly into the visualization. Categorical and quantitative attributes can be encoded in essen-

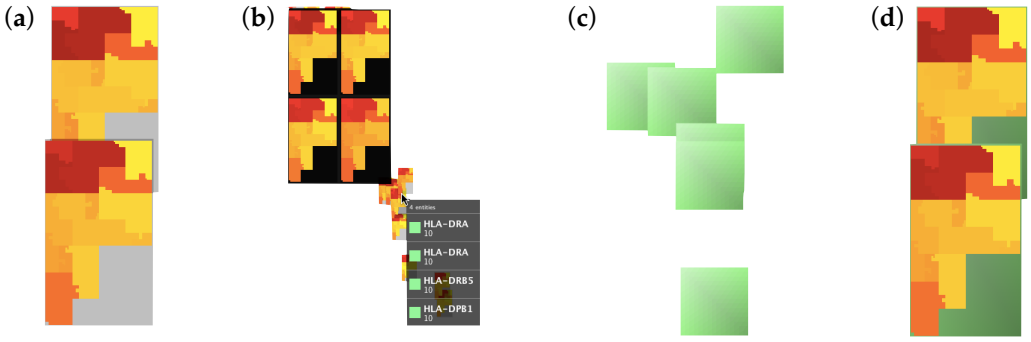

Figure 4.9: Strategies to distinguish individual glyphs and to combine color-coded attributes with expression profiles. (**a**) Glyphs are placed on semi-transparent tiles that are slightly larger than the glyphs themselves. This creates an outline that visually separates overlapping glyphs. (**b**) Moving the mouse over a set of overlapping glyphs shows all glyphs under the mouse pointer arranged on a grid in a tool tip in the top left corner. This figure also illustrates how labels are shown in tool tips in the bottom right corner. (**c**) Using smoothly shaded rectangles as marks for data points helps to distinguish them in dense clusters (Fekete and Plaisant, 2002). (**d**) A combination of (**a**) and (**c**) that shows both the expression profile and additional information by placing glyphs on colored tiles.

tially the same way as in scatter plots.

Quantitative attributes are encoded in the alpha channel or the size of glyphs or both. Categorical or class attributes are encoded by a color that is displayed as a frame around the expression profile, as shown in Figure 4.10 (**b**). The expression profile can also be removed temporarily to show only a colored glyph, as shown in Figures 4.10 (**b**), (**g**)-(**i**) and Figure 4.9 (**c**). With the grid-based layout described in Section 4.3.3 it is also possible to encode a gene attribute in an ordering of the glyphs, which can then be used to group them on the grid.

As illustrated in Figure 4.10, textual attributes can be shown as labels if they are reasonably short. Alternatively, they can be displayed on demand in a tool tip, as shown in Figure 4.9 (**b**).

4.3.5 Scaling of Color Maps for Expression Profiles

Dimensionality reduction methods yield different results depending on the measure used to determine the similarity of expression profiles. For instance, when a correlation measure such as Pearson's correlation coef-

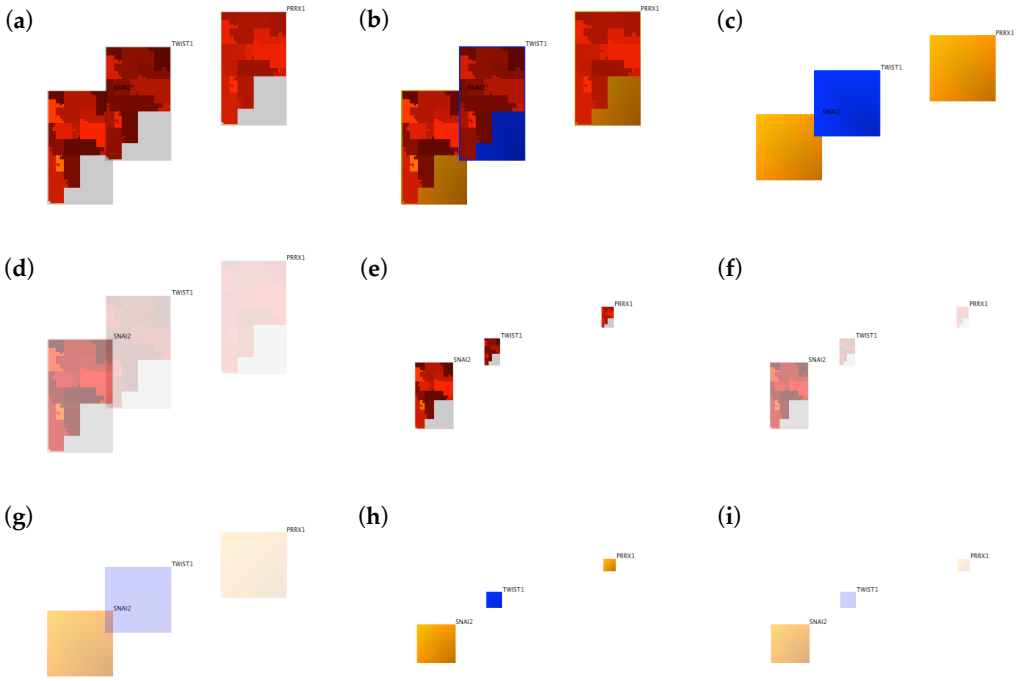

Figure 4.10: Integration of gene attributes into the visualization of glyphs. (**a**) Original glyphs for comparison. (**b**) Class attribute mapped to color. (**c**) Class attribute mapped to color. Expression profile hidden. (**d**) Quantitative attribute mapped to alpha channel. (**e**) Quantitative attribute mapped to glyph size using selective extent scaling. (**f**) Quantitative attribute mapped to both alpha channel and glyph size. (**g**) Quantitative attribute mapped to alpha channel of class glyph. (**h**) Quantitative attribute mapped to size of class glyph. (**i**) Quantitative attribute mapped on both alpha channel and size of class glyph.

ficient is used, genes whose expression levels have very different magnitudes may be found to be highly similar, whereas under another measure, such as Euclidean distance, they would be found to be very dissimilar.

Normally, Space Maps use a color mapping with global scaling to encode expression levels into colors used in the glyphs. In a global scaling the maximum and minimum values of the overall data set are used to determine the range that the color mapping covers. However, in cases where the expression profiles were projected based on a correlation measure that only takes into account the shape of the profiles but not their magnitude such a global color mapping is counterintuitive. This is because neighboring glyphs can look very different, as shown in Figure 4.11 (**a**), making it

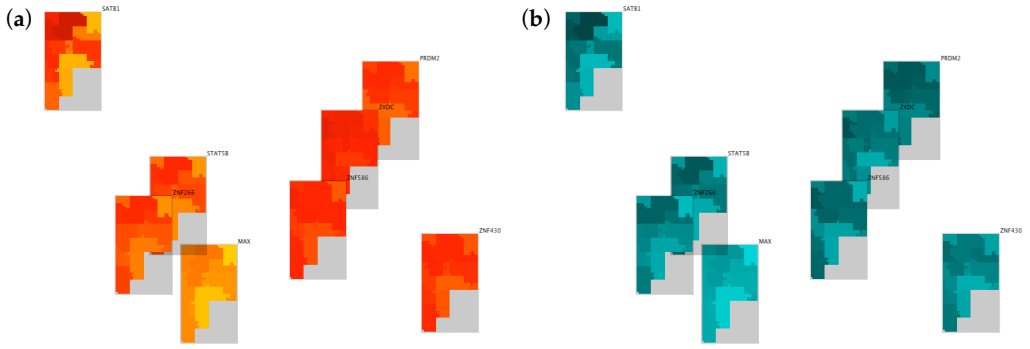

Figure 4.11: Color mappings using global and local color scaling for a projection based on Pearson's correlation coefficient. Different colors are used for local scaling to avoid confusion. (**a**) Global color scaling for positive expression levels. A single color mapping is used for all expression profiles. (**b**) Local color scaling. A separate color mapping is used for each expression profile.

difficult to determine the common pattern that they share.

A color mapping with a local scaling is more appropriate to visualize glyphs projected based on a correlation measure. In a local scaling the range of the color mapping is determined individually for each expression profile based on the maximum and minimum value of the profile. The local scaling emphasizes the shape of the profile and thus highlights the commonalities of the profiles, which is illustrated in Figure 4.11 (**b**).

4.3.6 Interaction and Navigation Techniques

Yang et al. (2007) present a comprehensive list of interaction techniques for the VaR display. The Space Maps visualization supports most of those and here only those are described that are different or new.

Zooming and extent scaling is supported as a means to focus on particular regions of the visualization. The two are independent, meaning that zooming into an area of the map will not change the size of the glyphs in the visualization. The size of the glyphs is controlled by extent scaling, which can be applied to scale the glyphs so that they use as much screen space as possible without overlapping.

The need to be able to switch quickly between global and local pat-

tern exploration is an important result of the task analysis. For instance, if the investigator identifies a cluster of genes somewhere in the expression space while looking at a dimensionality reduction-based projection of the expression profiles, a natural next step is to take a closer look at the expression profiles before continuing the exploration of the projection. It was observed that it is too tedious to adjust zoom and extent scaling factors manually every time the investigator wants to get a close up view of the expression profiles in a region of the visualization. To support this interaction pattern a "spring loaded zoom" feature was developed, which is illustrated in Figure 4.12: The investigator moves the mouse pointer over the area of interest and holds down a button on the keyboard or mouse. While the button is held down the region of interest will automatically be centered on the screen and then zoomed into with extent scaling coupled to the zoom factor. Once the region of interest is close enough, the investigator releases the button. The visualization will maintain the zoom and extent scaling factor while the glyphs are examined. By pressing the button again the visualization will return to the original state. Since panning, zooming and extent scaling are smoothly animated, a sense of orientation can be maintained.

The names of genes in the visualization can be retrieved by moving the mouse over the corresponding glyph. A tool tip shows the gene name or another textual attribute associated with the gene as selected by the investigator. If several glyphs overlap, a list of gene names will be shown, as illustrated in Figure 4.9 (**b**). The Space Maps visualization also supports adaptive labeling of glyphs. Rather than trying to label all glyphs, only a subset of them are labelled based on a quantitative gene attribute and an upper limit for the number of labels that can be controlled by the investigator. For instance, if the investigator is interested in genes that have a low p-value for differential expression, those genes are labeled preferentially. The limit set by the investigator is a limit with respect to the visible glyphs. If labeled glyphs move outside the visible area, their labels become available for glyphs that so far had been unlabeled. Adaptive labeling, in combination with the spring loaded zoom feature, enables the investigator

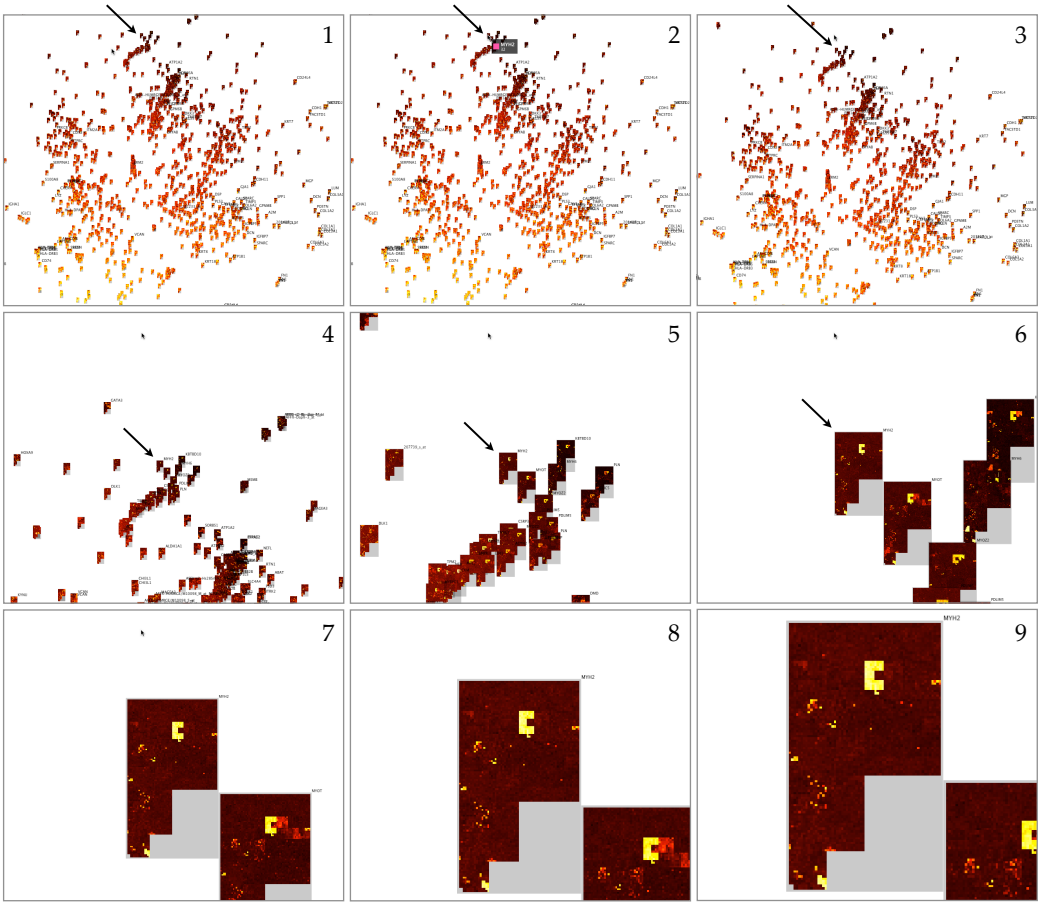

Figure 4.12: Illustration of the spring loaded zoom interaction. In Step 1 the investigator notices an interesting expression profile that is part of a bigger cluster. The investigator then moves the mouse into the area of interest in Step 2, in this case directly on the glyph, which brings up a tool tip with details about the gene. The investigator holds down the spring loaded zoom key in Steps 3 through 9 for animated zoom into the area of interest. Once the expression profile has been enlarged sufficiently, the investigator releases the key and examines the glyph. When the spring loaded zoom key is pressed again, the original state from Step 2 will be restored by reversing the animation.

look up the names of a set of genes very efficiently.

Moving the mouse over a set of overlapping glyphs not only shows their names, but also can show the glyphs arranged on a grid without overlap, as illustrated in Figure 4.9 (**b**). This enables the investigator to examine very dense areas in layouts that are based on dimensionality reduction without having to zoom in and scale the glyph extent.

Subsets of genes can be created through application of various fil-

ters on gene attributes as well as through selection of glyphs directly in the visualization. Saraiya et al. (2004) found in their evaluation of microarray visualization tools that this is one of the most important features required to gain insight from the data.

4.4 Prototype Implementation

A prototype of a visualization tool built around the Space Maps visualization was implemented. Its user interface is shown in Figure 4.13. The software is written in Java Version 1.6 (www.java.com) and uses the Java Bindings for OpenGL (JOGL, www.jogamp.org) library for rendering in OpenGL (www.opengl.org).

Data is loaded into the tool by parsing an XML-based *project file* that refers to additional files that contain expression matrices, hierarchies and attribute data. A sample project file is shown in Listing C.1 in Appendix C. Using an XML-based format to refer to tab-delimited text files is a very flexible solution to enable the import of results from a wide-range of external tools. This allows the investigator to work with the prototype in a more realistic data analysis setting.

As shown in Figure 4.13, the user interface consists of a central panel that shows the Space Maps visualization of the data set. Panels to the left and right contains controls that are used to set how the glyphs are generated, i.e. which expression matrix, hierarchy and kind of space-filling curve to use (marked with "Glyphs" in Figure 4.13), if and which class attributes are shown ("Categorial Attributes"), how quantitative attributes are mapped to the glyphs and what base size and alpha value to use ("Quantitative Attributes"), how the glyphs are laid out in 2D, e.g. by choosing a space-filling curve and an ordering or by choosing a precomputed projection ("2D Layout") and which labels to use ("Labels") and how many of them should be shown ("Label Count"). Furthermore, the investigator can control the brightness of the background in the central display ("Display Color") and if all glyphs under the cursor should

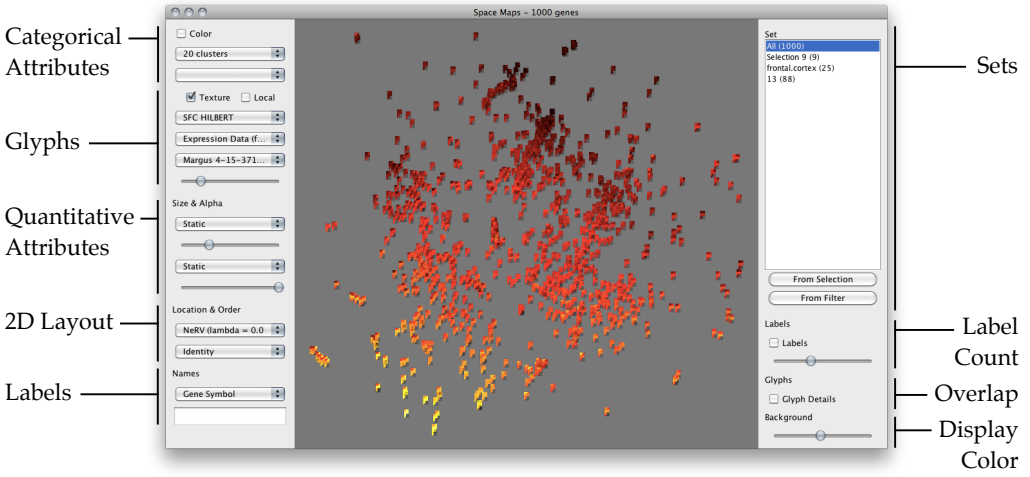

Figure 4.13: User interface of the Space Maps prototype implementation. The main display in the center shows the Space Maps visualization. The panels on the left and right contain controls for the mapping of attributes, application of filters and creation of sets. Further details are described in the text.

be shown without overlap in a pop up or not ("Overlap"). Additionally, filters can be applied by either selecting a specific class from a categorical attribute, by matching a regular expression against the currently selected labels or by using the mouse to select a region in the visualization. A set can be created from the items passing a filter ("Sets"). Selecting a set at later stage during the analysis has the same effect as reapplying the filter used to create the set.

It has to be emphasized that the focus for the development of the prototype was to implement a tool that enables the testing and evaluation of the Space Maps visualization method. Neither the arrangement of nor the interactions with the user interface controls of the prototype were optimized since the visualization is intended to be used as a component of gene expression data analysis software applications.

Like in the VaR display implementation by Yang et al. (2007), the glyphs are generated once and then stored as textures that are mapped to quadrilaterals as needed. Generation of glyphs is the most time-consuming process in the prototype implementation, but due to the small size of the glyphs and the fixed arrangement of samples in the glyphs, it

is possible to generate them once and then load them into memory when they are requested.

With 1,000 probe sets and over 5,000 samples, one of the largest test data set used to evaluate the visualization method contains more than 5,000,000 data points. With the prototype implementation, it is possible to render the corresponding visualization at a frame rate of 30 frames per second on a MacBook Pro with a 2.53 GHz processor, 4 GB of memory and a NVIDIA GeForce 9400M graphics processor. The prototype uses the OpenGL Vertex Buffer Object (VBO) extension to upload vertex and texture data directly into high-performance memory on the graphics card. This reduces the load on the CPU significantly and achieves much better frame rates. The high frame rate is required for the implementation and use of animated transitions and smooth zooming.

4.5 Case Studies

The Space Maps visualization technique for gene expression data sets with hundreds of samples was designed based on a list of tasks that have to be performed in most transcriptomics data analyses. To test how well the technique performs on these tasks, it was applied to a data set that contains measurements of around 20,000 human genes in 5,372 samples (Lukk et al., 2010). The data set is a "meta data set" in the sense that it consists of experimental data from many individual studies, which cover a wide range of conditions such as diseases, tissues and cell lines. Only data from studies with high quality measurements on Affymetrix Human Genome U133 microarrays were selected by Lukk et al. from the ArrayExpress Archive and the Gene Expression Omnibus repository. All data was normalized to remove systematic variation.

A hierarchy for the samples in this data set was manually created by Lukk et al. (2010). The hierarchy has five levels with one node at the root (L1), four nodes on the second level (L2), 15 nodes on the third level (L3), 371 nodes on the fourth level (L4) and 5,372 leaves (L5). As test data the

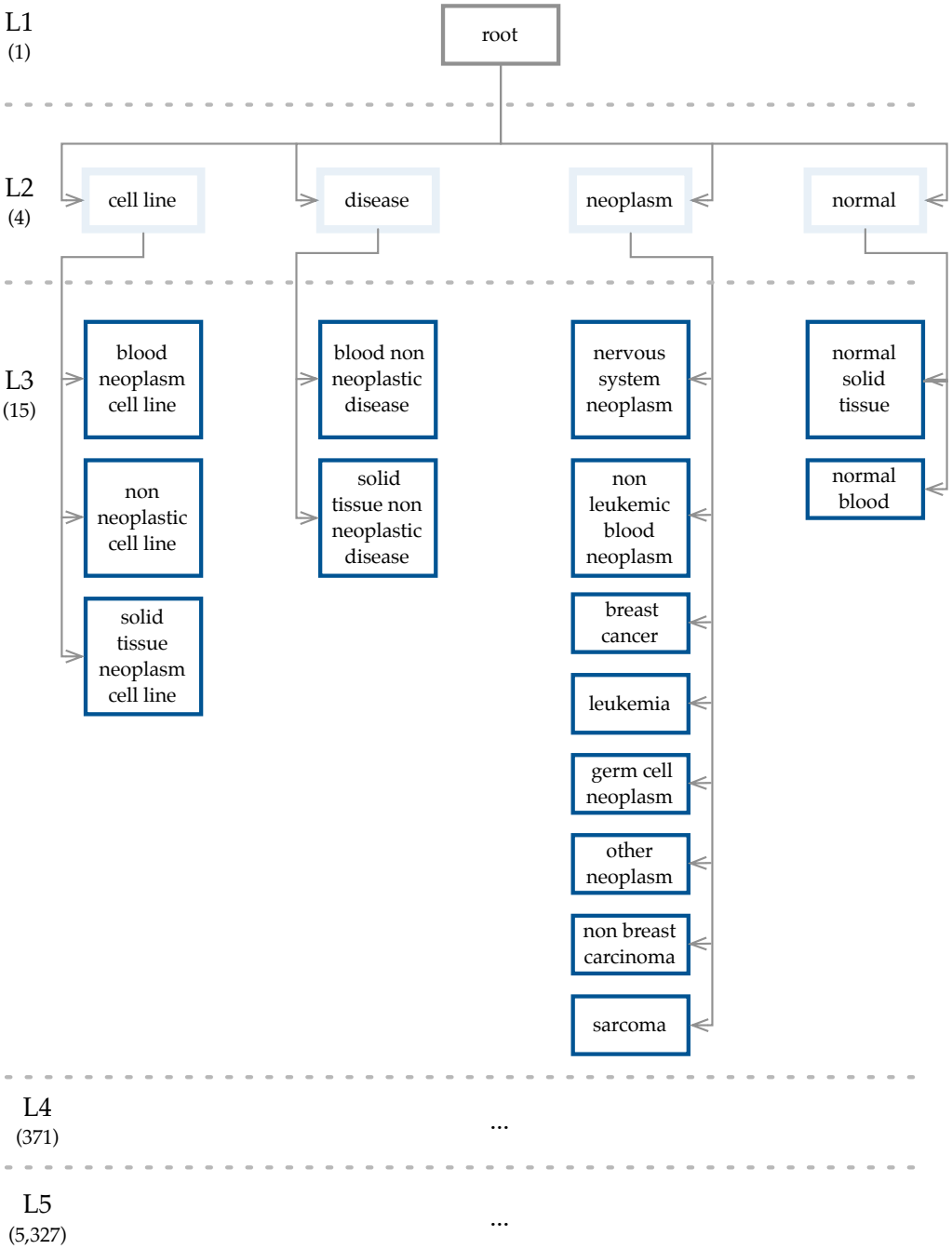

Figure 4.14: Hierarchy for the meta data set used in the case studies showing the categories of levels L1 through L3. The numbers in parentheses represent the number of categories on each level. Colors correspond to colors in Figure 4.15.

1,000 most variable probe sets across all samples were selected from the overall data set. An overview of the hierarchy is provided in Figure 4.14 and the structure of the corresponding glyphs for the data set is shown in Figure 4.15.

4.5.1 Analysis of Individual Expression Profiles

Figure 4.16 shows, at five levels of resolutions, the expression profile of Affymetrix probe set 202508_s_at, which maps to the human SNAP-25 gene. By examination of the expression profile it can be concluded that SNAP-25 is always highly expressed in the brain and that its expression levels elsewhere are comparatively low, indicating that this gene might play a role in processes within the brain. Since SNAP-25 encodes a synaptosome-associated protein (SNAP) of 25 kDa, this observation is indeed not surprising and the literature confirms that SNAP-25 is essential for exocytosis in neurons and neuroendocrine cells (Hodel, 1998). By comparing the L5 with the L4 glyph, it can be concluded that most of the variance in the expression profile is explained by change in conditions and that the within condition variance is rather low.

Analysis of the SNAP-25 expression profile shows that the Space Maps visualization is suitable for the identification and interpretation of local patterns (T1 - T3).

4.5.2 Finding Related Expression Profiles

Figure 4.17 (**a**) shows a NeRV (Neighborhood Retrieval Visualizer, Venna and Kaski, 2007a) projection based on the Euclidean distances of the 1,000 probe set expression profiles of the test data set. In this projection the two probes for the SNAP-25 gene have been highlighted as a result of a search for "SNAP25" across the labels of the probe sets. Glyphs corresponding to other probe sets are faded out. It can be seen that the two SNAP-25 probe sets are at the far ends of a dense cluster in the projection. Visual inspection reveals that the glyphs in this cluster exhibit a pattern that is

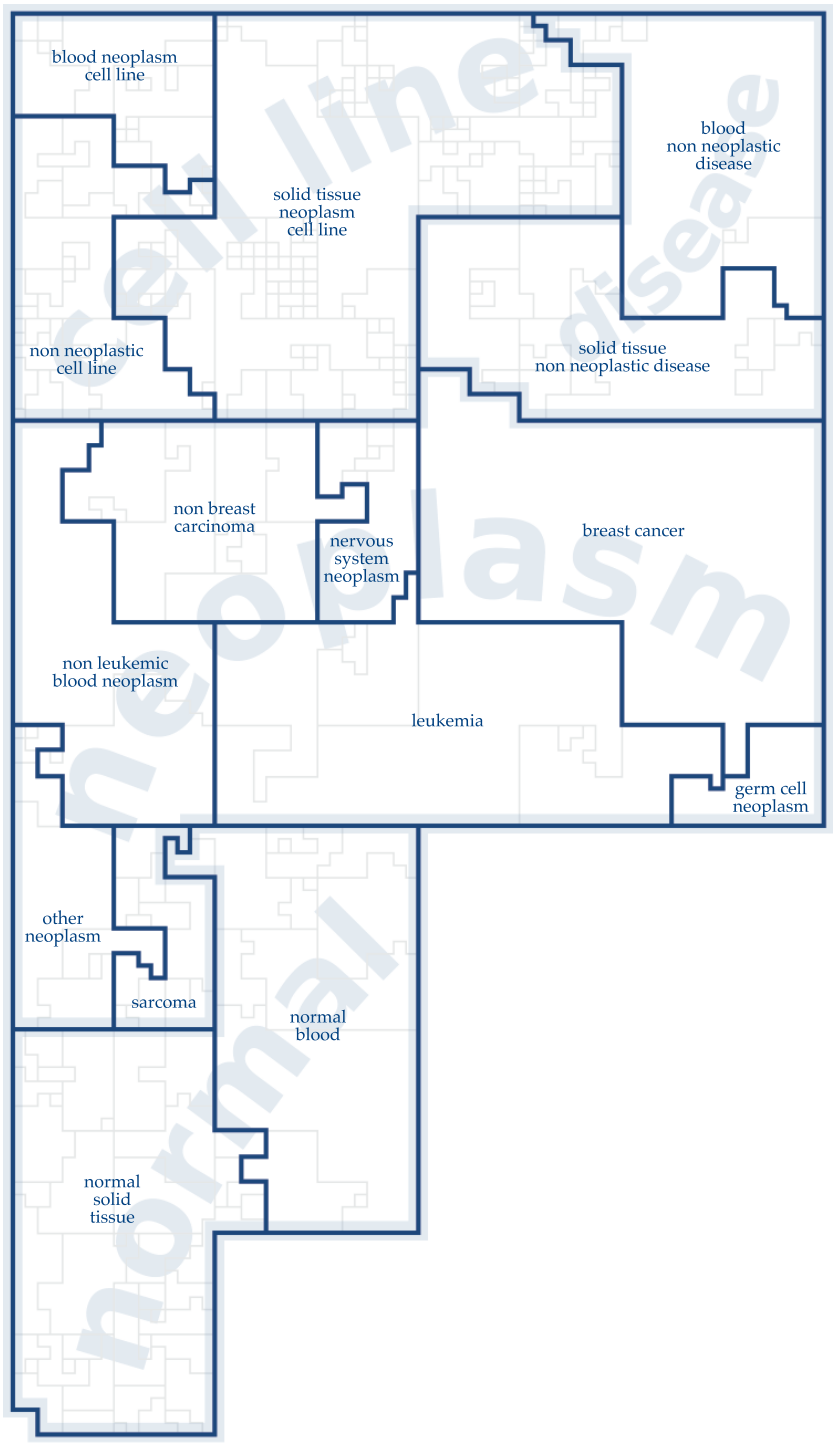

Figure 4.15: Structure of glyphs in the meta data set based on the hierarchy shown Figure 4.14. Outlines of the areas corresponding to categories on levels L2 (light blue), L3 (dark blue) and L4 (gray; parts only) are shown. Labels are provided for level L2 and L3.

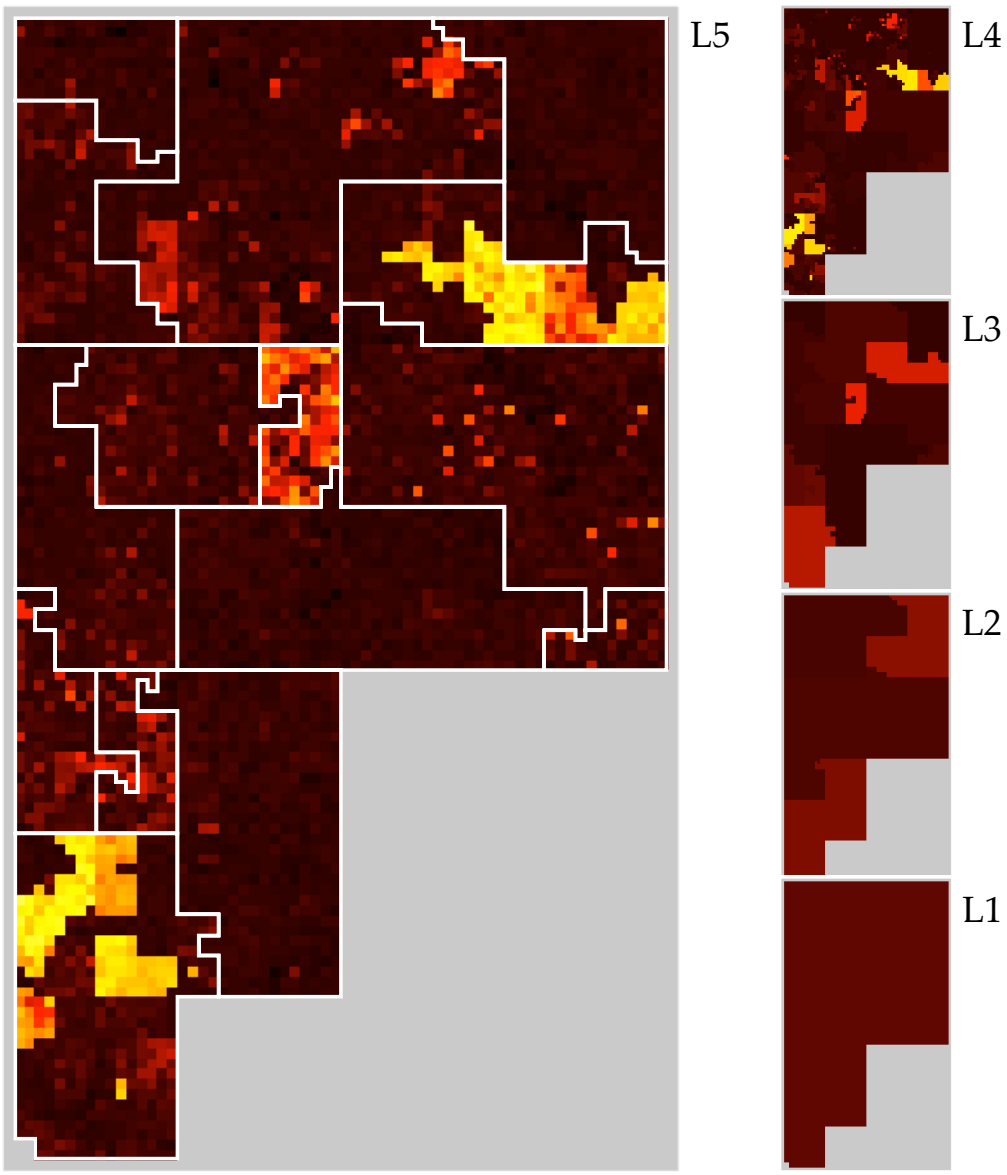

Figure 4.16: Expression profile of the human SNAP-25 gene across 5,372 samples, shown at levels L1 through L5 of the sample hierarchy. The L5 glyph is overlaid with the outlines of L3. In this data set all expression values are positive and the color scale ranges from black via red and yellow to white. Black represents the minimum and white the maximum expression level in the data set. The bright yellow areas in the L4 and L5 glyphs correspond to different parts of the brain and brain diseases. The bright red area in the center of the L3, L4 and L5 glyphs corresponds to neoplasms of the nervous system.

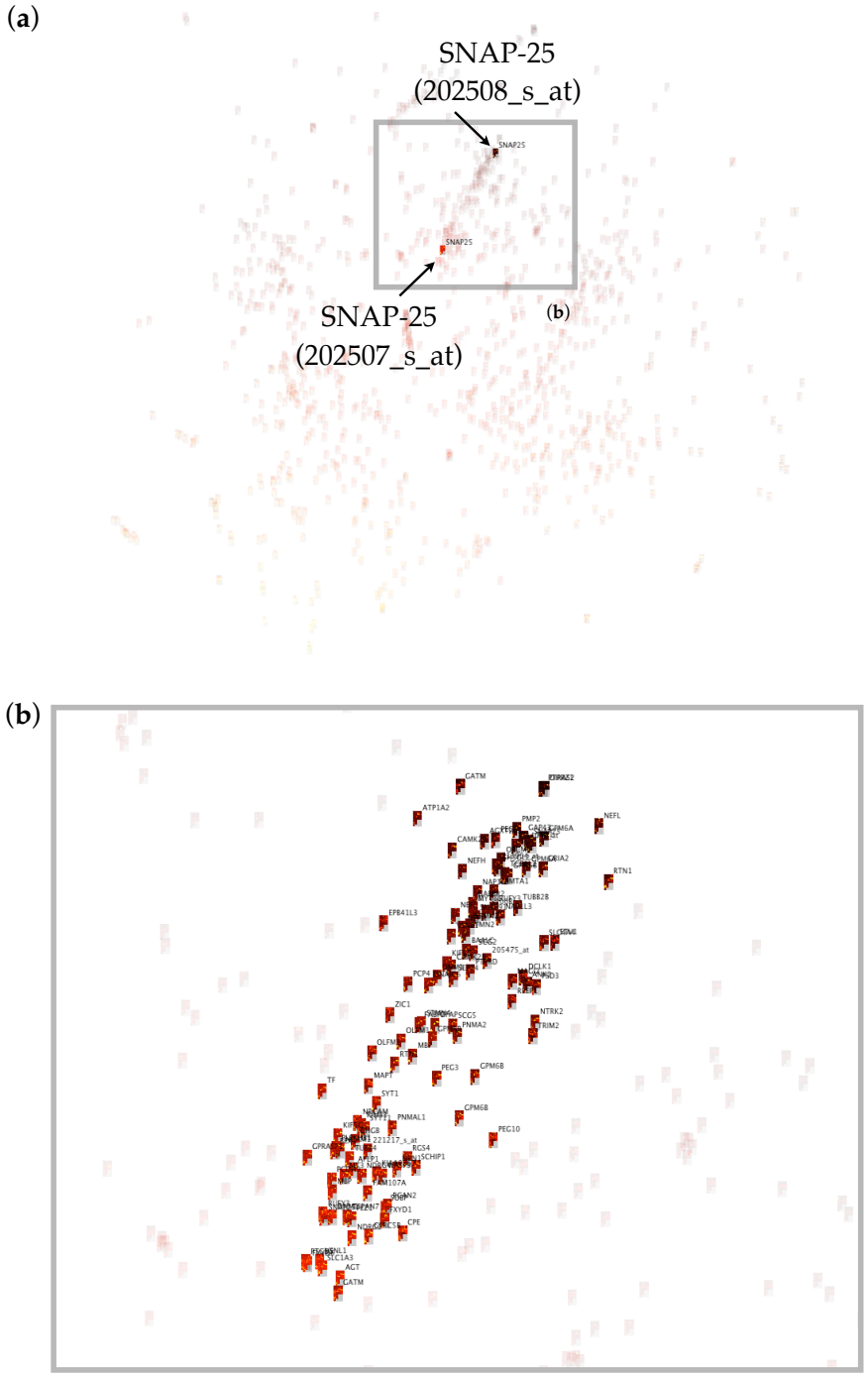

Figure 4.17: Cluster of expression profiles with high similarity to SNAP-25. (**a**) NeRV projection of 1,000 probe sets with glyphs for SNAP-25 probe sets highlighted. (**b**) Detailed view of the projections from (**a**) showing a cluster of expression profiles with high similarity to the SNAP-25 expression profile.

very similar to the expression pattern of SNAP-25 shown in Figure 4.16. A set can be created to represent this cluster by selecting the probe sets of the cluster using the mouse selection tool of the prototype implementation. The set is highlighted in Figure 4.17 (**b**), which is an enlargement of the NeRV projection shown in Figure 4.17 (**a**). Among others, the cluster contains probe sets for genes such as SYT1 (encoding synaptotagmin I protein), RTM1 (encoding reticulon-1 protein) and CAMK2B (encoding calcium/calmodulin-dependent protein kinase type II beta), which are all known to have functions in the brain, according to their corresponding WikiGenes entries (www.wikigenes.org, Hoffmann, 2008). The presence of such genes is a good indicator that the cluster indeed represents genes that are either specifically expressed in the brain or at least play a role in processes associated with the brain or nervous system.

To further study the expression profiles of the probe sets in the cluster, the glyphs are laid out on a 2-dimensional grid with the help of a Hilbert curve, as shown in in the top right corner of Figure 4.18. The ordering of the profiles used in the mapping is derived from a hierarchical clustering based on the Euclidean distances of the expression profiles. The grid layout is shown for the whole data set of 1,000 probe sets, but the glyphs of probe sets that are not part of the cluster have been faded out to highlight the glyphs of interest.

The glyphs from the brain-specific cluster are primarily located in two different parts of the grid in Figure 4.18, while a few glyphs are scattered across the map. The split of the cluster into two separate regions can be explained by the difference in overall expression levels in the two regions. The group on the left has a lower overall expression level, as indicated by the darker red, while the group on the right has a higher overall expression level, as indicated by the brighter red. This can also be observed in the projection in Figure 4.17 (**b**), where there is a gradient from darker red to brighter red from the upper right corner to the lower left corner. To confirm this observation, local color scaling was applied to the glyphs. Enlarged views of the two groups are shown in Figure 4.18 at

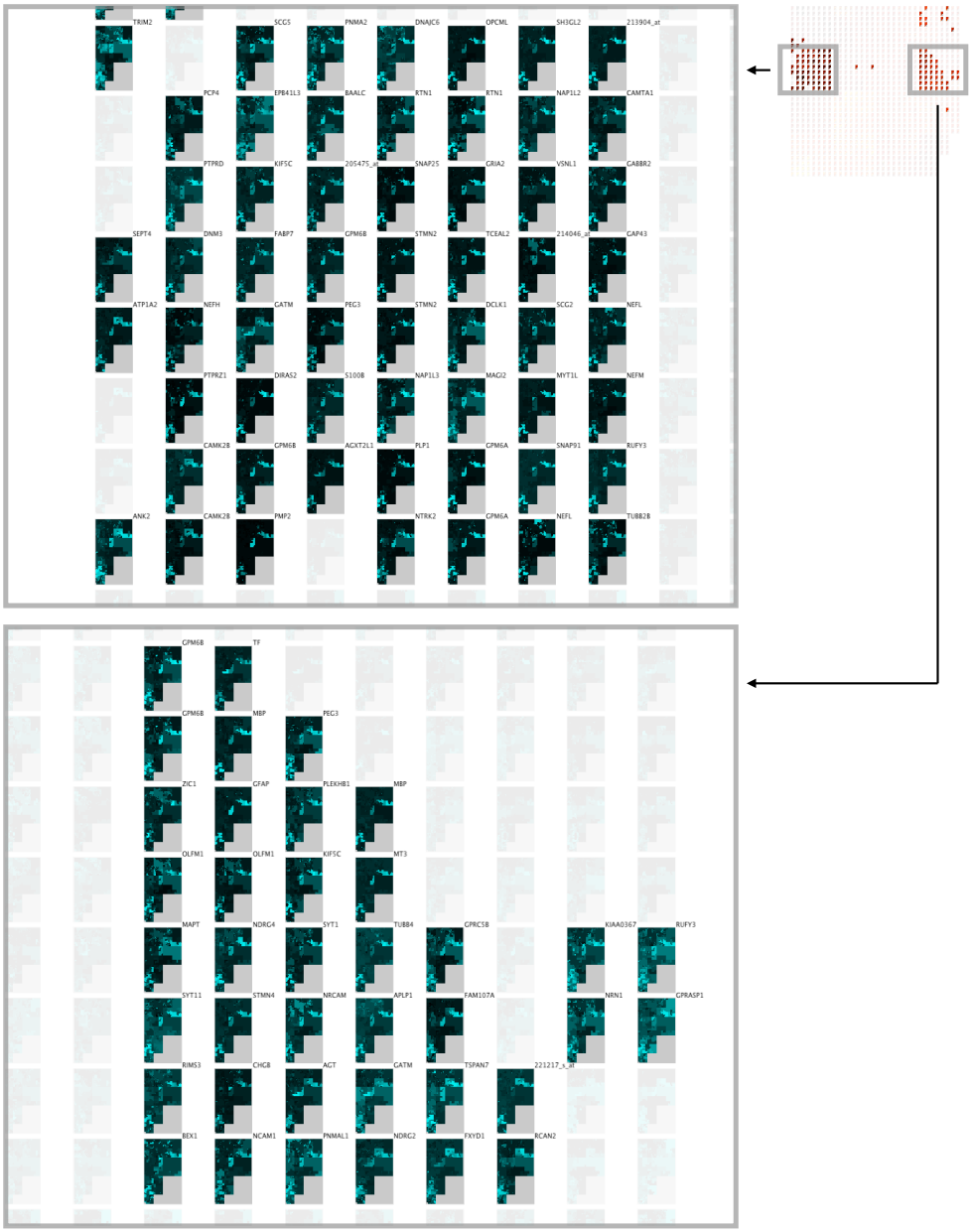

Figure 4.18: Detailed view of a cluster of expression profiles with high similarity to SNAP-25. A discussion of the figure is included in the main text.

the top (left group) and bottom (right group). The local color scaling emphasizes the regions where the probe sets have their highest expression levels, which is indicated by brighter cyan, confirming that a primarily brain-specific expression pattern can be observed in both groups. How-

ever, closer inspection also reveals that the group of genes shown at the bottom of Figure 4.18 are also expressed under other conditions besides the brain, whereas the group shown at the top appears to be more brain-specific.

The identification of a cluster of brain-associated probe sets based on the similarity to the SNAP-25 expression profile illustrates how the Space Maps technique can be applied to find profiles similar to a given profile, here SNAP-25 (T4).

4.5.3 Identification and Interpretation of Clusters

An overview of data set of 1,000 probe sets is provided in Figure 4.19, which shows the same NeRV projection as Figure 4.17 (**a**) with glyphs at level L3 of the sample hierarchy. It is evident that there is some structure in the data, including some clusters and outliers. Genes that are overall highly expressed (light yellow) are in the lower left corner of the projection and genes that are overall lowly expressed (dark red, black) are in the top right corner. Several observations were made by inspection of the visualization, two of which are marked in Figure 4.19 and discussed below.

Observation I is a dense cluster of probe sets that appear to be somewhat separated from the other probe sets, which is a hint that these probes might have a specific expression profile in common. An enlarged view of the region is shown in Figure 4.20 (**a**) and a grid layout of the cluster is shown in Figure 4.20 (**b**).

The cluster identified in Observation I contains expression profiles that indicate that the corresponding genes are highly specific for muscle tissues, as indicated by the labels (*H, HD, SM, SMD*) in Figure 4.20 (**b**). Examination of the gene symbol labels provides further support for this hypothesis, since many of them begin with "MY..." or "ACT...", hinting at involvement with myosin or actin, respectively. The gene symbols of the 23 probe sets in the cluster shown in Figure 4.20 (**b**) map to 20 unique genes: LDB3, MYL2, ACTN2, NEB, TNNC1, MYOT, ACTA1, MYH2,

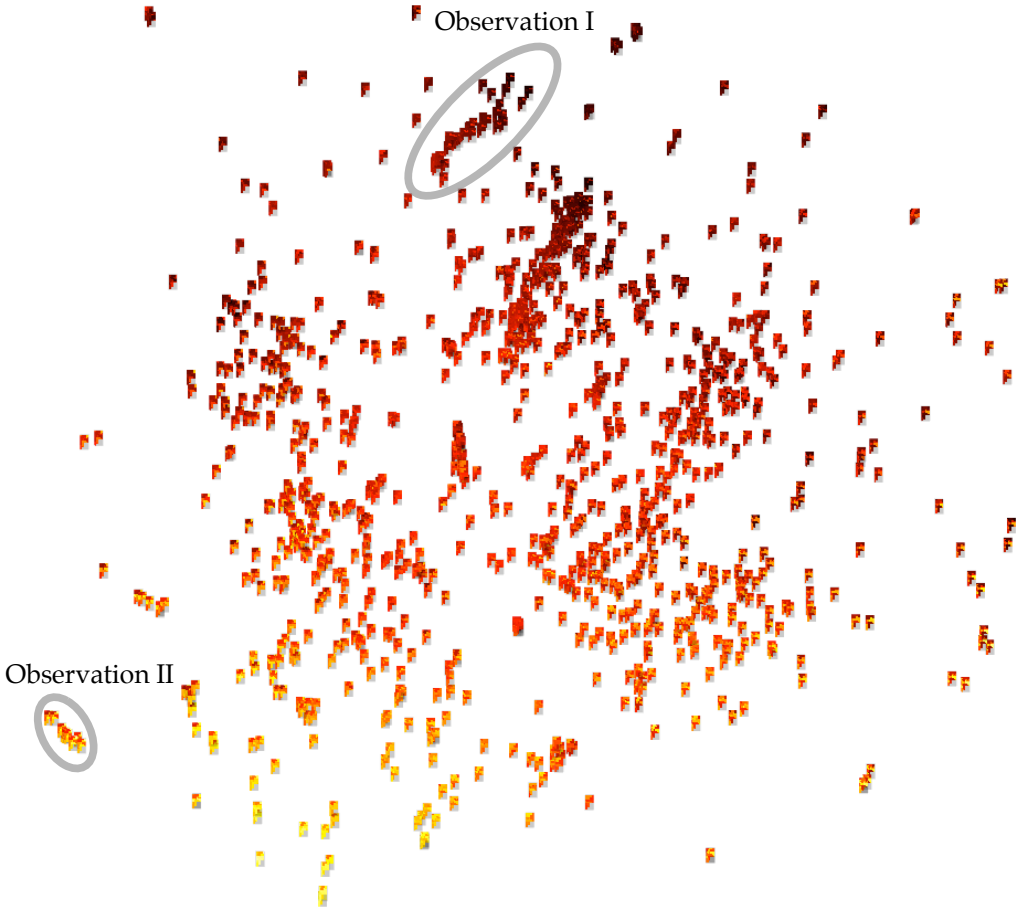

Figure 4.19: Layout of 1,000 expression profiles based on dimensionality reduction with the Neighborhood Retrieval Visualizer (NeRV) (Venna and Kaski, 2007a). Glyphs show level L3 of the sample hierarchy. A discussion of Observations I and II is included in the main text.

MYBPC1, PDLIM5, SMPX, PLN, CKM, TTN, CLIC5, MYH6, CSRP3, MYOZ2, KBTBD10 and PDE4DIP. Their Gene Ontology annotations are shown in Tables D.1, D.2 and D.3 in Appendix D. Analysis of the gene list with the "Functional Annotation Clustering" tool of the Database for Annotation, Visualization and Integrated Discovery (DAVID; Version 6.7; david.abcc.ncifcrf.gov; Huang et al., 2009) revealed that the list is highly enriched for annotation terms such as, for example, myofibril, contractile fiber, sarcomere, actin cytoskeleton (GO Cellular Component terms), structural component of muscle, actin binding, cytoskeletal protein bind-

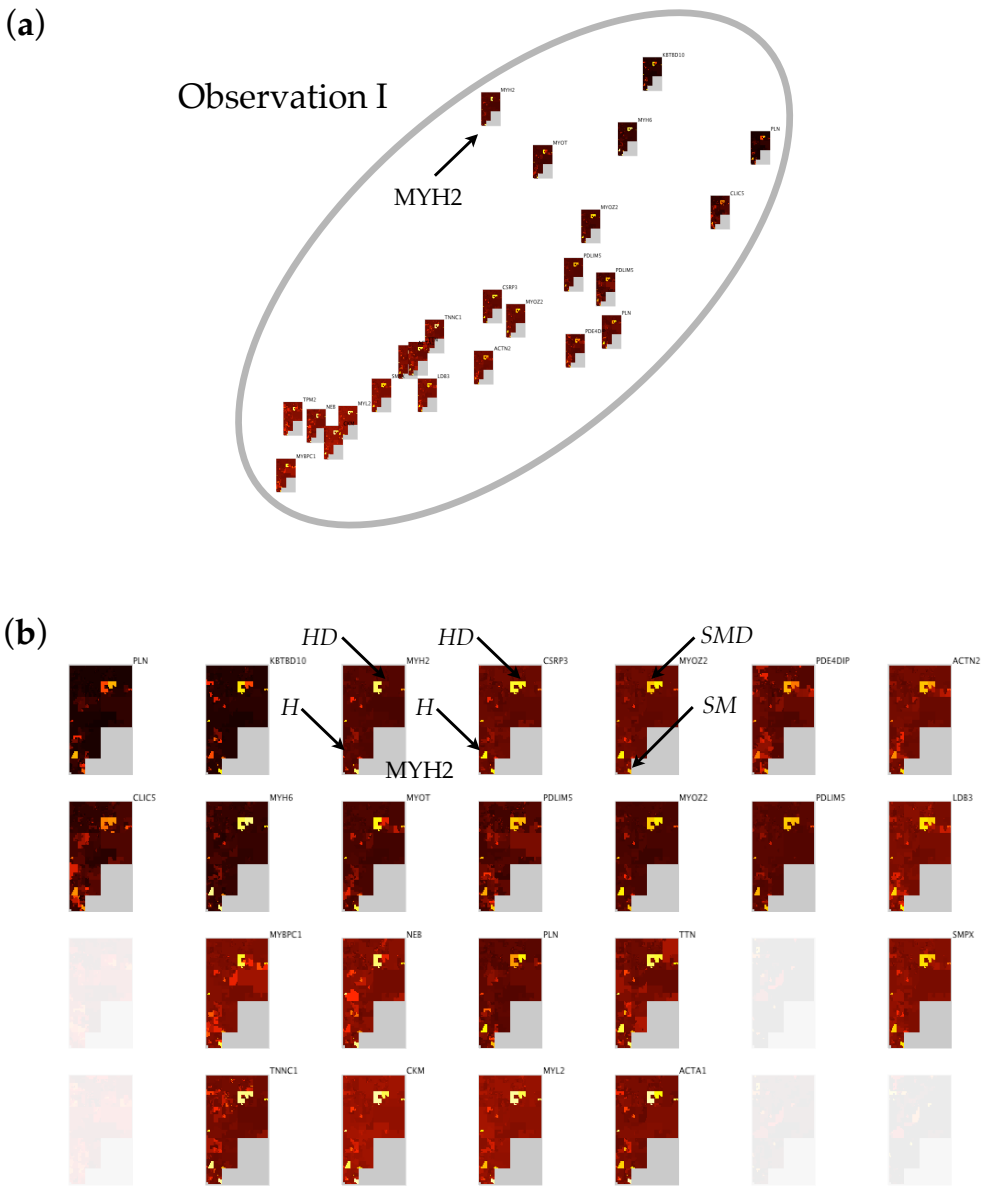

Figure 4.20: Detailed views of Observation I. A discussion is included in the main text. (**a**) Enlargement of the region from Figure 4.19. (**b**) Grid layout using a Hilbert curve and an ordering from a hierarchical clustering based on Euclidean distances. Labels are *H* = heart, *HD* = heart disease, *SM* = skeletal muscle and *SMD* = skeletal muscle disease. One outlier expression profile is not shown in this view.

ing (GO Molecular Function terms), as well as muscle contraction, muscle system process and muscle organ development (GO Biological Process terms). The method was run with the default parameters. The results of

the DAVID analysis confirm the initial hypothesis that the genes in the cluster are important for processes in muscle tissue.

One interesting outlier is the probe set for the MYH2 gene, which encodes the myosin-2 protein. Unlike all other genes in the cluster, this gene is not expressed in the heart or heart disease. One possible hypothesis is that the kind of myosin encoded by this gene is not used in heart tissue, since the transcript is absent both in normal and diseased heart tissue.

4.5.4 Detection of Outliers and Anomalies

Observation II marked in Figure 4.19 is an example of how the Space Maps technique can be used to identify anomalies and outliers (Task T7, see Table 4.1). By examining the group of outliers in the projection, it can be found, for instance, that the nine probe sets are only representing two different genes, HBA1 (hemoglobin alpha) and HBB (hemoglobin beta), which are measured by six and three probes, respectively, on the Affymetrix Human Genome U133 microarrays used to generate the data. The expression profiles are very similar and indicate that HBA1 and HBB are highly expressed in whole blood, both healthy and diseased; various blood cell types; bone; leukemia samples and some other tumors. However, neither HBA1 nor HBB appear to be expressed in cell lines, with the exception of the K562 myelogenous leukemia cell line, where HBA1 is highly expressed.

Another example of the identification of outliers is Observation III marked in Figure 4.21. Figure 4.21 shows a grid layout of the same 1,000 glyphs that was generated with a Hilbert curve and an ordering derived from a hierarchical clustering based on the Euclidean distance of the expression profiles. While it is possible to identify clusters of glyphs with distinct expression profiles, it is generally more challenging to identify outliers in this layout due to the lack of spatial cues. However, the better use of space makes it possible to render larger glyphs, which helps to spot interesting local patterns and in this case led to Observation III.

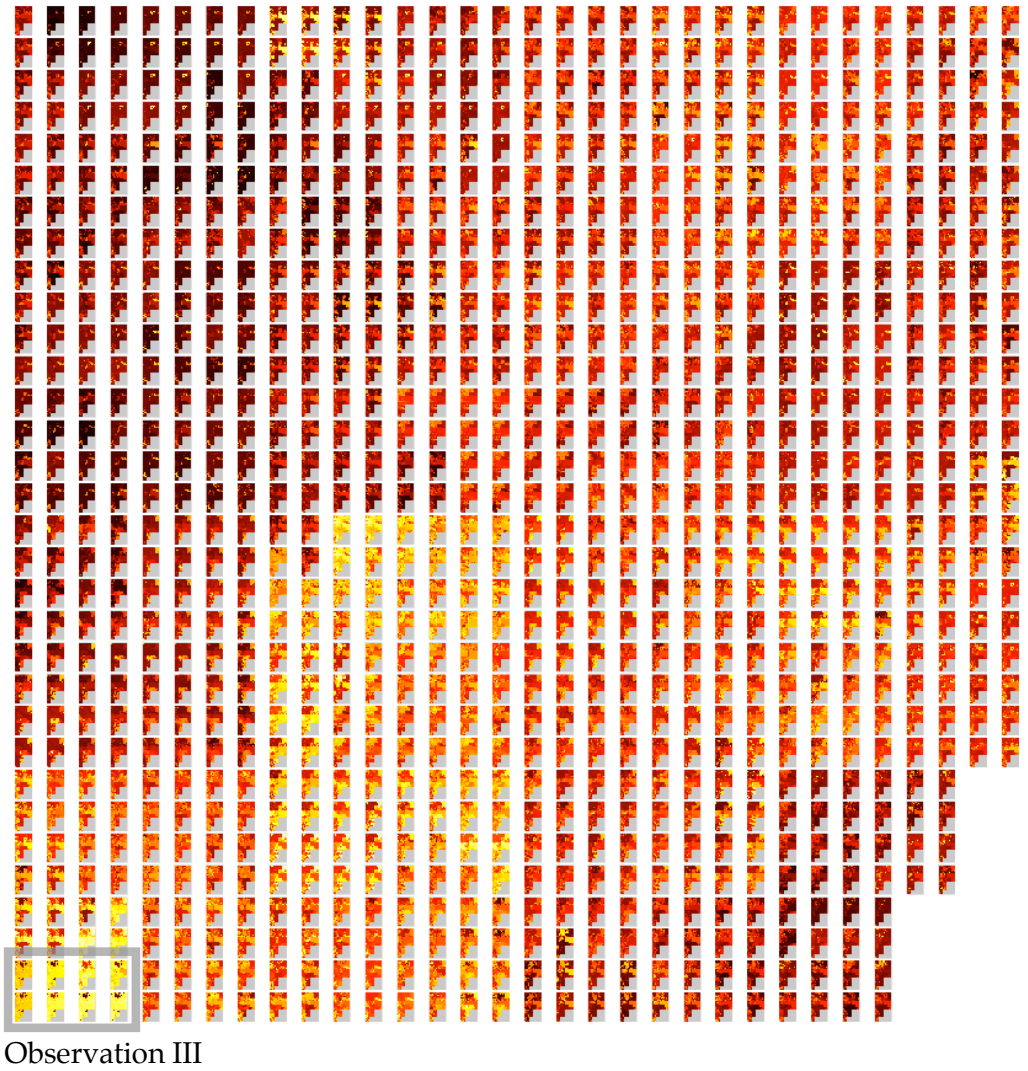

Figure 4.21: Layout of 1,000 expression profiles based on a Hilbert curve and an ordering derived from a hierarchical clustering based on the Euclidean distance of the expression profiles. Glyphs show level L4 of the sample hierarchy. Groups of glyphs with similar expression profiles can be identified despite the low resolution and the large number of expression profiles. A discussion of Observation III is included in the main text.

The eight expression profiles marked as Observation III stand out because of their very high overall expression indicated by the bright yellow and a small number of samples in which the corresponding transcripts appear not be expressed at all. A detailed view of the eight glyphs is shown in Figure 4.22.

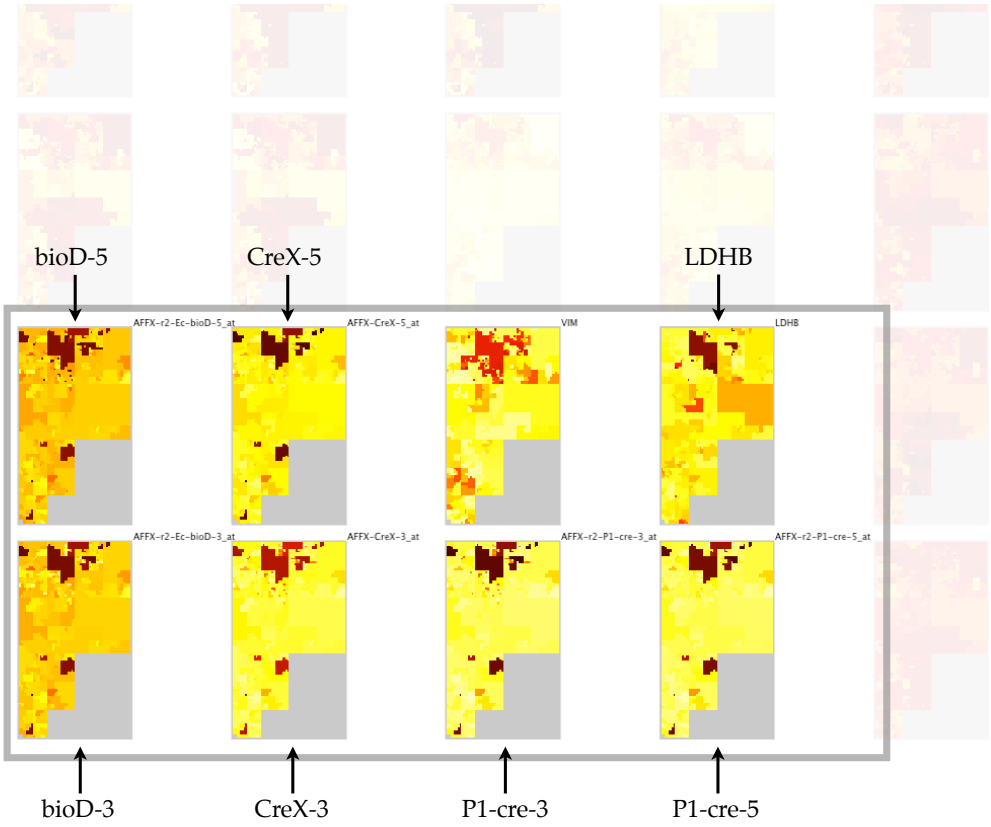

Figure 4.22: Detailed view of Observation III highlighted in Figure 4.21. Affymetrix control probe names are abbreviated, the full names are described in the main text.

Six of the eight probe sets that are part of Observation III represent control probes that are included on the Affymetrix arrays to monitor the hybridization quality. The control probes are designed to target non-human sequences that are spiked into the hybridization cocktail at different concentrations (www.affymetrix.com/support/help/faqs/ge_assays/faq_19.jsp). The control probes in Observation III are AFFX-r2-Ec-bioD-5_at (abbreviated as "bioD-5" in Figure 4.22), AFFX-r2-Ec-bioD-3_at ("bioD-3"), AFFX-CreX-5_at ("CreX-5"), AFFX-CreX-3_at ("CreX-3"), AFFX-r2-P1-cre-5_at ("P1-cre-5") and AFFX-r2-P1-cre-3_at ("P1-cre-3"). Of the sequences targeted by these control probes, the ones with the lowest concentration are targeted by the "bioD-*" probe sets and the ones with the highest concentration are targeted by the "P1-cre-*" and "CreX-*" probe

sets. In Figure 4.22 this is reflected in the slightly more orange yellow of the corresponding "bioD-*" glyphs.

Since the control sequences in Observation III should always be detectable according to Affymetrix protocols (www.affymetrix.com/support/help/faqs/ge_assays/faq_19.jsp), it has to be assumed that they were either not added to the hybridization cocktails used in the studies corresponding to the dark areas in the glyphs, or that the hybridizations failed or were of poor quality. Given the consistent absence of a strong signal in the affected samples in all six glyphs and the quality control performed by Lukk et al. (2010), it seems more likely that the sequences targeted by the control probes were simply not added.

Given that the control probes should have been expressed highly in all samples, it is probably a coincidence that the LDHB gene, which encodes the lactate dehydrogenase B protein, exhibits a highly similar expression profile, as shown in Figure 4.22. Specifically, LDHB is highly to very highly expressed under almost all conditions, with the prominent exception of the three MCF-7 breast cancer cell lines included in the data set. These are "MCF7 breast epithelial adenocarcinoma", "ssMCF7 breast cancer" and "mcf-7aro breast epithelial adenocarcinoma". This is surprising, given the important role of lactate dehydrogenase in glycolysis, where it catalyzes the reversible conversion of lactate and pyruvate and coenzymes NAD^+ (nicotinamide adenine dinucleotide) and NADH.

A literature search provides an explanation for this observation. Lactate dehydrogenase occurs as either a homo or hetero tetramer that is formed by a combination of two different subunits, A and B, encoded by the LHDA and LHDB genes, respectively. A total of five different enzyme isoforms are known (Markert, 1963). Burke et al. (1978) studied lactate dehydrogenase activity in MCF-7 cell lines and reported that only one of the five enzyme isoforms is present in this cell line. This isoform is a homo tetramer of four A subunits, which are encoded by LDHA. This explains why no activity of the LDHB gene is observed in MCF-7 cell lines.

4.6 Discussion

The growing number of gene expression studies that include hundreds or even thousands of samples prompted this study of the suitability of state-of-the-art visualization tools to support the analysis of such large data sets. Graphical and perceptual scaling problems were identified in these methods and, based on an analysis of low-level analytical pattern discovery tasks, the Space Maps visualization technique was devised to address these issues. The technique extends the Value and Relation (VaR) display by Yang et al. (2007) with hierarchical glyphs that enable a multi-resolution visualization. The glyphs that were designed for this purpose are a hybrid of pixel-oriented visualizations and tree maps. They allow the investigator to literally "drill down" into the data.

The Space Maps method has three major advantages: (i) The method can display complete gene expression profiles with thousands of samples on the screen without requiring the user to scroll or interact in any other way. (ii) Space Maps combine the advantages of dimensionality reduction methods with the visualization of expression profiles. (iii) The hierarchical glyphs generated by the method allow the investigator to explore expression profiles at different levels of detail.

By applying a prototype implementation of the technique to a data set of 1,000 genes measured across more than 5,000 samples, it was shown that the combination of hierarchical glyphs with knowledge-derived and data-derived layout techniques is an efficient approach to visualize and explore gene expression matrices of extreme dimensions. The fact that even the very basic analysis of an extremely large and complex test data set led to biologically meaningful findings indicates that the Space Maps visualization method is useful both for the identification and interpretation of local and global patterns in very large expression data sets.

The main goal of the work described here was to find out how gene expression profiles with hundreds or even thousands of samples can be visualized. In the previous sections it was shown that this design goal

was achieved. In the process, however, some weaknesses of the approach were identified that give room for improvement in the future.

A direction of research is to find a way to reduce the overlap of glyphs in the layout based on dimensionality reduction. Clustering of glyphs in the projection is a possible answer to this problem. However, the clustering algorithm and the visual representation of clusters will have to be chosen carefully so that it is still possible to infer the distribution of profiles in the expression space. A solution for this problem would possibly enable the Space Maps technique to visualize many thousands of expression profiles.

Another weakness that the analysis of the test data set revealed is that it is difficult to explore the data from the perspective of the samples because there is no explicit representation of the sample hierarchy. A future direction of research could be concerned with better integration of the sample hierarchy into the exploration process, for instance by visualization of the hierarchy and integration of the visualization with the generation of glyphs.

A formal user study would provide a better understanding of the advantages and disadvantages of the Space Maps visualization. Such a study could be used, for instance, to evaluate: (i) how difficult it is for investigators to grasp the concept of 2-dimensional, hierarchical expression profiles and (ii) if the interactions provided are sufficient for exploration and analysis of large and complex data sets.

While in this work the Space Maps visualization was applied only in the context of transcriptomics data, it has to be emphasized that as a general visualization technique it is applicable to many other types of data matrices. The only prerequisites for such additional applications are that a hierarchy can be defined on either rows or columns of the matrix and that a meaningful summary statistic can be computed for the inner nodes of the hierarchy.

Conclusions

Weinberg (2010) and Golub (2010) discuss in recent essays the fact that many areas of biology are currently undergoing a paradigm shift. In those areas, biology is changing from a hypothesis-driven science to a data-driven science. In hypothesis-driven science, investigators begin with a hypothesis and collect data to test this hypothesis. The opposite is the case in the data-driven paradigm, where large amounts of data are collected in an unbiased fashion. Not until all data has been collected do the investigators start looking for patterns in the data from which they can generate hypotheses about the system they are studying. This hypothesis generation process requires appropriate methods and tools for data exploration and visualization. In light of the paradigm shift in biology, the work presented in this dissertation is a timely contribution that addresses some of the challenges in data-driven biology and in biology as a "big data science".

To some, however, this dissertation might seem outdated already because the methods and tools presented here were developed for transcriptomics data generated with microarrays. Microarrays undoubtedly have seen their best days and are expected to be replaced by next-generation sequencing technologies in the near future. But it is exactly the fact that microarrays are an "old" technology and have been used for over a decade that made the research presented in this dissertation possible. The two main reasons are that over the years large amounts of data have accumulated in public repositories and that the standardization efforts for the representation of these data had enough time to mature and

become effective. It has to be emphasized that none of the work described in this dissertation would have been possible without these efforts. Since transcriptomics was one of the first fields to be faced with genome-wide data measured across many samples, it has had a strong influence on most other areas in functional genomics that have been established since. Thus transcriptomics microarray data are an excellent test case for the development of large scale data exploration and visualization tools.

At the same time, it is also important to study how other fields are dealing with the challenges of exploration and visualization in the context of large data sets. For instance, there are surprisingly many parallels between the situation in data-driven biology and the intelligence analysis field. Analysts in the intelligence field are also faced with enormous amounts of data and need to be able to efficiently and reliably identify relevant patterns in those data. "Visual analytics" (Thomas and Cook, 2005) is an emerging area of research that has its roots in the intelligence arena. Visual analytics research is concerned with how visualization fits into the larger process of understanding and interpreting data and how analytical and visual approaches can be combined to maximize the efficiency of both humans and computers in this process (Gehlenborg et al., 2010). Visual analytics is a promising area of research to develop solutions for biological "big data" problems, since the analysis of biological data typically involves both visual and computational methods that provide a starting point for visual analytics approaches.

Finally, a crucial insight from the development of the methods described in this dissertation is that not only visualization methods but also computational data exploration approaches in general must be designed with the human investigator in mind. This insight is becoming increasingly more relevant as more and more data is being generated.

Appendix A

Publications

A.1 Manuscripts Included in the Dissertation

1. **N. Gehlenborg**, S. I. O'Donoghue, N. Baliga, A. Goesmann, M. A. Hibbs, H. Kitano, O. Kohlbacher, H. Neuweger, R. Schneider, D. Tenenbaum and A.-C. Gavin. Visualization of Omics Data for Systems Biology. *Nature Methods* 7(3):S56-S68, 2010.

2. S. I. O'Donoghue, A.-C. Gavin, **N. Gehlenborg**, D. S. Goodsell, J. K. Heriche, C. B. Nielsen, C. North, A. J. Olson, J. B. Procter, D. W. Shattuck, T. Walter and B. Wong. Visualizing biological data now and in the future. *Nature Methods* 7(3):S2-S4, 2010.

3. J. Peltonen, H. Aidos, **N. Gehlenborg**, A. Brazma and S. Kaski. An Information Retrieval Perspective on Visualization of Gene Expression Data with Ontological Information. In *Proceedings of the IEEE 2010 International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, 2178-2181, 2010.

4. J. Caldas, **N. Gehlenborg**, A. Faisal, A. Brazma, and S. Kaski. Probabilistic retrieval and visualization of biologically relevant microarray experiments. *Bioinformatics* 25(12):i145-i153, 2009.

5. **N. Gehlenborg** and A. Brazma. Space Maps: Visualization of Gene Expression Profiles with Hundreds of Conditions. *In preparation.*

6. **N. Gehlenborg**, J. Caldas, A. Faisal, E. Kettunen, S. Knutilla, A. Brazma, and S. Kaski. Toward Interpretable Information Retrieval in Transcriptomics. *In preparation.*

A.2 Other Manuscripts Published during PhD

1. **N. Gehlenborg**, D. Hwang, I. Y. Lee, H. Yoo, B. Petritis, D. Baxter, R. Pitstick, B. Marzolf, S. J. DeArmond, G. A. Carlson and L. E. Hood. The Prion Disease Database: A Comprehensive Transcriptome Resource for Systems Biology Research in Prion Diseases. *Database* 2009:bap011, 2009.

2. D. Hwang, I. Y. Lee, H. Yoo, **N. Gehlenborg**, J. Cho, B. Petritis, D. Baxter, R. Pitstick, R. Young, D. Spicer, N. Price, J. G. Hohmann, S. J. DeArmond, G. A. Carlson and L. E. Hood. A Systems Approach to Prion Disease. *Molecular Systems Biology* 5:252, 2009.

3. **N. Gehlenborg**, W. Yan, I. Y. Lee, H. Yoo, K. Nieselt, D. Hwang, R. Aebersold and L. E. Hood. Prequips - An Extensible Software Platform for Integration, Visualization and Analysis of LC-MS/MS Proteomics Data. *Bioinformatics* 25(5):682-68, 2009.

4. A. Schmidt, **N. Gehlenborg**, B. Bodenmiller, L. N. Mueller, D. Campbell, M. Mueller, R. Aebersold and B. Domon. An integrated, directed mass spectrometric approach for in-depth characterization of complex peptide mixtures. *Molecular & Cellular Proteomics* 7:2138-2150, 2008.

Appendix B

Special Factors for the Retrieval of Experiments

B.1 Neutral Factors

Neutral factors are experimental factors that are ignored when comparisons are extracted from a data set. This is because they usually contain a very large number of factor values and in some cases there can be one distinct factor value for every sample in the data set, for example in the case of "individual".

List of Factors

individual
age
replicated
population
familyhistory
envhistory

B.2 Control Factor Values

Control factor values designate samples that were used as controls in a study. Within the context of the study they are representing the "normal" biological state against which other states are compared. The strings shown below were extracted directly from annotation of the data sets.

List of Factor Values

0
0 cm away from the tumor boundary
0 d
0 days
0 Gy
0 h
0 hours
0 hours per day access
0 IU
0 IU_per_ml
0 m
0 M
0 mg_per_kg
0 mg_per_kg_per_day
0 mg/kg
0 mg/kg x 2 doses per day
0 mg/kg/day
0 mM
0 mol_per_L
0 ng/ml
0 ng/mL
0 nM
0 nmol
0 ppm

0 U/kg
0 U/ml
0 ug
0 ug_per_kg
0 ug_per_mL
0 ug/kg
0 uM
0 umol
0 umol_per_kg
0 umol/kg
aortic banding - sham
av-fistula - sham
control - 37 degree_C
control - albumin
control - BSA
control - ethanol
control - IL-1b
control - interferon-gamma
control - keratinocyte growth factor
control - unsychronized
control - untreated
control - vector
control - vehicle
control diet
control for EGF
control for heregulin
control polyamide and dihydrotestoterone
control siRNA
control
empty vector
mock
mock infected
mock transfected

myocardial infarction - sham
non-smoker
none
normal
normal 2
normal 9
normal contralateral cartilage
normal diet
normal donor
normal growth media (bone marrow)
normal growth media (C85)
normal terminal duct lobular unit
normal tissue from invasive ductal carcinoma patient
normal tissue from invasive lobular carcinoma patient
normal1
normal2
normal3
normal4
normal5
placebo
reference
SCA1 wild_type
SCA7 wild_type
sham
sham
sham denervation
sham fracture
sham injury
sham surgery
sham surgery - contralateral right hind limb
sham surgery - ipsilateral left hind limb
sham surgery cartilage
uninduced

uninfected
untreated
wild type
wild type
wild type SOD1 transgenic
wild type T cell receptor
wild_type

Appendix C

Space Maps Project File

The Space Maps prototype implementation described in Section 4.4 requires a project file as input. A sample project file is shown in Listing C.1.

The prototype implements parsers for a range of different input file formats. The arguments of the parsers, such as the column to read from a tab-delimited text file or the number of rows to skip at the beginning of a file, are set in the project file. Additionally, the project file can be used to call generator functions in the prototype that dynamically create locations on various space-filling curves or random attributes for testing, rather than reading this information from a file.

Listing C.1: Sample Space Maps prototype project file. The project file refers to a range of other files that contain expression matrices as well as hierarchies and attributes. In the prototype implementation it is required that all attribute files assume the same order of genes.

```
<?xml version="1.0"?>

<model>

 <!-- general project settings -->
 <general>
   <path></path>
   <size>1000</size>
   <name>1000 genes</name>
 </general>

 <!-- collections of gene attributes -->
 <collections>
```

```
<!-- location attributes: projections and space filling curves -->
<location>
  <resource name="NeRV (lambda = 0.0, 5372 cond)" type="sompak_file" value="top1000.nerv_l00"/>
  <resource name="NeRV (lambda = 0.3, 96 cond)" type="sompak_file" value="nerv_l03.txt"/>

  <resource name="PC1/PC2" type="tab_file" value="pca.txt" skip="0" column1="0" column2="1"/>
  <resource name="PC2/PC3" type="tab_file" value="pca.txt" skip="0" column1="1" column2="2"/>

  <resource name="Random" type="random_generator" value="2"/>
  <resource name="Hilbert" type="curve_generator" value="hilbert"/>
  <resource name="Spiral" type="curve_generator" value="spiral"/>
</location>

<!-- expression matrices and associated hierarchies -->
<vector>
  <resource name="Expr. Data" type="tab_file" value="top1000.txt" skip="1" column="1" header="0">
    <hierarchy name="Margus 4-15-371-5372" type="indent_file" value="4-15-371-5372_hier.txt"/>
    <hierarchy name="Margus 7-97-5372" type="indent_file" value="7-97-5372_hier.txt"/>
  </resource>
</vector>

<!-- quantitative attributes -->
<ratio>
  <resource name="Static" type="static_generator" value="1"/>
  <resource name="Variance" type="tab_file" value="gene_variance.txt" skip="0" column="0"/>
  <resource name="Overall Expression" type="tab_file" value="gene_sum.txt" skip="0" column="0"/>
</ratio>

<!-- string attributes -->
<name>
  <resource name="Affymetrix ID" type="tab_file" value="id.txt" skip="0" column="0"/>
  <resource name="Gene Symbol" type="tab_file" value="id.txt" skip="0" column="1"/>
</name>

<!-- categorical attributes -->
<class>
  <resource name="GO CC" type="tab_file" value="go_cc_first.txt" skip="0" column="0"/>
  <resource name="GO MF" type="tab_file" value="go_mf_first.txt" skip="0" column="0"/>
  <resource name="GO BP" type="tab_file" value="go_bp_first.txt" skip="0" column="0"/>
  <resource name="10 clusters" type="tab_file" value="hcl10.txt" skip="0" column="0"/>
  <resource name="20 clusters" type="tab_file" value="hcl20.txt" skip="0" column="0"/>
</class>

<!-- rank attributes: used to determine ordering on space filling curves -->
<rank>
  <resource name="Hier. Clustering" type="tab_file" value="hcl_order.txt" skip="0" column="0"/>
</rank>

<!-- texture attributes: image files can be loaded to replace glyphs -->
```

```
    <texture>
      <resource name="Texture Test" type="tab_file" value="images.txt" skip="0" column="0"/>
    </texture>

  </collections>

</model>
```

Appendix D

Space Maps Case Study Data

Tables D.1, D.2 and D.3 contain the Gene Ontology annotation for the gene list discussed in the Space Maps case study in Section 4.5.3. The mapping from gene symbols to Gene Ontology terms was obtained from Ensembl BioMart (Release 59; Kasprzyk et al., 2004).

Table D.1: Gene Ontology Biological Process annotation for genes in Observation I.

ACTA1	cell growth, muscle contraction, response to extracellular stimulus, response to lithium ion, response to mechanical stimulus, response to steroid hormone stimulus, skeletal muscle fiber adaptation, skeletal muscle fiber development, skeletal muscle thin filament assembly
ACTN2	cell adhesion, focal adhesion assembly, microspike assembly, muscle contraction, protein homotetramerization, regulation of apoptosis
CKM	creatine metabolic process, phosphocreatine biosynthetic process
CLIC5	auditory receptor cell stereocilium organization, chloride transport, female pregnancy, ion transport, neuromuscular process controlling balance, protein localization, sensory perception of sound, transport
CSRP3	blood vessel remodeling, cardiac muscle tissue development, cell differentiation, cellular calcium ion homeostasis, multicellular organismal development, muscle organ development, regulation of the force of heart contraction, skeletal muscle tissue development
KBTBD10	DNA repair, regulation of lateral pseudopodium assembly, striated muscle contraction
LDB3	-
MYBPC1	cell adhesion
MYH2	muscle contraction, muscle filament sliding, plasma membrane repair, response to activity
MYH6	actin filament-based movement, adult heart development, ATP catabolic process, atrial cardiac muscle tissue morphogenesis, cardiac muscle fiber development, in utero embryonic development, muscle contraction, muscle filament sliding, myofibril assembly, regulation of ATPase activity, regulation of blood pressure, regulation of heart contraction, regulation of heart rate, regulation of the force of heart contraction, sarcomere organization, striated muscle contraction, ventricular cardiac muscle tissue morphogenesis, visceral muscle development
MYL2	cardiac myofibril assembly, heart contraction, negative regulation of cell growth, regulation of striated muscle contraction, ventricular cardiac muscle tissue morphogenesis
MYOT	muscle contraction
MYOZ2	biological process
NEB	muscle organ development, regulation of actin filament length, somatic muscle development
PDE4DIP	-
PDLIM5	regulation of dendritic spine morphogenesis, regulation of synaptogenesis
PLN	blood circulation, calcium ion transport, cardiac muscle tissue development, cellular calcium ion homeostasis, negative regulation of heart contraction, regulation of calcium ion transport, regulation of the force of heart contraction
SMPX	striated muscle contraction
TNNC1	cardiac muscle contraction, diaphragm contraction, regulation of ATPase activity, regulation of muscle contraction, regulation of muscle filament sliding speed, response to metal ion, ventricular cardiac muscle tissue morphogenesis
TTN	carbohydrate metabolic process, heart development, protein amino acid phosphorylation, proteolysis, sarcomere organization, somitogenesis

Table D.2: Gene Ontology Molecular Function annotation for genes in Observation I.

ACTA1	ADP binding, ATP binding, myosin binding, nucleotide binding, protein binding, structural constituent of cytoskeleton
ACTN2	actin binding, actin filament binding, calcium ion binding, FATZ 1 binding, identical protein binding, integrin binding, ligand-dependent nuclear receptor transcription coactivator activity, protein binding, protein dimerization activity, structural constituent of muscle, thyroid hormone receptor coactivator activity, titin binding, titin Z domain binding, ZASP binding
CKM	ATP binding, creatine kinase activity, kinase activity, nucleotide binding, transferase activity, transferring phosphorus-containing groups
CLIC5	chloride channel activity, protein binding, voltage-gated chloride channel activity, voltage-gated ion channel activity
CSRP3	metal ion binding, protein binding, zinc ion binding
KBTBD10	methylated-DNA-[protein]-cysteine S-methyltransferase activity, protein binding
LDB3	metal ion binding, protein binding, protein kinase C binding, zinc ion binding
MYBPC1	actin binding, structural constituent of muscle, titin binding
MYH2	actin binding, ATP binding, calmodulin binding, microfilament motor activity, motor activity, nucleotide binding, protein binding, structural constituent of muscle
MYH6	actin binding, actin-dependent ATPase activity, ATP binding, ATPase activity, calcium-dependent ATPase activity, calmodulin binding, microfilament motor activity, motor activity, nucleotide binding, protein binding, protein heterodimerization activity, protein homodimerization activity, structural constituent of muscle
MYL2	actin monomer binding, calcium ion binding, motor activity, myosin heavy chain binding, protein binding, structural constituent of muscle
MYOT	actin binding, protein binding, structural constituent of muscle
MYOZ2	protein binding, protein phosphatase 2B binding
NEB	actin binding, oxidoreductase activity, structural constituent of muscle
PDE4DIP	protein binding
PDLIM5	actin binding, actinin binding, metal ion binding, protein binding, protein kinase C binding, protein N-terminus binding, zinc ion binding
PLN	ATPase inhibitor activity, calcium channel regulator activity, protein binding
SMPX	-
TNNC1	actin filament binding, calcium ion binding, calcium-dependent protein binding, protein binding, protein homodimerization activity, troponin I binding, troponin T binding
TTN	ATP binding, cysteine-type endopeptidase activity, phosphotransferase activity, alcohol group as acceptor, protein kinase activity, protein serine/threonine kinase activity, protein tyrosine kinase activity, structural constituent of cell wall, structural constituent of muscle

Table D.3: Gene Ontology Cellular Component annotation for genes in Observation I.

ACTA1	actin cytoskeleton, actin filament, cytoplasm, sarcomere, stress fiber, striated muscle thin filament
ACTN2	actin cytoskeleton, actin filament, cytoplasm, cytoskeleton, cytosol, dendritic spine, filopodium, focal adhesion, nucleolus, pseudopodium, Z disc
CKM	cytoplasm, cytosol
CLIC5	actin cytoskeleton, cell cortex, chloride channel complex, cytoplasm, Golgi apparatus, insoluble fraction, integral to membrane, membrane, microtubule organizing center, stereocilium, stereocilium bundle
CSRP3	cytoplasm, cytoskeleton, nucleus, Z disc
KBTBD10	cytoplasm, cytoskeleton, plasma membrane, pseudopodium, ruffle
LDB3	cytoplasm, cytoskeleton, perinuclear region of cytoplasm, pseudopodium, Z disc
MYBPC1	myofibril, myosin filament
MYH2	A band, actomyosin contractile ring, cytoplasm, focal adhesion, Golgi apparatus, muscle myosin complex, myofibril, myosin complex, myosin filament, nucleolus, protein complex, sarcomere
MYH6	contractile fiber, cytoplasm, focal adhesion, muscle myosin complex, myofibril, myosin complex, myosin filament, nucleolus, nucleus, perinuclear region of cytoplasm, sarcomere
MYL2	actin cytoskeleton, cytoskeleton, myofibril, myosin complex, sarcomere
MYOT	actin cytoskeleton, cytoplasm, sarcolemma, Z disc
MYOZ2	actin cytoskeleton, cytoplasm, sarcomere, Z disc
NEB	actin cytoskeleton, cytoplasm, sarcomere, Z disc
PDE4DIP	cytoplasm, cytoskeleton, Golgi apparatus, microtubule organizing center, nucleus
PDLIM5	actin cytoskeleton, cell junction, cytoplasm, cytosol, membrane fraction, plasma membrane, postsynaptic density, postsynaptic membrane, synapse, synaptosome, Z disc
PLN	integral to membrane, membrane, mitochondrial membrane, mitochondrion, protein complex, sarcoplasmic reticulum, vesicle
SMPX	contractile fiber, costamere, cytoplasm, M band, muscle tendon junction, nucleus
TNNC1	contractile fiber, troponin complex
TTN	I band, myofibril

Bibliography

C. Ahlberg and B. Shneiderman. Visual information seeking: tight coupling of dynamic query filters with starfield displays. *CHI '94: Conference companion on Human factors in computing systems*, 1994.

A. E. Allen-Jennings, M. G. Hartman, G. J. Kociba, and T. Hai. The roles of ATF3 in glucose homeostasis. A transgenic mouse model with liver dysfunction and defects in endocrine pancreas. *The Journal of Biological Chemistry*, 276(31):29507–14, Aug. 2001.

D. B. Allison, X. Cui, G. P. Page, and M. Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*, 7(1):55–65, 2006.

R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 111–117. IEEE, 2005.

F. J. Anscombe. Graphs in Statistical Analysis. *The American Statistician*, 27 (1):17, 1973.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25 (1):25–9, 2000.

W.-Y. Au, K. K. Harod, and M.-F. Law. Cough mixture abuse, folate deficiency and acute lymphoblastic leukemia. *Leukemia Research*, 33(3):508–9, Mar. 2009.

A. Aula and H. Siirtola. Hundreds of Folders or One Ugly Pile Strategies for Information Search and Re-access. In M. F. Costabile and F. Paternò, editors, *Human-Computer Interaction - INTERACT 2005*, pages 954–957, Berlin, Germany, 2005. Springer.

S. Avraham, C.-W. Tung, K. Ilic, P. Jaiswal, E. A. Kellogg, S. McCouch, A. Pujar, L. Reiser, S. Y. Rhee, M. M. Sachs, M. Schaeffer, L. Stein, P. Stevens, L. Vincent, F. Zapata, and D. Ware. The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. *Nucleic Acids Research*, 36(Database issue): D449–54, 2008.

E. H. Baehrecke, N. Dang, K. Babaria, and B. Shneiderman. Visualization and analysis of microarray and gene ontology data with treemaps. *BMC Bioinformatics*, 5:84, 2004.

J. Bard, S. Y. Rhee, and M. Ashburner. An ontology for cell types. *Genome Biology*, 6(2):R21, 2005.

T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, and R. Edgar. NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Research*, 35:D760—-5, 2007.

T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muertter, and R. Edgar. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res*, 37(Database issue): D885—-90, 2009.

D. P. Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–97, 2004.

J. Bertin. *Semiology of Graphics*. Wisconsin University Press, Madison, WI, USA, 1983.

D. M. Blei and J. D. Lafferty. A correlated topic model of Science. *Annals of Applied Statistics*, 1(1):17–35, 2007.

D. M. Blei, M. I. Jordan, T. L. Griffiths, and J. B. Tenenbaum. Hierarchical Topic Models and the Nested Chinese Restaurant Process. In S. Thrun, L. Saul, and B. Schoelkopf, editors, *Advances in Neural Information Processing Systems 16*, 2003a.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003b.

B. M. Bolstad, R. A. Irizarry, M. Astrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

R. Bonneau. Learning biological networks: from modules to dynamics. *Nature chemical biology*, 4(11):658–64, 2008.

M. Bostock and J. Heer. Protovis: a graphical toolkit for visualization. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1121–8, 2009.

J.-P. Bourquin, A. Subramanian, C. Langebrake, D. Reinhardt, O. Bernard, P. Ballerini, A. Baruchel, H. Cavé, N. Dastugue, H. Hasle, G. L. Kaspers, M. Lessard, L. Michaux, P. Vyas, E. van Wering, C. M. Zwaan, T. R. Golub, and S. H. Orkin. Identification of distinct molecular phenotypes in acute megakaryoblastic leukemia by gene expression profiling. *Proceedings of the National Academy of Sciences of the United States of America*, 103(9):3339–44, 2006.

A. Brazma. Minimum Information About a Microarray Experiment (MIAME)–successes, failures, challenges. *The Scientific World Journal*, 9: 420–3, 2009.

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

M. Bruls, K. Huizing, and J. J. V. Wijk. Squarified Treemaps. In *Proceedings of the Joint Eurographics and IEEE TCVG Symposium on Visualization, 2000 (Data Visualization 2000)*, pages 33–42, 2000.

W. Buntine and A. Jakulin. Applying Discrete PCA in Data Analysis. In D. M. Chickering and J. Y. Halpern, editors, *Proceedings of the Twentieth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-04)*, pages 59–66, Arlington, VA, 2004. AUAI Press.

R. E. Burke, S. C. Harris, and W. L. McGuire. Lactate dehydrogenase in estrogen-responsive human breast cancer cells. *Cancer Research*, 38(9): 2773–6, 1978.

A. J. Butte and I. S. Kohane. Creation and implications of a phenome-genome network. *Nature Biotechnology*, 24(1):55–62, 2006.

S. K. Card, J. D. Mackinlay, and B. Shneiderman. *Readings in Information Visualization - Using Vision to Think*. Morgan Kaufmann, San Francisco, CA, USA, 1999.

H. Causton, A. Brazma, and J. Quackenbush. *Microarray Gene Expression Data Analysis: A Beginners Guide*. Wiley-Blackwell, 2003.

N. Chahin, S. R. Zeldenrust, K. K. Amrami, J. K. Engelstad, and P. J. B. Dyck. Two causes of demyelinating neuropathy in one patient: CMT1A and POEMS syndrome. *The Canadian Journal of Neurological Sciences*, 34(3): 380–5, 2007.

M. Chen and M. A. Hearst. Presenting Web site search results in context. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '98*, page 381, New York, NY, USA, 1998. ACM Press.

H. Chernoff. The Use of Faces to Represent Points in K-Dimensional Space Graphically. *Journal of the American Statistical Association*, 68(342):361, 1973.

G. A. Churchill. Fundamentals of experimental design for cDNA microarrays. *Nature Genetics*, 32 Suppl:490–5, 2002.

W. S. Cleveland and R. McGill. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984.

J. M. Coller, N. K. Gray, and M. P. Wickens. mRNA stabilization by poly(A) binding protein is independent of poly(A) and requires translation. *Genes & Development*, 12(20):3226–3235, Oct. 1998.

D. J. Craigon, N. James, J. Okyere, J. Higgins, J. Jotham, and S. May. NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic Acids Research*, 32(Database issue):D575–7, 2004.

J. F. Degner, J. C. Marioni, A. A. Pai, J. K. Pickrell, E. Nkadori, Y. Gilad, and J. K. Pritchard. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. *Bioinformatics*, 25(24):3207–12, 2009.

J. Dekker. Gene regulation in the third dimension. *Science*, 319(5871):1793–4, Mar. 2008.

D. DeRidder, O. Kouropteva, O. Okun, and M. Pietik. Supervised Locally Linear Embedding. In *Proceedings of the 2003 Joint International Conference on Artificial Neural Networks and Neural Information Processing*, pages 333–341. Springer, 2003.

P. Dyck and R. Chisholm. Disease ontology: structuring medical billing codes for medical record mining and disease gene association. *Proceedings of the Sixth Annual Bio-ontologies Meeting*, 2003.

R. Edgar and T. Barrett. NCBI GEO standards and services for microarray data. *Nature Biotechnology*, 24(12):1471–2, 2006.

R. Edgar, M. Domrachev, and A. E. Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–10, 2002.

S. G. Eick and A. F. Karr. Visual scalability. *Journal of Computational and Graphical Statistics*, 11:22–43, 2002.

M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25):14863–8, 1998.

J. A. Engelman. Targeting PI3K signalling in cancer: opportunities, challenges and limitations. *Nature Reviews Cancer*, 9(8):550–62, 2009.

A. Eulalio, E. Huntzinger, T. Nishihara, J. Rehwinkel, M. Fauser, and E. Izaurralde. Deadenylation is a widespread effect of miRNA regulation. *RNA*, 15(1):21–32, 2009.

T. A. Eyre, F. Ducluzeau, T. P. Sneddon, S. Povey, E. A. Bruford, and M. J. Lush. The HUGO Gene Nomenclature Database, 2006 updates. *Nucleic acids research*, 34(Database issue):D319–21, 2006.

J. D. Fekete and C. Plaisant. Interactive information visualization of a million items. *IEEE Symposium on Information Visualization*, 2002.

J. D. Fekete, J. J. V. Wijk, J. T. Stasko, and C. North. *The Value of Information Visualization*, pages 1–18. Springer, New York, 2008.

P. M. Fitts. The Information Capacity of the Human Motor System in Controlling the Amplitude of Movement. *Journal of Experimental Psychology*, 47:381–391, 1954.

P. Flaherty, G. Giaever, J. Kumm, M. I. Jordan, and A. P. Arkin. A latent variable model for chemogenomic profiling. *Bioinformatics*, 21(15):3286–93, 2005.

T. C. Freeman, L. Goldovsky, M. Brosch, S. van Dongen, P. Maziere, R. J. Grocock, S. Freilich, J. Thornton, and A. J. Enright. Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS Computational Biology*, 3(10):2032–2042, 2007.

W. Fujibuchi, L. Kiseleva, T. Taniguchi, H. Harada, and P. Horton. CellMontage: similar expression profile search server. *Bioinformatics*, 23(22): 3103–4, 2007.

N. Funel, S. Pelliccioni, M. Del Chiaro, L. Pollina, A. Michelucci, P. Simi, F. Mosca, U. Boggi, and D. Campani. PMP22 Gene Duplication in Pancreatic Ductal Adenocarcinoma. *Journal of the Pancreas*, 10(5 Suppl):616, 2009.

C. M. Gabriel, N. A. Gregson, and R. A. Hughes. Anti-PMP22 antibodies in patients with inflammatory neuropathy. *Journal of Neuroimmunology*, 104(2):139–46, 2000.

N. Gehlenborg, J. Dietzsch, and K. Nieselt. A Framework for Visualization of Microarray Data and Integrated Meta Information. *Information Visualization*, 4(3):164–175, 2005.

N. Gehlenborg, S. I. O'Donoghue, N. S. Baliga, A. Goesmann, M. A. Hibbs, H. Kitano, O. Kohlbacher, H. Neuweger, R. Schneider, D. Tenenbaum, and A.-C. Gavin. Visualization of omics data for systems biology. *Nature Methods*, 7(3 Suppl):S56–68, 2010.

R. Gentleman, V. J. Carey, W. Huber, S. Dudoit, and R. A. Irizarry. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, Berlin, 2005.

G. K. Gerber, R. D. Dowell, T. S. Jaakkola, and D. K. Gifford. Automated discovery of functional generality of human gene expression programs. *PLoS Computational Biology*, 3(8):e148, Aug. 2007.

G. Giambonini-Brugnoli, J. Buchstaller, L. Sommer, U. Suter, and N. Mantei. Distinct disease mechanisms in peripheral neuropathies due to altered peripheral myelin protein 22 gene dosage or a Pmp22 point mutation. *Neurobiology of Disease*, 18(3):656–68, 2005.

S. A. Glynn and D. Albanes. Folate and cancer: a review of the literature. *Nutrition and cancer*, 22(2):101–19, Jan. 1994.

T. Golub. Counterpoint: Data first. *Nature*, 464(7289):679, Apr. 2010.

C. Gotsman and M. Lindenbaum. On the metric properties of discrete space-filling curves. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, pages 98–102. IEEE, 1996.

T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101 Suppl: 5228–35, Apr. 2004.

C. Healey and J. Enns. Large datasets at a glance: combining textures and colors in scientific visualization. *IEEE Transactions on Visualization and Computer Graphics*, 5(2):145–167, 1999.

C. Healey, K. Booth, and J. Enns. Harnessing preattentive processes for multivariate data visualization. *Proceedings Graphics Interface '93*, pages 107–117, 1993.

M. A. Hearst. *Search User Interfaces*. Cambridge University Press, Cambridge, UK, 2009.

J. Heer and M. Agrawala. Multi-Scale Banking to 45 Degrees. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):701–708, 2006.

J. Heer and G. Robertson. Animated transitions in statistical data graphics. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1240–7, 2007.

M. A. Hibbs, D. C. Hess, C. L. Myers, C. Huttenhower, K. Li, and O. G. Troyanskaya. Exploring the functional landscape of gene expression: directed search of large microarray compendia. *Bioinformatics*, 23(20):2692–9, Oct. 2007.

D. Hilbert. Über die stetige Abbildung einer Linie auf ein Flächenstück. *Mathematische Annalen*, 38:459–460, 1891.

A. Hodel. SNAP-25. *The International Journal of Biochemistry and Cell Biology*, 30:1069–1073, 1998.

A. V. Hoffbrand, J. S. Stewart, C. C. Booth, and D. L. Mollin. Folate deficiency in Crohn's disease: incidence, pathogenesis, and treatment. *British Medical Journal*, 2(5597):71–5, Apr. 1968.

R. Hoffmann. A wiki for the life sciences where authorship matters. *Nature Genetics*, 40(9):1047–51, 2008.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *The Journal of Educational Psychology*, 24:417–441, 1933.

T. Hruz, O. Laule, G. Szabo, F. Wessendorp, S. Bleuler, L. Oertle, P. Widmayer, W. Gruissem, and P. Zimmermann. Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. *Advances in Bioinformatics*, 2008:420747, 2008.

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

D. W. Huang, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, 2009.

T. J. P. Hubbard, B. L. Aken, S. Ayling, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, L. Clarke, G. Coates, S. Fairley, S. Fitzgerald, J. Fernandez-Banet, L. Gordon, S. Graf, S. Haider, M. Hammond, R. Holland, K. Howe, A. Jenkinson, N. Johnson, A. Kahari, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, D. Rios, M. Schuster, G. Slater, D. Smedley, W. Spooner, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, S. Wilder, A. Zadissa, E. Birney, F. Cunningham, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, A. Kasprzyk, G. Proctor, J. Smith, S. Searle, and P. Flicek. Ensembl 2009. *Nucleic Acids Research*, 37(Database issue):D690–7, 2009.

J. Hubble, J. Demeter, H. Jin, M. Mao, M. Nitzberg, T. B. K. Reddy, F. Wymore, Z. K. Zachariah, G. Sherlock, and C. A. Ball. Implementation of GenePattern within the Stanford Microarray Database. *Nucleic Acids Research*, 37(Database issue):D898–901, 2009.

W. Huber, A. von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18 Suppl 1(1997): S96–104, 2002.

T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102: 109–126, 2000.

L. Hunter, R. C. Taylor, S. M. Leach, and R. Simon. GEST: a gene expression search tool based on a novel Bayesian similarity metric. *Bioinformatics*, 17 Suppl 1:S115–22, 2001.

J. S. Huo, R. C. McEachin, T. X. Cui, N. K. Duggal, T. Hai, D. J. States, and J. Schwartz. Profiles of growth hormone (GH)-regulated genes reveal time-dependent responses and identify a mechanism for regulation of activating transcription factor 3 by GH. *The Journal of Biological Chemistry*, 281(7):4132–41, 2006.

D. Hwang, I. Y. Lee, H. Yoo, N. Gehlenborg, J.-H. Cho, B. Petritis, D. Baxter, R. Pitstick, R. Young, D. Spicer, N. D. Price, J. G. Hohmann, S. J. Dearmond, G. A. Carlson, and L. E. Hood. A systems approach to prion disease. *Mol Syst Biol*, 5:252, 2009.

A. Inselberg. The Plane with Parallel Coordinates. *The Visual Computer*, 1: 69–92, 1985.

J. P. A. Ioannidis, D. B. Allison, C. A. Ball, I. Coulibaly, X. Cui, A. C. Culhane, M. Falchi, C. Furlanello, L. Game, G. Jurman, J. Mangion, T. Mehta, M. Nitzberg, G. P. Page, E. Petretto, and V. van Noort. Repeatability of published microarray gene expression analyses. *Nature Genetics*, 41(2): 149–55, 2009.

P. Jaccard. Nouvelles recherches sur la distribution florale. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 44:223–270, 1908.

K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

B. Johnson and B. Shneiderman. Tree maps: A space-filling approach to the visualization of hierarchical information structures. *Proceedings of the 2nd International IEEE Visualization Conference*, pages 284–291, 1991.

B. Jost and C. North. The Perceptual Scalability of Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 12:837–844, 2006.

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

M. Kapushesky, I. Emam, E. Holloway, P. Kurnosov, A. Zorin, J. Malone, G. Rustici, E. Williams, H. Parkinson, and A. Brazma. Gene expression atlas at the European bioinformatics institute. *Nucleic Acids Research*, 38: D690–8, 2010.

A. Kasprzyk, D. Keefe, D. Smedley, D. London, W. Spooner, C. Melsopp, M. Hammond, P. Rocca-Serra, T. Cox, and E. Birney. EnsMart: a generic system for fast and flexible access to biological data. *Genome Research*, 14 (1):160–9, 2004.

F. Katagiri and J. Glazebrook. Overview of mRNA expression profiling using DNA microarrays. *Current Protocols in Molecular Biology*, Chapter 22:Unit 22.4, 2009.

A. Kauffmann, T. F. Rayner, H. Parkinson, M. Kapushesky, M. Lukk, A. Brazma, and W. Huber. Importing ArrayExpress datasets into R/Bioconductor. *Bioinformatics*, 25(16):2092–4, 2009.

D. A. Keim. Pixel-oriented visualization techniques for exploring very large databases. *Journal of Computational and Graphical Statistics*, 5(1):58–77, 1996.

D. A. Keim. Designing Pixel-Oriented Visualization Techniques: Theory and Applications. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):59–78, 2000.

D. A. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):100–107, 2002.

M. K. Kerr and G. A. Churchill. Experimental design for gene expression microarrays. *Biostatistics*, 2(2):183–201, 2001.

S. Kilpinen, R. Autio, K. Ojala, K. Iljin, E. Bucher, H. Sara, T. Pisto, M. Saarela, R. I. Skotheim, M. Bjorkman, J.-P. Mpindi, S. Haapa-Paananen, P. Vainio, H. Edgren, M. Wolf, J. Astola, M. Nees, S. Hautaniemi, and O. Kallioniemi. Systematic bioinformatic analysis of expression levels of

17,330 human genes across 9,783 samples from 175 types of healthy and pathological tissues. *Genome Biology*, 9:R139, 2008.

R. Kincaid and H. Lam. Line Graph Explorer: scalable display of line graphs using Focus+Context. *AVI '06: Proceedings of the Working Conference on Advanced Visual Interfaces*, pages 404–411, 2006.

T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, 1982.

R. Kosara, S. Miksch, and H. Hauser. Semantic depth of field. *IEEE Symposium on Information Visualization*, pages 97–104, 2001.

I. Kozarewa, Z. Ning, M. A. Quail, M. J. Sanders, M. Berriman, and D. J. Turner. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nature Methods*, 6(4):291–5, 2009.

J. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–26, 1964.

D. H. Lackner and J. Bähler. Translational control of gene expression from transcripts to transcriptomes. *International Review of Cell and Molecular Biology*, 271(08):199–251, 2008.

J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong, S. J. Haggarty, P. A. Clemons, R. Wei, S. A. Carr, E. S. Lander, and T. R. Golub. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–35, Sept. 2006.

C. Langhi, C. Le May, V. Gmyr, B. Vandewalle, J. Kerr-Conte, M. Krempf, F. Pattou, P. Costet, and B. Cariou. PCSK9 is expressed in pancreatic delta-cells and does not alter insulin secretion. *Biochemical and Biophysical Research Communications*, 390(4):1288–93, 2009.

D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–91, 1999.

H. K. Lee, A. K. Hsu, J. Sajdak, J. Qin, and P. Pavlidis. Coexpression analysis of human genes across many microarray data sets. *Genome Research*, 14(6):1085–94, June 2004.

M. D. Lee, R. E. Reilly, and M. A. Butavicius. An Empirical Evaluation of Chernoff Faces , Star Glyphs , and Spatial Visualizations for Binary Data. In *Proceedings of the Asia-Pacific Symposium on Information Visualisation*, volume 24, pages 1–10, 2003.

T. I. Lee and R. A. Young. Transcription of eukaryotic protein-coding genes. *Annual Review of Genetics*, 34:77–137, 2000.

Y. H. Lee, S. Nair, E. Rousseau, D. B. Allison, G. P. Page, P. A. Tataranni, C. Bogardus, and P. A. Permana. Microarray profiling of isolated abdominal subcutaneous adipocytes from obese vs non-obese Pima Indians: increased expression of inflammation-related genes. *Diabetologia*, 48(9): 1776–83, 2005.

J. Li, J. Kleeff, I. Esposito, H. Kayed, K. Felix, T. Giese, M. W. Büchler, and H. Friess. Expression analysis of PMP22/Gas3 in premalignant and malignant pancreatic lesions. *The Journal of Histochemistry and Cytochemistry*, 53(7):885–93, 2005.

X. Lin. Map displays for information retrieval. *Journal of the American Society for Information Science*, 48(1):40–54, 1997.

A. Lindenmayer. Mathematical models for cellular interactions in development. *Journal of Theoretical Biology*, 18(3):280–299, 1968.

M. Lukk, M. Kapushesky, J. Nikkilä, H. Parkinson, A. Goncalves, W. Huber, E. Ukkonen, and A. Brazma. A global map of human gene expression. *Nature Biotechnology*, 28(4):322–4, 2010.

J. Mackinlay. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics*, 5(2):110–141, 1986.

J. Malone, T. F. Rayner, X. Zheng Bradley, and H. Parkinson. Developing an application focused experimental factor ontology: embracing the OBO Community. *Proceedings of the Eleventh Annual Bio-ontologies Meeting*, 2008.

J. Malone, E. Holloway, T. Adamusiak, M. Kapushesky, J. Zheng, N. Kolesnikov, A. Zhukova, A. Brazma, and H. Parkinson. Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, 26(8): 1112–8, 2010.

L. Mamanova, R. M. Andrews, K. D. James, E. M. Sheridan, P. D. Ellis, C. F. Langford, T. W. B. Ost, J. E. Collins, and D. J. Turner. FRT-seq: amplification-free, strand-specific transcriptome sequencing. *Nature Methods*, 7(2):130–2, 2010.

C. D. Manning, P. Raghavan, and H. Schütze. *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK, 2009.

G. Marchionini. *Information Seeking in Electronic Environments*. Cambridge University Press, Cambridge, UK, 1995.

G. Marchionini. From finding to understanding. *Communications of the ACM*, 49(4):41–46, 2006.

R. Margueron, P. Trojer, and D. Reinberg. The key to development: interpreting the histone code? *Current Opinion in Genetics & Development*, 15(2): 163–76, 2005.

C. L. Markert. Lactate Dehydrogenase Isozymes: Dissociation and Recombination of Subunits. *Science*, 140(3573):1329–30, 1963.

M. Mbikay, F. Sirois, J. Mayne, G.-S. Wang, A. Chen, T. Dewpura, A. Prat, N. G. Seidah, M. Chretien, and F. W. Scott. PCSK9-deficient mice exhibit impaired glucose tolerance and pancreatic islet abnormalities. *FEBS Letters*, 584(4):701–6, 2010.

R. McGill, J. W. Tukey, and W. A. Larsen. Variations of Box Plots. *The American Statistician*, 32(1):12, 1978.

G. Meyer zu Hörste, T. Prukop, K.-A. Nave, and M. W. Sereda. Myelin disorders: Causes and perspectives of Charcot-Marie-Tooth neuropathy. *Journal of molecular neuroscience : MN*, 28(1):77–88, 2006.

V. K. Mootha, C. M. Lindgren, K.-F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrå le, E. Laurila, N. Houstis, M. J. Daly, N. Patterson, J. P. Mesirov, T. R. Golub, P. Tamayo, B. Spiegelman, E. S. Lander, J. N. Hirschhorn, D. Altshuler, and L. C. Groop. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34(3):267–73, 2003.

C. J. Morris, D. S. Ebert, and P. Rheingans. Experimental analysis of the effectiveness of features in Chernoff faces. In *28th AIPR Workshop: 3D Visualization for Data Exploration and Decision Making*, pages 12–17, 2000.

J. R. Nevins and A. Potti. Mining gene expression profiles: expression signatures as cancer phenotypes. *Nature Reviews Genetics*, 8(8):601–609, 2007.

R. Niedermeier, K. Reinhardt, and P. Sanders. Towards optimal locality in mesh-indexings. *Discrete Applied Mathematics*, 117(1-3):211–237, 2002.

I.-L. Nilsson, J. Zedenius, L. Yin, and A. Ekbom. The association between primary hyperparathyroidism and malignancy: nationwide cohort analysis on cancer incidence after parathyroidectomy. *Endocrine-Related Cancer*, 14(1):135–40, 2007.

NIST. NIST/SEMATECH e-Handbook of Statistical Methods, 2010. URL http://www.itl.nist.gov/div898/handbook.

S. Oba, M.-A. Sato, I. Takemasa, M. Monden, K.-I. Matsubara, and S. Ishii. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16):2088–96, 2003.

A. P. Oron, Z. Jiang, and R. Gentleman. Gene set enrichment analysis using linear models and diagnostics. *Bioinformatics*, 24(22):2586–91, 2008.

A. Oshlack and M. J. Wakefield. Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct*, 4:14, 2009.

A. M. Parissenti, S. L. Hembruff, D. J. Villeneuve, Z. Veitch, B. Guo, and J. Eng. Gene expression profiles as biomarkers for the prediction of chemotherapy drug response in human tumour cells. *Anti-cancer drugs*, 18 (5):499–523, 2007.

P. J. Park, Y. A. Cao, S. Y. Lee, J.-W. Kim, M. S. Chang, R. Hart, and S. Choi. Current issues for DNA microarrays: platform comparison, double linear amplification, and universal RNA reference. *Journal of Biotechnology*, 112 (3):225–45, 2004.

H. Parkinson, M. Kapushesky, M. Shojatalab, N. Abeygunawardena, R. Coulson, A. Farne, E. Holloway, N. Kolesnykov, P. Lilja, M. Lukk, R. Mani, T. Rayner, A. Sharma, E. William, U. Sarkans, and A. Brazma. ArrayExpress–a public database of microarray experiments and gene expression profiles. *Nucleic Acids Research*, 35:D747–D750, 2007.

H. Parkinson, M. Kapushesky, N. Kolesnikov, G. Rustici, M. Shojatalab, N. Abeygunawardena, H. Berube, M. Dylag, I. Emam, A. Farne, E. Holloway, M. Lukk, J. Malone, R. Mani, E. Pilicheva, T. F. Rayner, F. Rezwan, A. Sharma, E. Williams, X. Z. Bradley, T. Adamusiak, M. Brandizi, T. Burdett, R. Coulson, M. Krestyaninova, P. Kurnosov, E. Maguire, S. G. Neogi, P. Rocca-Serra, S.-A. Sansone, N. Sklyar, M. Zhao, U. Sarkans, and A. Brazma. ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Research*, 37(Database issue):D868–72, 2009.

H.-O. Peitgen and D. Saupe. *The Science of Fractal Images*. Springer, New York, NY, USA, 1988.

J. Peltonen, H. Aidos, and S. Kaski. Supervised nonlinear dimensionality reduction by Neighbor Retrieval. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1809–1812. IEEE, Apr. 2009.

J. Peltonen, H. Aidos, N. Gehlenborg, A. Brazma, and S. Kaski. An information retrieval perspective on visualization of gene expression data with ontological annotation. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2178–2181. IEEE, Mar. 2010.

H. Peng and W. Hagopian. Environmental factors in the development of Type 1 diabetes. *Reviews in Endocrine & Metabolic Disorders*, 7(3):149–62, 2006.

P. Pirolli, S. K. Card, and M. M. Van Der Wege. The effect of information scent on searching information. In *Proceedings of the Working Conference on Advanced Visual Interfaces - AVI '00*, pages 161–172, New York, NY, USA, 2000. ACM Press.

J. Quackenbush. Computational analysis of microarray data. *Nature Reviews Genetics*, 2(6):418–427, 2001.

J. Quackenbush. Microarrays - guilt by association. *Science*, 302(5643): 240–241, 2003.

J. Quackenbush. Microarray analysis and tumor classification. *N Engl J Med*, 354(23):2463–2472, 2006.

L. Rahib, N. K. MacLennan, S. Horvath, J. C. Liao, and K. M. Dipple. Glycerol kinase deficiency alters expression of genes involved in lipid metabolism, carbohydrate metabolism, and insulin signaling. *European journal of Human Genetics*, 15(6):646–57, 2007.

T. Ranheim, M. Mattingsdal, J. M. Lindvall, O. L. Holla, K. E. Berge, M. A. Kulseth, and T. P. Leren. Genome-wide expression analysis of cells expressing gain of function mutant D374Y-PCSK9. *Journal of Cellular Physiology*, 217(2):459–67, 2008.

T. F. Rayner, P. Rocca-Serra, P. T. Spellman, H. C. Causton, A. Farne, E. Holloway, R. A. Irizarry, J. Liu, D. S. Maier, M. Miller, K. Petersen, J. Quackenbush, G. Sherlock, C. J. Stoeckert, J. White, P. L. Whetzel,

F. Wymore, H. Parkinson, U. Sarkans, C. A. Ball, and A. Brazma. A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics*, 7:489, 2006.

M. Reich, T. Liefeld, J. Gould, J. Lerner, P. Tamayo, and J. P. Mesirov. GenePattern 2.0. *Nature Genetics*, 38(5):500–1, 2006.

D. R. Rhodes, S. Kalyana-Sundaram, V. Mahavisno, R. Varambally, J. Yu, B. B. Briggs, T. R. Barrette, M. J. Anstet, C. Kincead-Beal, P. Kulkarni, S. Varambally, D. Ghosh, and A. M. Chinnaiyan. Oncomine 3.0: Genes, Pathways, and Networks in a Collection of 18,000 Cancer Gene Expression Profiles. *Neoplasia*, 9:166, 2007.

G. G. Robertson, S. K. Card, and J. D. Mackinlay. Information visualization using 3D interactive animation. *Communications of the ACM*, 36(4):57–71, 1993.

S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–6, 2000.

N. Salomonis, K. Hanspers, A. C. Zambon, K. Vranizan, S. C. Lawlor, K. D. Dahlquist, S. W. Doniger, J. Stuart, B. R. Conklin, and A. R. Pico. GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics*, 8:217, 2007.

P. Saraiya, C. North, and K. Duca. An Evaluation of Microarray Visualization Tools for Biological Insight. In *IEEE Symposium on Information Visualization*, pages 1–8. IEEE, 2004.

D. Sean and P. S. Meltzer. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, 23(14):1846–7, 2007.

E. Segal, N. Friedman, D. Koller, and A. Regev. A module map showing conditional activity of expression modules in cancer. *Nature genetics*, 36 (10):1090–8, Oct. 2004.

A. J. Shatkin. Capping of eucaryotic mRNAs. *Cell*, 9(4 PT 2):645–53, 1976.

L. Shi, L. H. Reid, W. D. Jones, R. Shippy, J. A. Warrington, S. C. Baker, P. J. Collins, F. De Longueville, E. S. Kawasaki, K. Y. Lee, Y. Luo, Y. A. Sun, J. C. Willey, R. A. Setterquist, G. M. Fischer, W. Tong, Y. P. Dragan, D. J. Dix, F. W. Frueh, F. M. Goodsaid, D. Herman, R. V. Jensen, C. D. Johnson, E. K. Lobenhofer, R. K. Puri, U. Schrf, J. Thierry-Mieg, C. Wang, M. Wilson, P. K. Wolber, L. Zhang, S. Amur, W. Bao, C. C. Barbacioru, A. B. Lucas, V. Bertholet, C. Boysen, B. Bromley, D. Brown, A. Brunner, R. Canales, X. M. Cao, T. A. Cebula, J. J. Chen, J. Cheng, T.-M. Chu, E. Chudin, J. Corson, J. C. Corton, L. J. Croner, C. Davies, T. S. Davison, G. Delenstarr, X. Deng, D. Dorris, A. C. Eklund, X.-H. Fan, H. Fang, S. Fulmer-Smentek, J. C. Fuscoe, K. Gallagher, W. Ge, L. Guo, X. Guo, J. Hager, P. K. Haje, J. Han, T. Han, H. C. Harbottle, S. C. Harris, E. Hatchwell, C. A. Hauser, S. Hester, H. Hong, P. Hurban, S. A. Jackson, H. Ji, C. R. Knight, W. P. Kuo, J. E. Leclerc, S. Levy, Q.-Z. Li, C. Liu, Y. Liu, M. J. Lombardi, Y. Ma, S. R. Magnuson, B. Maqsodi, T. Mcdaniel, N. Mei, O. Myklebost, B. Ning, N. Novoradovskaya, M. S. Orr, T. W. Osborn, A. Papallo, T. A. Patterson, R. G. Perkins, E. H. Peters, R. Peterson, K. L. Philips, P. S. Pine, L. Pusztai, F. Qian, H. Ren, M. Rosen, B. A. Rosenzweig, R. R. Samaha, M. Schena, G. P. Schroth, S. Shchegrova, D. D. Smith, F. Staedtler, Z. Su, H. Sun, Z. Szallasi, Z. Tezak, D. Thierry-Mieg, K. L. Thompson, I. Tikhonova, Y. Turpaz, B. Vallanat, C. Van, S. J. Walker, S. J. Wang, Y. Wang, R. Wolfinger, A. Wong, J. Wu, C. Xiao, Q. Xie, J. Xu, W. Yang, L. Zhang, S. Zhong, Y. Zong, and W. Slikker. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24 (9):1151–61, 2006.

B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343. IEEE Computer Society Press, 1996.

J. Siegel, E. Farrell, R. Goldwyn, and H. Friedman. The surgical implication

of physiologic patterns in myocardial infarction shock. *Surgery*, 72:126–141, 1972.

D. Slonim. From patterns to pathways: gene expression data analysis comes of age. *Nature Genetics*, 32:502–508, 2002.

E. M. Southern. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology*, 98(3):503–17, 1975.

K. Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972.

C. Spearman. The Proof and Measurement of Association between Two Things. *American Journal of Psychology*, 15(1):72–101, 1904.

P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, 9(12):3273–3297, 1998.

P. T. Spellman, M. Miller, J. Stewart, C. Troup, U. Sarkans, S. Chervitz, D. Bernhart, G. Sherlock, C. Ball, M. Lepage, M. Swiatek, W. L. Marks, J. Goncalves, S. Markel, D. Iordan, M. Shojatalab, A. Pizarro, J. White, R. Hubley, E. Deutsch, M. Senger, B. J. Aronow, A. Robinson, D. Bassett, C. J. Stoeckert, and A. Brazma. Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biology*, 3(9): RESEARCH0046, 2002.

I. Spence and S. Lewandowsky. Displaying proportions and percentages. *Applied Cognitive Psychology*, 5(1):61–77, 1991.

L. K. Stewart, J. L. Soileau, D. Ribnicky, Z. Q. Wang, I. Raskin, A. Poulev, M. Majewski, W. T. Cefalu, and T. W. Gettys. Quercetin transiently increases energy expenditure but persistently decreases circulating markers of inflammation in C57BL/6J mice fed a high-fat diet. *Metabolism*, 57(7 Suppl 1):S39–46, 2008.

J. D. Storey. The positive false discovery rate: a Bayesian interpretation and the q -value. *The Annals of Statistics*, 31(6):2013–2035, 2003.

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

A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–50, 2005.

S. Suthram, J. T. Dudley, A. P. Chiang, R. Chen, T. J. Hastie, and A. J. Butte. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS Computational Biology*, 6(2):e1000662, 2010.

D. Swarbreck, C. Wilks, P. Lamesch, T. Z. Berardini, M. Garcia-Hernandez, H. Foerster, D. Li, T. Meyer, R. Muller, L. Ploetz, A. Radenbaugh, S. Singh, V. Swing, C. Tissier, P. Zhang, and E. Huala. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Research*, 36(Database issue):D1009–14, 2008.

A. Sweet-Cordero, S. Mukherjee, A. Subramanian, H. You, J. J. Roix, C. Ladd-Acosta, J. Mesirov, T. R. Golub, and T. Jacks. An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. *Nature Genetics*, 37(1):48–55, 2005.

D. Talantov, A. Mazumder, J. X. Yu, T. Briggs, Y. Jiang, J. Backus, D. Atkins, and Y. Wang. Novel genes associated with malignant melanoma but not benign melanocytic lesions. *Clinical Cancer Research*, 11(20):7234–42, 2005.

A. Tanay, I. Steinfeld, M. Kupiec, and R. Shamir. Integrative analysis of genome-wide experiments in the context of a large high-throughput data compendium. *Molecular Systems Biology*, 1:2005.0002, 2005.

J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–23, 2000.

J. J. Thomas and K. A. Cook. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visual Analytics Center & IEEE, Richland, WA, 2005.

M. J. Tovée. *Introduction to the Visual System*. Cambridge University Press, Cambridge, UK, 1996.

J. P. Townsend. Multifactorial experimental design and the transitivity of ratios with spotted DNA microarrays. *BMC Genomics*, 4(1), 2003.

O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6):520–5, 2001.

Y. Tu and H.-W. Shen. Visualizing changes of hierarchical data using treemaps. *IEEE Transactions on Visualization and Computer Graphics*, 13(6): 1286–93, 2007.

J. W. Tukey. *Exploratory Data Analysis*. Addison Wesley, Reading, MA, USA, 1977.

A. Tutt and A. Ashworth. The relationship between the roles of BRCA genes in DNA repair and cancer predisposition. *Trends in Molecular Medicine*, 8(12):571–6, Dec. 2002.

H. D. Van Guilder, K. E. Vrana, and W. M. Freeman. Twenty-five years of quantitative PCR for gene expression analysis. *BioTechniques*, 44(5):619–26, 2008.

I. Vastrik, P. D'Eustachio, E. Schmidt, G. Joshi-Tope, G. Gopinath, D. Croft, B. de Bono, M. Gillespie, B. Jassal, S. Lewis, L. Matthews, G. Wu, E. Birney, and L. Stein. Reactome: a knowledge base of biologic pathways and processes. *Genome Biology*, 8(3):R39, 2007.

V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler. Serial analysis of gene expression. *Science*, 270(5235):484–7, 1995.

J. Venna and S. Kaski. Nonlinear Dimensionality Reduction as Information Retrieval. *Proceedings of AISTATS 2007*, 2007a.

J. Venna and S. Kaski. Comparison of visualization methods for an atlas of gene expression data sets. *Information Visualization*, 6(2):139–154, 2007b.

A. J. Vilella, J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, and E. Birney. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research*, 19(2):327–35, 2009.

R. Vliegen, J. J. van Wijk, and E.-J. van Der Linden. Visualizing business data with generalized treemaps. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):789–96, 2006.

C. T. Walsh, S. Garneau-Tsodikova, and G. J. Gatto. Protein posttranslational modifications: the chemistry of proteome diversifications. *Angewandte Chemie*, 44(45):7342–72, 2005.

F. Wang, M. Herrington, J. Larsson, and J. Permert. The relationship between diabetes and pancreatic cancer. *Molecular Cancer*, 2:4, 2003.

Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

M. O. Ward. A taxonomy of glyph placement strategies for multidimensional data visualization. *Information Visualization*, 1(3-4):194–210, 2002.

C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann, San Francisco, 2nd edition, 2004.

M. Wattenberg. A note on space-filling visualizations and space-filling curves. *IEEE Symposium on Information Visualization*, pages 181–186, 2005.

R. Weinberg. Point: Hypotheses first. *Nature*, 464(7289):678, Apr. 2010.

J. N. Weinstein. An Information-Intensive Approach to the Molecular Pharmacology of Cancer. *Science*, 275(5298):343–349, 1997.

J. N. Weinstein. A postgenomic visual icon. *Science*, 319(5871):1772–3, 2008.

K. P. White. Functional genomics and the study of development, variation and evolution. *Nature Reviews Genetics*, 2(7):528–537, 2001.

L. Wilkinson and M. Friendly. The History of the Cluster Heat Map. *The American Statistician*, 63(2):179–184, 2009.

B. Wong. Points of view: Color coding. *Nature Methods*, 7(8):573–573, 2010.

J. Yang, A. Patro, S. Huang, N. Mehta, M. Ward, and E. Rundensteiner. Value and Relation Display for Interactive Exploration of High Dimensional Datasets. In *IEEE Symposium on Information Visualization 2004 (INFOVIS 2004)*, pages 73–80. IEEE, 2004.

J. Yang, D. Hubball, M. O. Ward, E. A. Rundensteiner, and W. Ribarsky. Value and relation display: interactive visual exploration of large data sets with hundreds of dimensions. *IEEE Transactions on Visualization and Computer Graphics*, 13(3):494–507, 2007.

Y. H. Yang and T. Speed. Design issues for cDNA microarray experiments. *Nature Reviews Genetics*, 3(8):579–88, 2002.

S. Q. Ye, T. Lavoie, D. C. Usher, and L. Q. Zhang. Microarray, SAGE and their applications to cardiovascular diseases. *Cell Research*, 12(2):105–15, June 2002.

Y. Zhu, S. Davis, R. Stephens, P. S. Meltzer, and Y. Chen. GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus. *Bioinformatics*, 24(23):2798–800, 2008.