

**Analysing functional genomics data using novel ensemble,  
consensus and data fusion techniques**

by  
**Enrico Glaab**

Thesis submitted to The University of Nottingham  
for the Degree of Doctor of Philosophy

School of Computer Science and Information Technology  
The University of Nottingham  
Nottingham, United Kingdom

March 2011

# Abstract

**Motivation:** A rapid technological development in the biosciences and in computer science has made it possible to analyse high-dimensional biological datasets, containing measurements for entire genomes, transcriptomes and proteomes, on standard desktop computers. New experimental and computational technologies provide ample opportunities to improve our basic understanding of biological systems and develop better methods for the monitoring, diagnosis and treatment of genetic diseases. However, common properties of the new high-throughput experimental data, like small sample sizes in relation to the number of features, high noise levels and outliers, also pose novel challenges.

**Goal:** Ensemble and consensus machine learning techniques and data integration methods can address some of these issues, but they often provide very complex models of the system of interest, which lack interpretability and overfit the data. The goal behind this thesis was therefore to develop new approaches to combine algorithms and large-scale biological datasets, including novel approaches to integrate analysis types from different domains (e.g. statistics, topological network analysis, machine learning and text mining), which are designed not only to exploit the diverse information content in the data sources and the strength of different algorithms, but to provide compact and interpretable models, which enable the extraction of new biological knowledge.

**Approach:** As the main contribution of this thesis, a novel framework and software collection for integrative analysis of gene expression data, gene/protein sets, cellular pathway and protein interaction data was developed and applied to real-world biological datasets from collaborating institutions, focussing on problems in cancer biology.

The framework takes advantage of cross-study normalisation and cross-domain data fusion methods, and enables both the comparison and modular combination of algorithms for different statistical learning tasks (feature selection, classification and clustering). Ensemble and consensus analysis techniques employed for this purpose are re-designed such that the model generation does not only seek to maximise predictive accuracy and model robustness, but also to create compact and interpretable models. More importantly, novel integrative analysis techniques have been developed, which combine algorithmic techniques from different domains (machine learning, network topological analysis, literature mining and optimisation) and use information from multiple data sources (gene expression data, protein interactions, cellular pathway

---

definitions and gene/protein sets).

**Main results:** The key deliverables of the doctoral project are new ensemble, consensus and cross-domain bioinformatics algorithms, and new analysis pipelines combining these and classical data mining techniques within a general framework. This framework contains methods for the integrative analysis of both large-scale gene and protein expression data (including the tools *ArrayMining*, *Top-scoring pathway pairs* and *RNAnalyze*) and general gene and protein sets (including the tools *TopoGSA*, *EnrichNet* and *PathExpand*).

Among the biological findings obtained with these new software tools, a central result was the identification of a novel tumour marker gene in collaboration with the Nottingham Queens Medical Centre, facilitating the distinction between two clinically important breast cancer subtypes (*framework tool: ArrayMining*). Other biomedically relevant findings resulted from a co-operation with the Spanish National Cancer Centre, predicting novel candidate disease genes for Alzheimer's disease and pancreatic cancer using an integrative analysis combining cellular pathway definitions and protein interaction data (*framework tool: PathExpand*). Moreover, associations between disease-related processes, including the verification of functional associations between prostate cancer development and different cellular processes, were identified using a new rule-based classification method integrating gene expression and cellular pathway data (*framework tool: Top-scoring pathway pairs*).

Apart from these results obtained from data fusion techniques, new insights were also gained from the combination of diverse analysis techniques, as illustrated by a combined microarray gene selection and network topological analysis, which identified genes that are differentially expressed in different cancers and have outstanding topological properties when being mapped to a molecular interaction network (*framework tool: TopoGSA*). Finally, new techniques for interactive visualisation and exploration of functional associations in biological data facilitated the interpretation of different real-world datasets, with successful applications in agriculture (analysis of gene regulation in a plant model organism) and biomedicine (analysis of cancer gene expression data; *framework tool: VRMLGen*).

# Acknowledgements

I would like to thank my supervisors, Prof. Natalio Krasnogor and Dr. Jonathan Garibaldi, for the opportunity to pursue my doctoral studies at Nottingham University and for all the support and advice I received during the last three years.

Prof. Krasnogor has helped me to gain the scientific skills that have been essential for my work, provided numerous valuable comments and insightful discussions, and organised external placements for me at two prestigious research institutes, the Spanish National Cancer Centre and the Weizmann Institute of Science. Dr. Garibaldi organised and managed all administrative matters with great dependability and efficiency, helped to secure additional funding, and made the necessary arrangements that allowed me to participate in cross-national fellowship meetings and workshops. In short, the training and opportunities I received will help me beyond the completion of my thesis.

Above all, I am grateful for the support of my family, whose understanding, advice and encouragement has been invaluable for me, not only during the course of my studies. My parents and sister have shown much patience with me and my research, and I hope that we will be able to see each other more often in the future.

Many special thanks go to my current and previous office mates Dr. German Terrazas Angulo, Dr. Pawel Widera, Dr. Jaume Bacardit, Dr. James Smaldon, Dr. Azhar Shah, Dr. Leong Ting Lui and Jack Chaplin, and to all fellow PhD students and research colleagues in the ASAP group for many inspiring discussions and making the School of Computer Science a great place to work at.

I am also grateful for the advice and suggestions I received from collaborating scientists during several research projects, in particular Dr. Anaïs Baudot who has supervised my work at the CNIO in Madrid with remarkable attention and competence, my placement host Prof. Alfonso Valencia, for his invaluable expert advice, and the whole Structural Computational Biology Group at the CNIO for a great time in Madrid. Similarly, I thank Prof. Doron Lancet's group at the Weizmann Institute of Science for their kind hospitality while hosting me in their lab and for an exciting and memorable stay in Israel.

I am also indebted to our collaborating researchers at Nottingham University, namely Dr. George Bassel and Dr. Michael Holdsworth from the Division of Plant and Crop Sciences, the Department of Histopathology

at the Queens Medical Centre, Dr. Ali Mobasheri from the School of Veterinary Medicine, and Dr Maria Toledo-Rodriguez from the Institute of Neuroscience.

Finally, I would like to thank the organisations that have provided financial support through research fellowships, in particular the Marie-Curie Early Stage-Training programme (grant MEST-CT-2004-007597) and the Bridging the Gaps initiative.

# Contents

<b>Abstract</b>	<b>ii</b>
<b>Acknowledgements</b>	<b>iv</b>
<b>Contents</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background and Motivation . . . . .	1
1.2 Aims and Scope . . . . .	4
1.3 Thesis Organisation . . . . .	6
1.4 Methodology . . . . .	7
1.5 Main Results . . . . .	8
<b>2 Biological Background and Fields of Application</b>	<b>15</b>
2.1 Genetic Disorders and Cancer Diseases . . . . .	16
2.2 Biomarker Discovery and Outcome Prognosis . . . . .	18
2.3 Role of Systems Biology in Elucidating the Basis of Complex Diseases . . . . .	20
<b>3 Literature Review</b>	<b>22</b>
3.1 Low Level Analysis: Pre-Processing, Normalisation and Quality Checking . . . . .	22
3.2 Higher Level Analysis - Introduction . . . . .	34
3.3 Dimensionality reduction and feature selection . . . . .	35
3.4 Class Discovery (Unsupervised Machine Learning) . . . . .	48
3.5 Class Prediction (Supervised Machine Learning) . . . . .	59
3.6 Data Integration 1: Cross-Study Analysis . . . . .	71
3.7 Data Integration 2: Integrating Cellular Pathway Data . . . . .	73
3.8 Data Integration 3: Integrating Molecular Interaction Data . . . . .	75
<b>4 Comparison of Standard Machine Learning Techniques and Integrative Extensions</b>	<b>78</b>
4.1 Comparative Evaluation of Feature Selection Methods . . . . .	82
4.2 Comparative Evaluation of Classification Methods . . . . .	85

---

4.3	Comparative Evaluation of Clustering Methods . . . . .	99
<b>5</b>	<b>Integrative Framework for Gene/Protein Expression Data Analysis (ArrayMining)</b>	<b>112</b>
5.1	Automatic Data Processing and Gene/Protein Name Normalisation . . . . .	115
5.2	Ensemble and Consensus Analysis of Microarray Data . . . . .	117
5.3	Specialised Analysis Methods for Microarray Data . . . . .	122
5.4	Modular Combination of Analysis Techniques / Novel Analysis Pipelines . . . . .	125
<b>6</b>	<b>Integrative Analysis of Gene/Protein Sets</b>	<b>127</b>
6.1	Network Topological Analysis of Gene/Protein Sets (TopoGSA) . . . . .	128
6.2	Integrative Enrichment Analysis of Gene and Protein Sets (EnrichNet) . . . . .	132
6.3	Integrative Extension of Cellular Pathway Definitions (PathExpand) . . . . .	137
<b>7</b>	<b>Simplifying Classification Rules to Enhance Model Interpretability</b>	<b>149</b>
7.1	Integrative Rule Learning for High-Dimensional Biological Data (TSPP) . . . . .	151
7.2	Evaluation of Integrative Rule Learning on Microarray Data . . . . .	155
<b>8</b>	<b>Visualisation of High-Dimensional Biological Data (VRMLGen)</b>	<b>161</b>
<b>9</b>	<b>Main Biological Contributions</b>	<b>167</b>
<b>10</b>	<b>Conclusions</b>	<b>174</b>
10.1	General Summary and Discussion . . . . .	174
10.2	Outlook on Future Work . . . . .	178
<b>11</b>	<b>Appendix</b>	<b>198</b>
11.1	Glossary . . . . .	198
11.2	Example flowcharts for new integrative analysis pipelines . . . . .	203

# List of Figures

1.1	Overview of data and analysis types . . . . .	3
1.2	Modules and workflow of the data analysis framework . . . . .	9
2.1	Example inheritance pedigree . . . . .	17
2.2	The drug development pipeline . . . . .	19
3.1	Higher level analysis - overview . . . . .	23
3.2	Microarray data analysis workflow . . . . .	24
3.3	Example MA-plot . . . . .	26
3.4	Overview of generic feature selection methodologies . . . . .	36
3.5	Bias/variance trade-off in statistical learning . . . . .	38
3.6	Example Markov blanket . . . . .	42
3.7	Overview of unsupervised learning methods . . . . .	49
3.8	Example heat map . . . . .	53
3.9	Overview of supervised learning methods . . . . .	60
4.1	Comparative algorithm evaluation - Experimental procedure . . . . .	79
4.2	B-cell lymphoma - 3D PCA visualisation . . . . .	81
4.3	Prostate cancer - 3D PCA visualisation . . . . .	81
4.4	Breast cancer - 3D PCA visualisation . . . . .	82
4.5	Example for a BioHEL classification rule set . . . . .	86
4.6	Prostate cancer - Rule-based sample assignment intervals . . . . .	95
4.7	B-cell lymphoma - Rule-based sample assignment intervals . . . . .	96
4.8	Visualisation of clustering results - Calinski-Harabasz index . . . . .	102
4.9	Visualisation of clustering results - kNN-Connectivity index . . . . .	105
4.10	Visualisation of clustering results - Silhouette plot . . . . .	106
4.11	Visualisation of clustering results - PCA plot . . . . .	106
4.12	Histogram of adjusted rand indices - VF dimensionality reduction . . . . .	107
4.13	Histogram of adjusted rand indices - GSA dimensionality reduction . . . . .	108
4.14	Principal Component Analysis - Breast cancer dataset . . . . .	109



---

4.15	Independent Component Analysis - Breast cancer dataset . . . . .	109
4.16	Isomap Analysis - Breast cancer dataset . . . . .	109
4.17	Locally Linear Embedding - Breast cancer dataset . . . . .	110
5.1	ArrayMining: The main component of the integrative analysis framework . . . . .	113
5.2	ArrayMining - main interface . . . . .	118
5.3	ArrayMining: Example box plot visualisation . . . . .	119
5.4	ArrayMining: Example heat map visualisation . . . . .	119
5.5	VRMLGen: Example 3D-visualisation of an Independent Component Analysis . . . . .	121
5.6	Breast cancer gene co-expression network analysis . . . . .	124
5.7	Cross-study analysis - density plots . . . . .	125
6.1	TopoGSA, EnrichNet and PathExpand . . . . .	128
6.2	TopoGSA: Example cancer gene set analysis . . . . .	131
6.3	EnrichNet: Example Results (BioCarta REIA pathway) . . . . .	134
6.4	EnrichNet: Network visualisation (Sarcoma and Bladder cancer mutated genes) . . . . .	135
6.5	PathExpand: Visualisation of graph-based filtering criteria . . . . .	140
6.6	PathExpand: Semantic similarity analysis of protein annotations . . . . .	142
6.7	PathExpand: Extended Alzheimer's pathway . . . . .	143
6.8	PathExpand: Crosstalk between interleukin signalling pathways . . . . .	145
6.9	PathExpand: Cell cycle G1/S check point sub-network . . . . .	146
7.1	Top-scoring pathway pair method . . . . .	150
7.2	TSPP: Overview of the workflow . . . . .	153
7.3	TSPP: Analysing differentially regulated pathway-pairs in a protein interaction network . . . . .	159
8.1	VRMLGen: Main functions and features . . . . .	163
8.2	VRMLGen: Example visualisation of breast cancer microarray data . . . . .	165
8.3	VRMLGen: Example code for breast cancer data analysis . . . . .	165
9.1	New breast cancer marker gene RERG - Gene expression box plot . . . . .	169
9.2	SeedNet: Gene co-expression network visualisation . . . . .	172
11.1	ArrayMining - Example flowchart . . . . .	203
11.2	ArrayMining/TopoGSA - Example flowchart . . . . .	203
11.3	PathExpand/EnrichNet/TSPP - Example flowchart . . . . .	203

# List of Tables

2.1	Genetic disorders and inheritance types . . . . .	17
2.2	Application of bioinformatics tools in biomedicine . . . . .	19
4.1	Datasets used for comparative evaluation . . . . .	80
4.2	Comparison of feature selection methods . . . . .	85
4.3	Parameters used for BioHEL . . . . .	88
4.4	10-fold CV classification results . . . . .	89
4.5	LOOCV classification results . . . . .	90
4.6	Comparison of prediction results from the literature (DLBCL dataset) . . . . .	91
4.7	Comparison of prediction results from the literature (Prostate cancer dataset) . . . . .	92
4.8	Comparison of prediction methods (Friedman test) . . . . .	92
4.9	List of high scoring genes (Prostate cancer dataset) . . . . .	94
4.10	List of high scoring genes (DLBCL dataset) . . . . .	94
4.11	List of high scoring genes (Breast cancer dataset) . . . . .	97
4.12	Comparison of clustering methods (classical standardisation) . . . . .	103
4.13	Comparison of clustering methods (robust standardisation) . . . . .	104
6.1	PathExpand: Statistics on added proteins . . . . .	141
6.2	PathExpand: Topological properties of extended pathways . . . . .	142
6.3	PathExpand: Cellular processes enriched in pancreatic mutated genes . . . . .	148
7.1	TSPP: Leave-one-out cross-validation results (KEGG database) . . . . .	155
7.2	TSPP: Leave-one-out cross-validation results (GO database) . . . . .	156
7.3	TSPP: Leave-one-out cross-validation results (alternative method) . . . . .	156
7.4	TSPP: Top-ranked pathway pairs (Prostate cancer data) . . . . .	158
7.5	TSPP: Top-ranked pathway pairs (B-cell lymphoma data) . . . . .	158

# Chapter 1

## Introduction

### Chapter abstract

This introductory chapter will provide a bird's eye view of the goals behind the thesis and the data sources and methods used to achieve them. It will delineate the scope of the project and discuss in general terms how the work compares to and departs from previous biological data analysis and integration approaches. Finally, it will guide the reader through the different sections of the thesis and present a summary of the main results.

### 1.1 Background and Motivation

The spread of high-throughput technologies in the biosciences in recent years, including high-throughput sequencing methods, DNA and protein microarrays, has led to an exponential increase of public biological databases. The large amount of freely available data has raised hopes that researchers will in the long run be able to obtain a more holistic understanding of the molecular mechanisms in living cells by analysing complete gene-, protein- and metabolite networks, instead of considering their individual components separately. However, although the data from high-throughput experiments offers new opportunities for the biosciences, at the same time a multitude of new challenges arise from its typical characteristics: Large numbers of features in relation to small numbers of samples pose several problems in the statistical analysis, which have been extensively discussed in the literature under the headings “curse of dimensionality” [1], “multiple testing” [2,3] and “feature redundancy and dependence” [4,5]. Moreover, with regard to a specific biological question of interest, the majority of features in a dataset might be irrelevant, hence, extracting only the informative sub-structures can be akin to finding the proverbial needle in a haystack. Additionally, single measurements within high-throughput experimental methods are often affected by different types of noise, providing the experimenter only with scaled and shifted versions of the original signals and with outliers both among the samples and features. Apart from these problems affecting the statistical analysis and evaluation, various computational difficulties arise commonly, spanning from general issues concerning runtime complexity and memory management to data access efficiency problems in database and web-server applications.

In summary, the most prevalent problems and research questions that have emerged in the field of high-dimensional biological data analysis and which are addressed in this PhD are the following:

- *“Curse of dimensionality” problem*: How can robust clustering or supervised prediction results be attained for small sample-size datasets in which the number of features is by two orders of magnitude larger than the number of samples (in particular, in the case of an approximately uniform data distribution [6])?
- *Noise problem*: How can genes, proteins or metabolites that are significantly differentially regulated across different biological conditions be reliably identified, if the expression values of a large proportion of genes/proteins/metabolites are masked or dominated by noise (with both technical and biological sources of noise)?
- *Multiple testing problem*: How can spurious rejections of a null-hypothesis be avoided effectively by reducing the dimensionality of the input data or adjusting hypothesis tests to account for repeated hypothesis testing?
- *Evaluation problem*: Given that evaluation methods to estimate the generalization error of machine learning models tend to have limitations either in accurately estimating the variance or the bias on microarray datasets with small sample size [7], how should a reliable validation pipeline be built?
- *Methodological problem and “no-free-lunch” problem*: Do any “methods of choice” exist for microarray data analysis, or which algorithms should be compared or combined to solve a specific analysis problem efficiently and effectively? If several methods have been shown to have different strengths and weaknesses on different datasets (e.g. datasets for different cancer types), how can a robust analysis system be built, attaining a high performance across many diverse datasets?

In order to address these statistical and computational challenges in the analysis of high-dimensional data, several new algorithms and data structures, tailored to specific analysis problems and experimental platforms, have been developed in recent years. However, using methods optimised for a single data source type does often not suffice to exploit the information content of multiple available datasets from diverse platforms and biological domains, or to reach the model accuracy and significance that might be obtained from combining the benefits of multiple search methodologies, scoring functions or data structures.

In many areas of computer science and biology, *integrative analysis methods*, which combine different data sets and/or algorithms, have not only been shown to effectively increase robustness and accuracy of an analysis but are often essential requirements to verify a given biological hypothesis, to obtain sufficiently robust prediction models or to solve a computational problem in a given time-frame. For example, in several applications of statistics and machine learning, ensemble and consensus approaches, which exploit the synergies of diverse algorithms for the same problem type, provide significant improvements in terms of robustness and accuracy on large-scale datasets with high noise levels [8–13]. Similarly, at the data collection and pre-processing level, cross-study normalisation and data fusion techniques have been employed successfully to obtain more stable prediction models or clustering results [14–16]. However, especially the application of ensemble techniques, and in some cases also data integration methods, often tends to generate very complex biological models, lacking interpretability and sometimes even overfitting the data. Cross-platform normalisation methods often result in a significant loss of information due to the normalisation process, and the quality of the outcome highly depends on the size of the intersection set between the features of the considered input datasets [17]. Thus, the above list of problems to be addressed in high-dimensional biological data analysis has to be extended by the following challenges in *integrative data analysis*:

- *Cross-platform data integration problem*: How can data sets obtained with different experimental platforms (for the same cell types and phenotypes) be combined, if the overlap in the measured

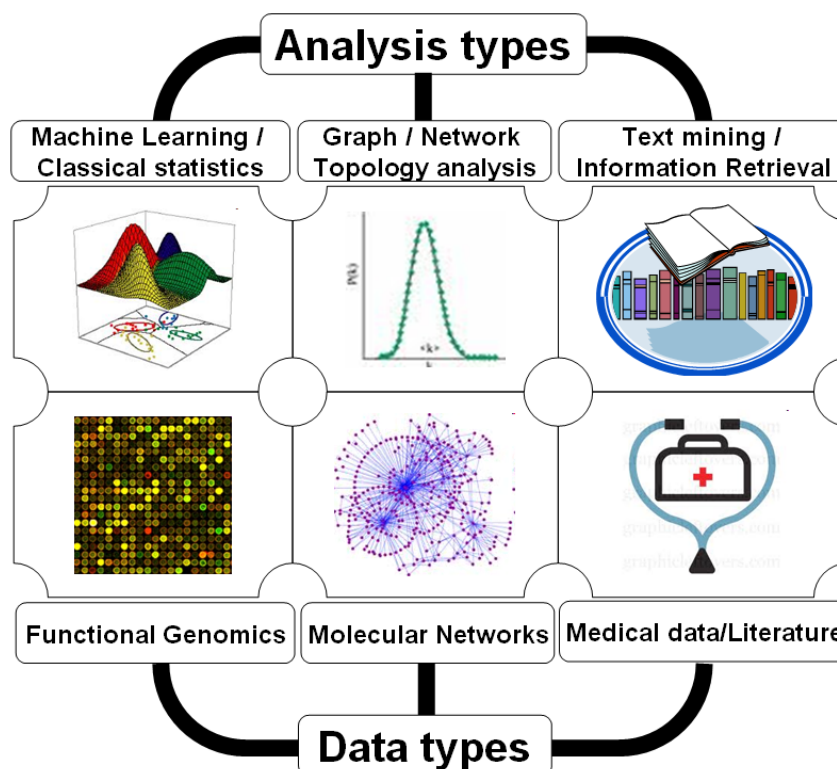


Figure 1.1: Overview of data and analysis types combined as part of integrative methods developed in this thesis to obtain a better understanding of biological processes of interest

features (e.g. genetic probes on DNA chips) is small and the experiments are affected by different systematic biases?

- *General data integration problem:* How can gene/protein expression data and other biological data sources, e.g. molecular interaction, genomic, epigenetic and metabolic data, be combined to obtain new insights or improve accuracy and robustness, while retaining a high level of interpretability?

These challenges and the previously mentioned problems in large-scale biological data analysis provide the primary motivation for this thesis to investigate the potential of new integrative analysis techniques to obtain improvements in terms of classical performance criteria (e.g. prediction accuracy, adjusted rand index in clustering, model robustness), but at the same time also in terms of model interpretability and biological insights gained. Moreover, as the main contribution of this thesis, new cross-domain analysis methods will be presented, which combine both diverse data sources (gene/protein expression data, gene/protein set data, protein interaction data and cellular pathway data) and analysis types (statistics, network analysis, machine learning, literature mining). These novel integrative analysis techniques have been presented in dedicated publications and are discussed in separate chapters of this thesis. Figure 1.1 provides a general overview of the data and algorithm types considered as part of integrative biological analysis methods in this thesis.

Molecular networks are shown as the central data type in this figure, because their common representation as graphs with nodes (corresponding to molecules) and edges (corresponding to molecular interactions) provides a suitable data structure to map other data sources onto a single graph-based model (with nodes corresponding to molecules and edges to associations). The other input data sources considered here, consisting of large-scale functional genomics data (e.g. gene/protein expression datasets), and clinical records and literature data, are typically only analysed by highly specialised algorithms. However, synergistic ef-

fects cannot only be attained by investigating multiple input data sources independently, with different dedicated analysis methods, and only interpreting their outputs together, but also by combining the analysis techniques directly, either in a modular fashion or by integrating them into new algorithms operating on a unified data structure for all biological inputs. Achieving these synergistic effects, while at the same time retaining a high level of model interpretability and facilitating the extraction of new biological knowledge, is the primary goal for this thesis. Details on the more specific objectives and the novel approaches to achieve them will be provided in the following sections.

## 1.2 Aims and Scope

The current limitations in the statistical power and interpretability of computational models built from high-dimensional biological data (see Background and Motivation section), motivate the main objective behind the research conducted for this thesis: Improving the statistical analysis of these datasets in terms of accuracy, robustness and interpretability by developing new integrative techniques to exploit the synergies of diverse data sources, algorithms and data structures.

The methods developed for this purpose will mainly be applied to the study of microarray gene expression data, representing a prime example for the opportunities and challenges arising when analysing high-dimensional and noisy real-world data. Moreover, for the study of microarray data a multitude of datasets and algorithms are already publicly available, providing enough material for the comparison and combination of datasets and algorithms within the analysis framework proposed in this thesis. In spite of this focus on a particular data type, most of the approaches presented here are equally applicable to protein expression data and other functional genomics data sources, and one chapter of this thesis will therefore be dedicated to the more general analysis of gene and protein lists obtained from any type of biological experiment (see chapter 6).

Similarly, the biological applications of the framework will focus on a specific range of problems, mostly associated with cancer biology, but examples for other biological problem types will be given to highlight the wide range of further potential applications. The choice of this biological focus is motivated by the expectation that due to the characteristics of complex genetic disorders and cancer diseases, which often depend on multiple genetic and epigenetic influences, the analysis of this data is particularly likely to benefit from integrative methods (see chapter 2 for details about the biological background and motivation behind this dissertation). Thus, the biological scope of the thesis is limited to biomedically relevant and representative example applications of the proposed integrative analysis methods, but at the same time seeks to provide researchers in related scientific fields with sufficient information to transfer methodological guidelines to their specific area of research.

Moreover, since the current restraints in the analysis of microarray gene expression data and similar large-scale data sources, which are addressed in this thesis, have already been tackled by previous methods, these will be discussed in detail in a literature survey as part of this thesis (see chapter 3). Due to the large number of published analysis techniques, especially for the study of microarray gene expression data, this survey will focus on state-of-the-art methods tailored to classical machine learning tasks (feature selection, prediction and clustering) and on previous integrative analysis approaches combining data from multiple biological sources (e.g. cellular pathway, molecular interaction and gene/protein expression data). These existing approaches from the literature already address several common problems in the analysis of high-dimensional biological data, but still have several limitations in terms of interpretability, robustness and

applicability to a wide range of platforms. Therefore, the main goals and characteristics of the integrative approach presented here, which differ from previous analysis and data integration systems, have been chosen as follows:

- Methods are presented to combine *algorithms and data types from diverse domains*, in addition to employing cross-study normalisation methods to combine datasets of the same type, and ensemble/consensus methods to combine algorithms for the same problem
- Novel *modular* combinations of previously published approaches interconnect both single algorithms for specific analysis types and corresponding ensemble/consensus methods
- Different algorithms and different datasets are integrated in a *unified approach*, rather than considering data fusion methods and ensemble/consensus approaches separately. In this context, *unified* means that instead of combining already existing analysis techniques based on diverse datasets in a modular or sequential fashion, new methods are developed that operate directly on all input data sources by exploiting different data representations, search methodologies and scoring functions combined in a single algorithm.
- Instead of focussing only on maximising accuracy and robustness, the methods are designed to create *compact and human-interpretable* models (maximising interpretability and accuracy/robustness at the same time)
- Wherever possible, data analysis and parameter selection tasks are *automated* in order to provide non-expert users with a simple way to access and configure the methods within the framework and combine them to a valid statistical analysis pipeline (in most cases using an installation-free, operating system independent and web-based interface)
- *Interactive* means to explore the data and statistical results from analysis, including navigable low-dimensional data visualisations and sortable tables with dynamic and expandable content, are embedded into the implementation of the framework

The last three aspects, which are all related to model interpretability, knowledge management and ease-of-use, are realized by integrating automatic parameter selection methods, visualisation methods and various approaches for enhancing model interpretability into the framework. These include automatic dimensionality reduction and feature selection methods, automatic model parameter selection using penalty terms for model complexity, and novel self-devised rule-based classification methods generating small sets of decision rules. Moreover, automatic methods to handle class imbalances among the samples and redundancy among the input features, and an automatic analysis of the statistical properties of features in ensemble methods are part of the framework.

In addition to these functions and features provided for single analysis modules addressing a specific analysis type (gene selection, clustering, prediction, gene set analysis, etc.), different analysis types are also linked together within the framework, providing multiple possibilities to forward the output from one analysis module to another. Correspondingly, to increase the impact and extend the framework's range of applications, its modules can also exchange data with external web-services, including the "Gene Cards" web-server by the Weizmann Institute of Science and the "Gene Expression Omnibus" (GEO) data repository by the National Center for Biotechnology Information (NCBI). A further benefit of adopting a very modular approach is that, although it requires more time for the initial implementation, it will facilitate extensions of the framework in the future.

In summary, the two main motivations behind this thesis are to exploit the synergies of different information sources, both from the biological input data and from information extracted by different analysis methods combined to an ensemble or consensus, and to enhance the interpretability of the resulting models. Throughout the thesis, the methods employed to achieve these general objectives will not only be applied on classical benchmark datasets for performance evaluation but also on novel real-world datasets to solve specific biological problems, analysed in collaboration with external research groups.

## 1.3 Thesis Organisation

This thesis will first provide the reader with a background on data mining methods for high-dimensional biological datasets and some of their most important biological applications, and then present new integrative analysis methods and techniques to increase model interpretability. Therefore, the dissertation is grouped into the following chapters (excluding this introductory chapter):

**Chapter 2** describes the biological background for this dissertation and the main applications for the bioinformatics methods developed as part of the PhD. It will present some of the main analysis tasks in the study of genetic disorders and cancer diseases, and then explain why there is a need for novel integrative “Systems Biology” approaches. The specific challenges that have to be addressed will be discussed, as well as the opportunities that integrative approaches provide for improvements.

**Chapter 3** contains a literature review about current methods for (low-level) pre-processing and normalisation and (high-level) computational analysis of high-dimensional biological data, with a focus on integrative analysis techniques, including ensemble, consensus and cross-domain analysis methods. In this context, microarray gene expression data will be discussed in particular detail as a prime example for noisy, high-throughput experimental data. The review will start with an overview of classical machine learning analysis approaches (feature selection, classification and clustering) and then present extensions obtained by employing ensemble, consensus and data integration methods.

**Chapter 4** contains a comparative evaluation of algorithms for the classical machine learning tasks feature selection, prediction and clustering on high-dimensional microarray gene expression data. Three different types of algorithms are considered in this comparison: a) single algorithm based classical machine learning methods (e.g. support vector machines for classification), b) ensemble classification and consensus clustering methods, and c) integrative methods using cellular pathway data in addition to the microarray data.

**Chapter 5** discusses the main component of the integrative data analysis framework developed in the PhD project: The *ArrayMining.net* [18] tool set and web-application for microarray data analysis, consisting of six analysis modules for *Gene Selection* (feature selection), *Class Discovery* (clustering), *Class Assignment* (prediction), *Gene Set Analysis*, *Co-Expression Network Analysis* and *Cross-Study normalisation*. Each module contains multiple algorithms, which can be compared and/or combined, and additionally, different modules are interlinked to enable cross-domain integrative analyses (e.g. Gene Set Analysis combined with Class Discovery Analysis). Although this tool set is applicable to many high-dimensional biological datasets, example applications shown in this chapter will focus on the analysis of gene expression cancer datasets, which led to the discovery of a novel, experimentally validated breast cancer marker gene [19].

**Chapter 6** introduces new integrative analysis methods for the general analysis of gene and protein lists



obtained from an experiment. As further data sources, public molecular interaction data and cellular pathway definitions are used within these tools. Specifically, *TopoGSA* [20], a web-application for network topological analysis of gene and protein lists, *EnrichNet* [21], a web-server and algorithm for network-based functional enrichment analysis, and *PathExpand* [22] an algorithm for extending cellular pathway definitions using molecular interaction data will be presented, as well as modular combinations with other tools from the framework.

**Chapter 7** is dedicated to a new integrative rule learning method for increasing the interpretability of machine learning models for biological systems. This approach, termed *Top-scoring pathway pair (TSPP)* algorithm [23], relies on cross-domain data fusion, evolutionary learning and the extraction of robust classification rules representing relations between the gene expression values in pairs of cellular pathways. It generates compact sample classification models consisting of a combination of easily interpretable decision rules.

**Chapter 8** presents a software package for low-dimensional visualisation and interactive exploration of biological data, *VRMLGen* [24], which is used in different modules of the integrative analysis framework. *VRMLGen* generates interactive, web-based 3D data visualisations that are specifically tailored towards the analysis of biological data, integrating functional annotation data into the plots, highlighting regions of high data density, and interlinking data points with information from external biological databases.

**Chapter 9** will summarise the main biological results of the PhD project, including the identification of an experimentally validated breast cancer marker gene, the proposal of candidate disease genes based on computationally extended disease pathway definitions, and the prioritization of putative associations between cancer mutated genes, cellular pathways and different disease processes.

**Chapter 10** will provide a concluding summary and general discussion of the novel methods and results presented in the thesis, as well as an outlook on possible future work.

## 1.4 Methodology

In order to obtain significant, biologically relevant and reproducible results, a methodology consisting of multiple preparatory steps, and following widely accepted standards and guidelines during the development of the data analysis and evaluation pipeline was employed for this project.

First, a survey of the relevant literature was conducted to identify knowledge gaps and promising target applications and methods within the field of large-scale biological data analysis for the development of novel analysis approaches that significantly complement or extend existing algorithms and procedures. An updated version of this survey is also part of this thesis (see chapter 3).

While studying the literature, several high-dimensional microarray datasets were collected, including benchmark datasets, recently published real-world datasets and data from collaborating institutes. These datasets were used to compare current feature selection, clustering and classification methods by setting up an evaluation pipeline, consisting of multiple widely recognized performance measures and cross-validation methods (e.g. using external cross-validation with nested cross-validation parameter optimisation [25] for classification and multiple cluster validity indices for clustering). A comparative evaluation of machine learning methods using this pipeline is part of this dissertation (see chapter 4).

This preparation and new ideas gathered while studying previous methods helped to design a framework and implement corresponding software modules to combine previous machine learning and other data analysis methods using new ensemble learning and consensus clustering techniques, and to explore new modular combinations between analysis types. The analysis pipeline and software was designed in a manner that would allow biological scientists and clinical practitioners without previous background in computer science to apply the methods easily to their data, using a platform-independent web-application, automatic parameter selection mechanisms, and simple, interactive interfaces and visualisations to explore the results. More importantly, users have the possibility to both apply classical machine learning methods they might already be familiar with, but also to compare and combine them with new methods. As the central components of the framework, several new integrative analysis approaches were developed, which combine both diverse algorithms and datasets, using unified graph-based representations of data from multiple biological sources. Finally, in addition to the data pre-processing and analysis modules implemented for this framework, several statistical methods for evaluation purposes and to improve the interpretability of the obtained models were integrated into the analysis system.

To promote data exchange and prevent the framework from becoming an isolated system, the software was interlinked with a multitude of external public data repositories and complementary analysis systems (Gene Cards [26], GEO [27], ENSEMBL [28], DAVID [29], Gene Ontology [30], KEGG [31], and BioCarta [32], among others).

Since one of the main goals of the doctoral project was to develop software of direct practical utility for the biosciences, the data analysis system was also employed for the analysis of several current real-world datasets in co-operation with different partner institutes, focussing mainly on applications in cancer biology. Specifically, collaborations were set up with the Spanish National Cancer Institute (CNIO, Madrid, Spain), the Weizmann institute of Science (Rehovot, Israel), the Queens Medical Centre, the School of Veterinary Medicine and the Institute of Neuroscience at Nottingham University. These institutes provided access to new data to address current research questions of biomedical relevance.

Thus, this plan to build an integrative analysis framework, evaluate different methods and obtain relevant data to approach specific analysis problems provided the foundation for the milestones of the doctoral project, which are reflected by the structure of this thesis and which were pre-defined in advance and monitored throughout the entire duration of the project in regular meetings and progress reports.

## 1.5 Main Results

This final introductory chapter will provide an overview of the most important results obtained from the doctoral project, including the implemented software packages and web-applications, the publications and the main biological results. Detailed discussions of all the software tools and biological contributions will be provided in dedicated chapters (see chapters 4 to 8).

### 1.5.1 The integrative analysis framework

The central result of the project is a software and algorithm framework for integrative analysis of large-scale gene and protein expression data and general gene/protein lists, consisting of multiple interlinked analysis modules, illustrated in figure 1.2.

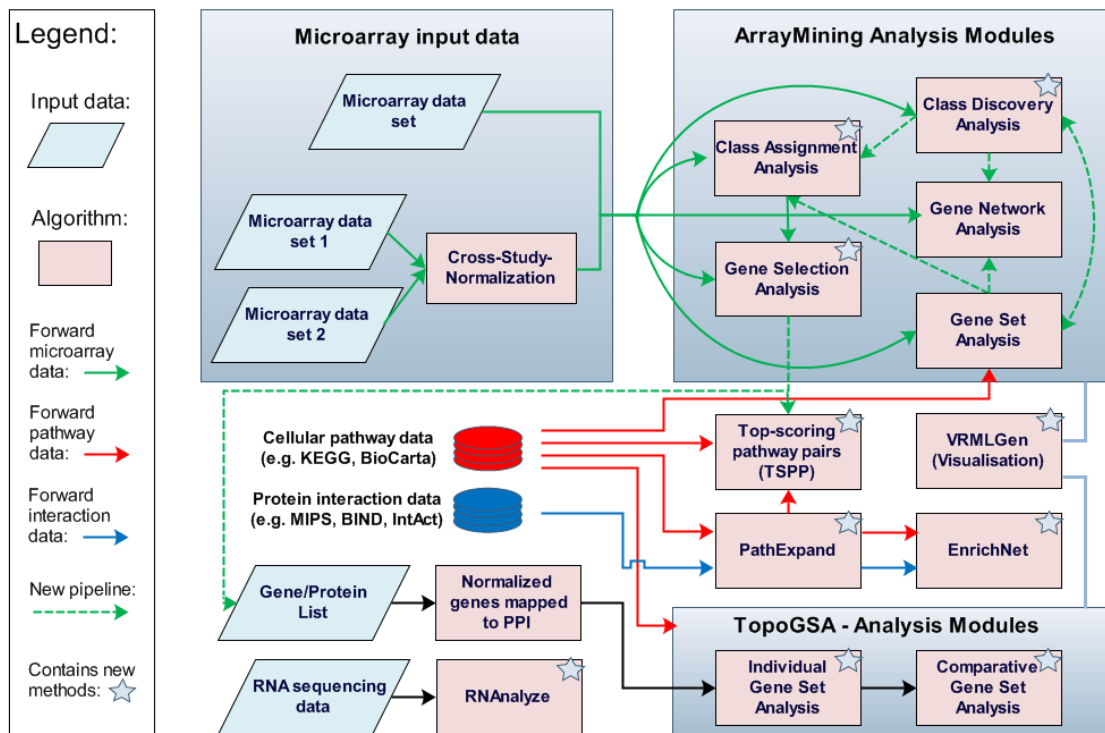


Figure 1.2: Modules and workflow of the integrative data analysis framework presented in this thesis: The *ArrayMining* system for integrative microarray gene expression analysis (top right, consisting of a Gene Selection, Class Discovery, Class Prediction, Gene Set Analysis, Network Analysis and Cross-Study normalisation module), the *PathExpand* method for integrative cellular pathway and molecular interaction data analysis (centre), the *Top-scoring Pathway Pairs (TSPP)* method combining pathway definitions and gene/protein expression data (centre), the network-based enrichment analysis method *EnrichNet* (centre right), the *VRMLGen* software package for creating interactive low-dimensional visualisations of biological (centre right) and the network topological analysis method, *TopoGSA* (bottom right). Modules containing new methods are highlighted by a star symbol, and new pipelines are indicated by dotted lines.

Each module provides access to multiple algorithms for an analysis task, enabling users to compare and/or combine them using novel ensemble or consensus techniques. The modules enable the integration of external biological data, including functional annotation data, cellular pathway definitions and molecular interactions, into the analysis, and can be combined sequentially into different user-configured cross-domain analysis pipelines. Automatic parameter selection and optimisation mechanisms, the generation of compact rule-based models, and interactive means to visualise, explore and statistically analyse the results obtained from an algorithm, facilitate the usage of the software and the interpretation of the data.

Different components of this framework have been developed as independent software projects and presented in dedicated publications. The main component is the *ArrayMining.net* [18] tool set and web-application for DNA- and protein-microarray data analysis, which is freely available on the internet since February 2009, and has been accessed more than 31,000 times by more than 10,000 users. It consists of six analysis modules (see the large blue boxes “Microarray input data” and “ArrayMining Analysis Modules” in figure 1.2) for combining datasets from different microarray studies (Cross-Study normalisation), classical machine learning tasks (Feature Selection, Clustering and Classification) including ensemble and consensus methods, and new integrative and specialized analysis methods (Gene Set Analysis and Co-Expression Analysis). Additionally, these modules are interlinked with the other algorithms in the framework and external data repositories and web-services. *ArrayMining.net* will be discussed in detail in chapter 5.

A further component for microarray data analysis in the framework is the self-devised *Top-scoring pathway pairs* (TSSP, [23]) algorithm, which integrates cellular pathway information into the machine learning analysis of gene expression data to identify compact and easy-to-interpret decision rules for sample classification (see details in chapter 7).

Two other integrative framework tools, which also exploit information from pathway definitions and additionally make use of molecular interaction networks, are the new algorithm *PathExpand* [22], which extends pathway definitions using a graph analysis on protein-protein interaction data, and *EnrichNet* [21], an approach to enhance classical functional enrichment analysis, estimating the significance of the functional association between pairs of gene/protein sets using distance information from molecular networks (see chapter 6).

The combined information content from interaction network and pathway data is also exploited by the web-application *TopoGSA* [20], which maps a gene or protein set provided by the user to a molecular network for the corresponding species and analyses its topological properties in comparison to gene/protein sets representing known cellular processes, complexes and functional annotations (see chapter 6). Although *EnrichNet* and *TopoGSA* are not designed to make use of expression level measurements from gene or protein expression data, these methods have a significantly wider applicability than the microarray-related analysis modules, and can be employed to investigate gene/protein sets obtained from any biological experiment or bioinformatics analysis in more detail.

In order to keep pace with recent developments, the framework additionally provides a method to analyse the most recent and accurate type of large-scale gene expression data, RNA sequencing data. This software, *RNAnalyze* [33], which has also been made available as a web-application, combines the information from multiple gene selection methods within a gene set enrichment analysis approach.

Finally, methods for web-based interactive data exploration and low-dimensional visualisation are used within multiple components of the framework, and have also been published in a dedicated software package, *VRMLGen* [24].

### 1.5.2 Collaborative project results

Using combinations of the software tools within the framework, a wide range of biological analyses have been conducted in collaboration with external institutes. One of the key results was obtained in co-operation with the department of Histopathology at the *Queens Medical Centre* in Nottingham, using the gene set analysis, ensemble feature selection and classification methods in ArrayMining to analyse data from a large-scale microarray cohort study with samples from 128 breast cancer patients. In this pre-clinical study a new candidate marker gene for a special breast cancer tumour subtype (the oestrogen-receptor positive luminal-like subtype) was identified and experimentally validated by immunohistochemistry using tissue microarrays. A dedicated manuscript, proposing the corresponding human gene *REG* (“Ras-related and oestrogen-regulated growth inhibitor”) as a new tumour marker, has recently been published [19].

A second cancer-related project, implemented in co-operation with the *Spanish National Cancer Institute* (CNIO, Madrid) during a three-months secondment in Madrid, analysed sets of human genes known to be mutated in more than 30 different cancer types using the tools PathExpand, TopoGSA and EnrichNet from the framework. By combining molecular network and cellular pathway data, as well as the lists of cancer-mutated genes, several disease-related pathways enriched in these genes were extended using the PathExpand tool, and enrichment in cancer-mutated genes was also detected among the genes added by the extension procedure [22]. Applying these analyses to cellular disease pathways enabled the proposal of new putative disease genes, e.g. for Alzheimer’s disease the prediction of two candidates was corroborated by previously published experiments.

In a related study, the complete set of known human cancer genes [34] was analysed with respect to its topological properties when mapped to a molecular network using the TopoGSA tool. After assembling a large-scale human protein interaction network in collaboration with the CNIO and mapping the genes onto the network, TopoGSA determined their topological properties (e.g. their centrality and their tendency to form clusters) and compared them to known gene sets representing cellular processes, complexes and functional annotations. The final results revealed that cancer genes have markedly distinct topological properties in comparison to gene sets representing metabolic and regulatory processes, even after accounting for biases resulting from the inclusion of small-scale studies in the construction of the interaction network [20]. To extract further information on cancer gene set associations from molecular network data in a follow-up project, the EnrichNet approach and web-application [21] was developed, as an extension of classical functional enrichment analysis. In contrast to existing methods, which assess the significance of the overlap between the datasets (e.g. using the Fisher exact test) or compute the enrichment of “true positive” disease genes among an experimentally derived ranking list of candidate disease genes (e.g. using the Kolmogorov-Smirnov test), the EnrichNet approach makes use of network distance information to obtain more sensitive estimates of the functional associations (see a detailed discussion in chapter 6). Both EnrichNet and TopoGSA have been interlinked with the GeneCards web-service from the co-operating Weizmann institute of Science (Rehovot, Israel) and used several hundred times by external visitors on the web.

Apart from these ongoing co-operations, various short-term biomedical research projects have been conducted based on software from the framework. In a three-month collaborative project with the *School of Veterinary Medicine* at Nottingham University, funded by the “Bridging the Gaps” initiative, microarray analysis methods from the framework and various literature mining tools were combined to compare gene expression samples from horse cartilage tissue under different drug treatments against arthritis. The goal was to study the drug effect on a molecular basis, in order to develop more targeted therapies in the future. To maximise the robustness and accuracy in the ranking of genes in terms of their differential regulation

across diverse treatment types, special analysis methods exploiting per-gene replicate measurements on the microarray chips were used and information from the literature was integrated into the analysis. This methodology provided a set of 17 high-confidence target genes for further study of the cellular response to drug treatment (the majority of which had previously been implicated in arthritis).

Similarly, in a still ongoing co-operative study with the *Institute of Neuroscience* at Nottingham University, the effects of nicotine on the brain were analysed using gene expression data and rat brain cells as an animal model, ensemble feature selection methods from the framework were employed to identify the genes with the most significantly differential expression profiles in the nicotine and control samples.

A central part of the integrative analysis framework is dedicated to network analysis methods, including the Gene Co-Expression Network Analysis module in ArrayMining, which builds a network of genes (represented as nodes), which are connected by edges if the corresponding gene expression values are significantly correlated. This module was used in collaboration with the *Division of Plant and Crop Sciences* at Nottingham University, to generate a genome-wide network model describing transcriptional interactions in dormant and germinating seeds, in the model plant organism *Arabidopsis thaliana*. The model enabled the prediction of genes regulating the plant germination process with higher accuracy than previous alternative methods. To allow external researchers to explore this data using their own query gene set, an interactive network visualisation was developed and made publicly available on the web [35].

As part of the above projects, a new web-based 3D-visualisation software to inspect low-dimensional representations of high-dimensional biological datasets was implemented additionally, to enable direct visual inspection of the results for different analyses online. This software package, “VRMLgen”, enables users to create interactive scatter plots, bar charts and 3D mesh visualisations in web-ready formats, and interlinks biological features (e.g. gene and protein names) in the data via hyperlinks with public biological databases. VRMLGen has been presented in a dedicated publication [24] and as an open-source software package in the “Comprehensive R Archive Network” (CRAN, <http://cran.r-project.org>).

Finally, in addition to the biological applications of the algorithms and software tools developed in this PhD project, some of the machine learning methods were also used to enter different data mining competitions. By participating in these competitions, the performance of the ensemble machine learning methods on the ArrayMining.net prediction server could be evaluated against a large number of other recent methods in a fair and unbiased setting. In the KDD Cup 2009 machine learning competition (<http://www.kddcup-orange.com>), among the 4921 complete valid entries from 465 entrants the best-performing ensemble method developed during the PhD was ranked 53rd in the “Slow Track” category. In a smaller data-mining competition dedicated specifically to microarray analysis, the RSCTC Discovery Challenge 2010, the automatic ensemble learning approach was ranked 26th out of 100 participants, with an average classification accuracy across six data sets within 3% of the three top-ranked methods in the competition. These rankings suggest that automated ensemble prediction methods can achieve competitive accuracies in relation to other state-of-the-art approaches.

In summary, the results of this thesis highlight the benefits of ensemble, consensus and cross-domain integrative analysis methods both for classical machine problems and new biology-inspired analysis and modelling tasks, and cast light on the obstacles that have to be overcome to exploit these opportunities. Detailed background information, methodological and biological results and critical method comparisons will be provided in the next chapters.

### 1.5.3 Software tools and web-applications



*ArrayMining*: A web-server for ensemble and consensus analysis of microarray data, [www.arraymining.net](http://www.arraymining.net) (BMC Bioinformatics, 2009, accessed more than 31,000 times, as of January 2011), interlinked with Gene Cards [26], DAVID [29], GEO [27], KEGG [36] and Gene Ontology [30]



*TopoGSA*: A web-tool for comparative network topological analysis of gene sets, [www.infobiotics.net/topogsa](http://www.infobiotics.net/topogsa) (Bioinformatics, 2010, accessed more than 2,900 times, as of January 2011), interlinked with Gene Cards [26], KEGG [36], BioCarta [32], Reactome [37] and Gene Ontology [30]



*TSPP*: A software using pairwise relations between pathway expression fingerprints for supervised classification of microarray sample classification (German Conference on Bioinformatics, 2010)



*VRMLGen*: A software package in R for 3D-visualisation of high-dimensional biological datasets on the web, <http://cran.r-project.org/web/packages/vrmlgen> (Journal of Statistical Software 2010)



*PathExpand*: A visualisation of network representations of extended cellular pathways and processes, [www.infobiotics.net/pathexpand](http://www.infobiotics.net/pathexpand) (BMC Bioinformatics and RECOMB Computational Cancer Biology Workshop 2010)



*SeedNet Online*: An interactive gene co-expression network visualisation to investigate transcriptional interactions between dormant and germinal *Arabidopsis thaliana* seeds (Proc. Natl. Acad. Sci. USA 2011)



*EnrichNet*: A web-application for network-based gene set enrichment analysis, [www.infobiotics.net/enrichnet](http://www.infobiotics.net/enrichnet) (manuscript in preparation), interlinked with Gene Cards [26], KEGG [36], BioCarta [32] and Gene Ontology [30]



*RNAalyze*: A web-server for gene set enrichment analysis of RNA sequencing data (manuscript in preparation), interlinked with KEGG [36] and Gene Ontology [30]


### 1.5.4 Publications

[1] **E. Glaab**, J.M. Garibaldi, and N. Krasnogor. ArrayMining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalisation. *BMC Bioinformatics*, 10(1):358, 2009.

[2] **E. Glaab**, A. Baudot, N. Krasnogor, and A. Valencia. TopoGSA: network topological gene set analysis. *Bioinformatics*, 26(9):1271–1272, 2010.

[3] **E. Glaab**, A. Baudot, N. Krasnogor, and A. Valencia. Extending pathways and processes using molecular interaction networks to analyse cancer genome data, 2010. *BMC Bioinformatics* (BioMed Central designation: **highly accessed**, RECOMB Computational Cancer Biology 2010), 11(1):597, 2010.

- [4] **E. Glaab**, J.M. Garibaldi, and N. Krasnogor. Learning pathway-based decision rules to classify microarray cancer samples. In D. Schomburg and A. Grote, editors, *German Conference on Bioinformatics 2010*, volume 173 of *Lecture Notes in Informatics*, pages 123–134. Gesellschaft fuer Informatik, 2010.
- [5] **E. Glaab**, J.M. Garibaldi, and N. Krasnogor. vrmngen: An R package for 3d data visualisation on the web. *Journal of Statistical Software*, 36(8):1–18, 2010.
- [6] H. O. Habashy, D. G. Powe, **E. Glaab**, N. Krasnogor, J. M. Garibaldi, E. A. Rakha, G. Ball, A. R. Green, C. Caldas, and I. O. Ellis. RERG (Ras-related and oestrogen-regulated growth-inhibitor) expression in breast cancer: A marker of ER-positive luminal-like subtype. *Breast Cancer Research and Treatment*, 128(2):315–326, 2011.
- [7] G. W. Bassel, H. Lan, **E. Glaab**, D. J. Gibbs, T. Gerjets, N. Krasnogor, A. J. Bonner, M. J. Holdsworth, N. J. Provart. A genome-wide network model capturing seed germination reveals co-ordinated regulation of plant cellular phase transitions *Proc. Natl. Acad. Sci. USA*, 108(23):9709–9714, 2011.
- [8] **E. Glaab**, J. Bacardit, J.M. Garibaldi, and N. Krasnogor. Using rule-based evolutionary learning for candidate disease gene prioritization and sample classification of cancer gene expression data (in preparation, title might change).



Please note: Terms which appear underlined when they first occur in the thesis are defined in the glossary (see Appendix at the end of this dissertation). Moreover, paragraphs which are highlighted by a blue vertical stripe (see example on the left) contain key messages of the corresponding chapters.



## Chapter 2

# Biological Background and Fields of Application

### Chapter abstract

The development of new high-throughput technologies in the biosciences in recent decades, including High-Throughput-Sequencing (HTS), DNA and protein microarrays, has provided the scientific community with almost complete genome sequences for several species and a multitude of transcriptome, proteome and metabolome datasets for diverse cell types under numerous biological conditions of interest. This enormous amount of publicly available data has raised hopes that researchers will in the long run be able to obtain a more holistic understanding of the molecular mechanisms in living cells, especially under disease conditions, by analysing complete gene-, protein- and metabolite networks and not only their individual components. These new possibilities have led to the establishment of *systems biology* as a novel scientific discipline, aiming at a holistic interpretation of interactions within biological systems. This discipline is also accompanied by a novel scientific methodology, *discovery science*, in which the discovery process, including hypothesis formation, is automated to a great extent and operates on large volumes of data. This development can be seen as a complement to classical *reductionism*, “the practice of describing or explaining a complex phenomenon in terms of relatively simple or fundamental concepts, especially when this is said to provide a sufficient description or explanation” (Oxford English Dictionary, 3rd edition, 2009). Importantly, reductionism does not neglect the interactions between the different components of a systems, but enables the filtering of variables, while still providing explanations for some system-wide properties.

This chapter will provide an overview on some of the most important biomedical applications and challenges for systems biology based bioinformatics approaches. It will briefly describe genetic disorders and cancer diseases which are difficult to treat partly due to their complex systemic properties and influences, and explain how new large-scale data sources enable a more comprehensive and rapid biomarker discovery and drug target screening. More importantly, the final section of this chapter will discuss how these data analysis tasks can profit from novel network-based systems biology approaches and integrative analysis techniques. In combination with the literature survey on classical machine learning methods and novel integrative data

mining approaches (see chapter 3), this overview will provide the background and motivation for the new integrative bioinformatics approaches presented in this dissertation.

## 2.1 Genetic Disorders and Cancer Diseases

Diseases with genetic components cover a multitude of malignancies and genetic disorders. Hereditary single gene disorders alone account for more than 4000 different human diseases, and although cancer is often referred to as a single condition, more than 200 different cancer diseases exist, differing in terms of the affected organs and tissue types [38]. While most genetic disorders are mainly hereditary diseases and occur relatively rarely (affecting one individual in several thousands or millions), cancer diseases are frequently caused by environmental factors and often occur spontaneously. They also belong to the main causes of deaths, with 13% of all world-wide human deaths in 2007 resulting from cancer [39]. Importantly, although people at all ages can be affected by cancer, the risk tends to increase with age [40]. Given an aging population in industrialised countries and a lack of causative treatments against cancers, the search for effective therapies is therefore growing in significance.

In this section a brief overview of diseases with genetic components will be given, focussing on current therapeutic approaches and their limitations. The next sections will discuss how new bioinformatics methods for rapid identification of drug targets, virtual drug screening and rational drug design, as well as integrative systems biology analysis approaches can help to overcome some of these limitations.

In general, genetic disorders can be grouped into *single gene disorders*, resulting from a single mutated gene, and *multifactorial* or *polygenic disorders*, which are associated with multiple genes, as well as external environmental factors.

Single gene disorders are mostly inherited diseases with a multitude of possible inheritance patterns, sometimes affected by epigenetic influences like genomic imprinting [41] (e.g. silencing of an allele by CpG-methylation of the corresponding promotor region). Generally, the inheritance patterns can be grouped into recessive and dominant patterns, depending on whether the corresponding trait needs to be expressed on two alleles (recessive) or only one allele (dominant) to determine the final phenotype (special cases of co-dominant and semi-dominant diseases, where both alleles have the same influence on the phenotype, exist additionally). Moreover, inheritance patterns can be distinguished according to whether the corresponding gene is located on an autosomal or gonosomal (i.e. sex-linked) chromosome. Prominent examples for different single gene disorders and their inheritance patterns are shown in table 2.1. An example inheritance pedigree for an autosomal dominant disease is shown in figure 2.1.

For most of these diseases, the precise genetic cause is well-known, e.g. an extension of a nucleotide sequence region rich in CAG-repeats in the Huntingtin gene is known to cause Huntington's disease, but a causative treatment would require the development of a new gene therapy. However, bioinformatics data analysis approaches can support the development of cheaper and more sensitive and specific methods for diagnosis, prognosis and disease progression monitoring, and the identification and design of new drugs.

Interestingly, although most of these disorders have a relatively low prevalence of one in a few thousand of cases and affected carriers would be expected to have a selective disadvantage, these genetic diseases have not disappeared over thousands of years [42]. A likely reason for this long survival of low-prevalence genetic diseases is that some disorders can also endow carriers with a selective advantage, e.g. heterozygous carriers of sickle-cell anaemia have an increased resistance against malaria [43].

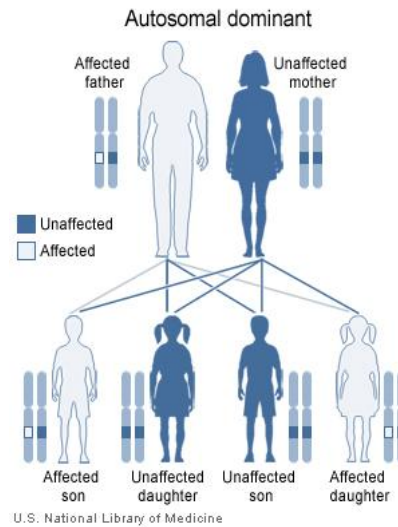


Figure 2.1: Example inheritance pedigree for an autosomal dominant disease with only one heterozygous carrier of the disease in the parental generation, resulting in a 50% chance for members of the first filial generation to carry the disease (source: U. S. National Institutes of Health).

Table 2.1: **Overview on genetic inheritance types and representative single-gene genetic disorders**

Inheritance types	Representative diseases
Autosomal dominant	Huntington's disease, familial hypercholesterolemia, Marfan syndrome
Autosomal recessive	sickle-cell anaemia, cystic fibrosis, phenylketonuria, spinal muscular atrophy
Gonosomal, X-linked, dominant	Aicardi syndrome, Rett syndrome
Gonosomal, X-linked, recessive	red-green colour blindness, Hemophilia A, Duchenne muscular dystrophy
Gonosomal, Y-linked	hypertrichosis pinnae (human hairy ears trait), male infertility
Mitochondrial (maternal)	Leber's optic atrophy

However, complex, multifactorial disorders like the majority of cancers are much more widespread than the above single-gene disorders. Importantly, multifactorial or polygenic disorders also include many diseases which are not commonly associated with genetic disorders, e.g. heart disease, obesity, asthma, hypertension, inflammatory bowel disease and multiple sclerosis, which are often only partly influenced by genetic pre-dispositions and partly by environmental and lifestyle factors. The complexity of these diseases results both from the involvement of multiple genes in the disease and from the fact that defects in different genes and proteins can result in similar deregulations of a cellular pathway or process.

In spite of the detailed knowledge available for many genetic disorders, for most of these diseases with genetic components currently no cure exists and most therapies only attempt to reduce the severity of the symptoms. For the future, scientists hope to develop somatic gene therapies [44], which compensate the effect of mutated genes by inserting copies of the non-mutated gene into the genome of affected cells (without affecting any germ cells for ethical reasons). However, several obstacles have to be overcome until such gene therapies will become feasible. An effective therapy would not only require a targeted delivery of the corresponding gene's DNA to many affected cells, but also the integration of the DNA carrier into these cells and into the cell nucleus, the controlled release of the DNA and the integration into the genome. Although many DNA viruses operate similarly and integrate their DNA into the genome of the host cell, the uncontrolled replication of viruses might rather represent a risk to the host organism than an opportunity for a new gene therapy. Thus, there is a need for new, controlled drug delivery systems, a better understanding of complex polygenic diseases and a new rapid drug target identification and rational drug design process.

Bioinformatics methods can assist biological and clinical researchers both in improving the basic understanding of genetic disorders, and developing diagnosis systems and therapies for these diseases using *in silico* identification of new drug targets and techniques for rapid virtual drug screening, computational drug design and optimisation of given drug lead structures. The next section will discuss techniques for discovering biomarkers, which are useful for disease diagnosis and monitoring, as well as methods for predicting the disease subtype and the long-term prognosis.

## 2.2 Biomarker Discovery and Outcome Prognosis

Classically, biomarkers are blood tests which measure the abundance of a protein that reflects the presence and progression of a certain disease state, e.g. the presence of an antibody can point to an infection [45,46]. However, biomarkers can also be genes, RNA, cells, enzymes, hormones or other molecules (also known as “molecular biomarkers”), imaging biomarkers (MRI, CT, PET), and even a classic laboratory parameter like the body temperature can be regarded as a simple biomarker, e.g. for fever. Biomarkers are used for monitoring the progression of a disease, but more importantly, they can also be used for the early detection of a disease or disease risk. Early diagnosis is vital for the effective treatment of many diseases, especially cancers, because prior to metastasis tumours which are confined to one location can often be removed completely. Without biomarkers, early and reliable disease detection is often impossible, because several diseases like rheumatoid arthritis and Alzheimer's start with a symptom-free phase [47, 48]. Similarly, if biomarkers are used as a risk indicator for a disease, they can assist even in preventing the onset of a disease completely (“preventive medicine”), e.g. monitoring cholesterol levels can help to prevent coronary diseases [49]. Moreover, biomarkers are particularly important tools for a more *personalised medicine*, enabling doctors to adjust biomedical decision making to an individual's specific disease state by obtaining a molecular portrait of the individual (“molecular profiling”), improving the diagnosis and tailoring the

treatment plan and therapy to the individual. In drug development, biomarkers can promote the identification of new disease targets, and both directly and indirectly the improvement of drugs and design of novel active compounds. For example, to evaluate a potential drug therapy, and in particular when testing the safety of a drug in addition to its clinical effect, biomarkers can also be used as “surrogate endpoints” for survival [50]. At a later stage of the drug development process, biomarkers can be employed in pharmacokinetics and pharmacodynamics studies (as part of clinical phase I studies), to identify a suitable dose, dose interval and application type for a drug [51]. For this reason, clinical researchers distinguish between *disease-related* and *drug-related* biomarkers, depending on whether a biomarker is used for predictive and diagnostic purposes, or for studying drug effects.

High-throughput genomics, transcriptomics and proteomics experiments and bioinformatics methods analysing these data can help to speed up the discovery of new molecular markers. As illustrated in table 2.2, providing an overview of bioinformatics applications in biomedicine, *in silico* biomarker discovery represents just one group of analysis approaches among a multitude of bioinformatics sub-disciplines linked to biomedical research. Moreover, figure 2.2 shows an overview of typical stages in a biopharmaceutical drug development pipeline, highlighting the stages in which bioinformatics methods can be used to speed up and rationalize the development process (blue colour).

Table 2.2: Application of bioinformatics tools in biomedicine

Biomedical applications / clinical phases	Supporting bioinformatics disciplines
Disease model generation	high-throughput data mining, pathway & network analysis
Drug target identification	<i>in silico</i> target screening (docking [52], feature trees [53])
Drug screening	<i>in silico</i> drug screening (docking [52], feature trees [53], ligand superposition [54])
Lead optimisation	QSAR [55], CoMFA/CoMSIA [56], QM/MM [57]
Disease diagnosis and prognosis	<i>in silico</i> biomarker discovery [58]
Drug pharmacokinetics & safety	<i>in silico</i> ADME/Tox models [59]

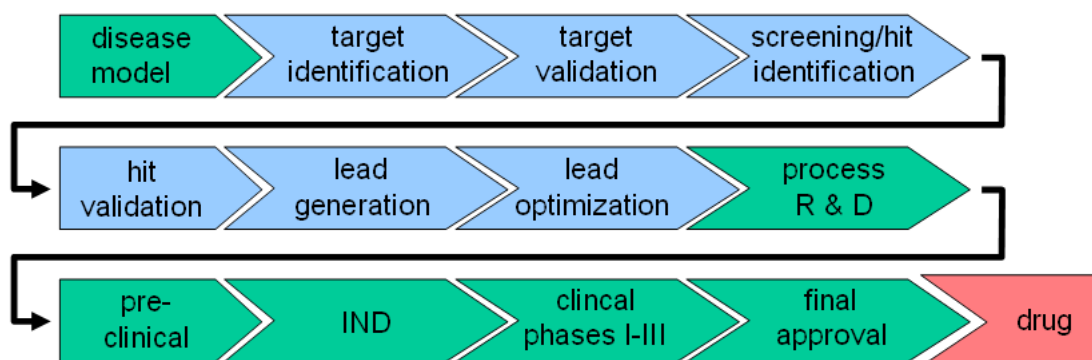


Figure 2.2: Overview of common stages in a drug development pipeline. All stages in which bioinformatics methods are applied to speed up the research and development process are highlighted in blue colour.

The classical biomarker discovery approach is applied to transcriptomics data like DNA microarrays or data collected with other gene expression measurement techniques (qPCR = real-time quantitative polymerase chain reaction, RNA sequencing, SAGE = serial analysis of gene expression, Northern blotting), to identify molecular markers that enable a discrimination between a healthy control state and different disease states, drug effects etc. Proteomics techniques are mostly using mass spectrometry and protein microarray tech-

niques, e.g. LC/MS (liquid chromatography mass spectrometry), MALDI-TOF MS (matrix-assisted laser desorption ionisation - time-of-flight mass spectrometry), 2D-PAGE (2D polyacrylamide gel electrophoresis), tissue microarrays and antibody (Ab) microarrays. More recently, new “-omics”-techniques analysing all metabolites (Metabolomics), lipids (Lipidomics) and the entire complement of sugars (Glycomics), have been developed to find new biomarkers.

All of these experimental platforms and measurement techniques for the discovery of biomarkers have in common, that their measurements are affected by various sources of noise, and that multiple measurements are made in parallel, requiring adjustments for multiple hypothesis testing in the statistical analysis as well as special techniques to deal with the analysis of high-dimensional noisy data with small sample sizes (many attributes in relation to few observations for each attribute). Therefore, bioinformatics analyses methods, and in particular integrative techniques which exploit the synergies of complementary data sources and different types of statistical, graph-based and information-theoretic analysis, are needed the robust identification and prioritisation of candidate biomarkers. Importantly, bioinformatics analysis and experimental validation should go hand in hand, because noisy high-throughput data alone cannot provide absolutely reliable evidence for the verification of biological models, and the models are always simplified representations of reality. Thus, a typical biomarker discovery pipeline will involve high-throughput data collection from multiple sources (e.g. gene expression microarray, protein expression and molecular interaction data), data integration and computational analysis (which is the main focus of this thesis), and experimental validation using small-scale high-sensitivity experiments (e.g. repeated qPCR experiments). Chapter 3 will provide an overview on previous computational analysis techniques for this type of data, as well as a motivation for new integrative analysis methods. However, before covering the *in silico* analysis of the data, the next section will first discuss the role systems biology approaches in the analysis of new large-scale datasets.

### 2.3 Role of Systems Biology in Elucidating the Basis of Complex Diseases

The analysis of data from high-throughput experiments requires different approaches than those used in the classical reductionist methodology. The measurements for single molecules (genes, proteins and metabolites) in these experiments are frequently affected by high levels of noise, and often only when considering properties of whole cellular pathways, processes and complexes, robust patterns and structures emerge in the data, which can be verified by other experimental procedures and platforms. This observation also matches with the basic rationale behind systems biology and synthetic biology, according to which many functions within a biological system can only be explained by looking at the multitude of interactions between the components of the system, rather than only at single components in isolation. Studying single structures and isolated processes in detail according to a reductionist method continues to be an indispensable scientific methodology, required to verify specific hypotheses at the single-molecule level. However, the complementary systems biology approach, looking at the interplay between a multitude of components in a biological system and studying patterns emerging on a system-wide level, is becoming as practically useful as the traditional methodology.

Importantly, systems biology approaches are still consistent with the scientific method [60], i.e. experience is used to form a conjecture (a testable hypothesis or model), and this theory/model is used to make predictions, which are falsified by an experiment. The knowledge gained from this procedure can again be used to refine the model and repeat the whole process in an iterative, cyclic process. The main difference to

the reductionist approach is that a multitude of hypotheses are formed, and advanced analysis methods are required to filter these hypotheses to identify those providing the most relevant and significant information, adjusting the results for the multiple testing problem [2, 3]). Moreover, systems biology accounts for new interdisciplinary aspects of biology, combining knowledge and analysis techniques from molecular biology, biophysics, biochemistry, biopharmacy, microbiology, mathematics, computer science, graph/network analysis and other disciplines. This collaboration between experimental and quantitative scientists is often an essential requirement, because many large-scale experimentation and data acquisition techniques provide noisy and high-dimensional data, which requires special computational analysis techniques to filter out irrelevant information and reduce the dimensionality of the data. Furthermore, effective search methodologies often have to be employed to find the best model parameter combinations, providing a good fit to the data and compact and interpretable explanations for the relations between variables. For this reason, systems biology has not only been driven by new system-wide experimental high-throughput techniques, but also by new computational technologies, and bioinformatics algorithms and data structures, which enable the analysis of high-dimensional data on standard desktop computers.

Systems biology methods do not only analyse large-scale “omics” datasets measuring abundances and activities of many molecules in parallel, but a special focus lies on interactions, and dynamic relations between different molecules and processes. The classical “omics” have therefore been extended by *Interactomics* (the analysis of molecular interactions, like protein-protein interactions) and *Fluxomics* (the quantitative study of the biomolecular fluxes, i.e. the rates of passage of biomolecular substances through a given metabolic pathway, and the dynamic synthesis and conversion of these substances in metabolic networks). Importantly, in agreement with findings showing that many genetic disorders are complex, multifactorial and polygenic (see previous section), systems biology focuses on the network-based analysis of high-throughput data and the multivariate analysis of activity and expression data, to account for the combined influence of multiple genes, proteins and metabolites on a biological process of interest. Indeed, assays for the diagnosis and monitoring of complex diseases might often require combinations of multiple biomarkers, and a personalised medicine tailored to a specific patient can even necessitate conducting high-throughput analysis for individuals (e.g. by devising a cheap diagnostic microarray, representing only replicates of the most relevant biomarker genes; an example is the “MammaPrint” breast cancer molecular diagnostic test based on a 70-gene signature [61]). Classical reductionist approaches would often not be adequate to analyse data for developing diagnostic assays for complex diseases, unless a sufficiently reliable single-gene or -protein marker exists for the disease under consideration.

In summary, systems biology approaches complement rather than replace the traditional reductionist methodology, and provide new opportunities for the rapid identification of potential drug targets, and screening for biomarkers and novel drug candidates. Apart from the new high-throughput experimental technologies which promote the spread of systems biology studies, integrative bioinformatics analyses methods represent one of driving forces behind systems biology, helping to uncover complex relations between attributes in the data and speeding up screening processes by improved candidate biomarker, drug target and drug substance rankings.

In order to introduce new integrative analysis methods and compare them to traditional algorithmic strategies, the next chapter will provide a literature survey and overview of both classical machine learning methods and new biology-tailored and integrative approaches for the computational analysis of high-throughput data.

# Chapter 3

## Literature Review

### Chapter abstract

The analysis of complex, high-dimensional and noisy biological datasets has become a common challenge in both clinical and basic biological research. To address the statistical and computational difficulties associated with new large-scale data sources, a multitude of highly specific and tailored computational biology approaches for the investigation of large gene and protein expression datasets and other high-throughput data sources have been developed in recent years.

This chapter contains a literature survey discussing current methods for high-dimensional biological data analysis, with a special focus on ensemble/consensus and cross-domain integrative analysis techniques, in order to provide the reader with a representative overview on current state-of-the-art methods in this field and the limitations and opportunities for the development of novel analysis tools. Since data normalisation and quality checking are important pre-requisites for a successful higher level statistical analysis, corresponding lower level pre-processing methods will be discussed first. However, the main goal of the chapter is to guide the reader through a representative selection of current higher-level machine learning analysis techniques and their integrative extensions combining information from multiple datasets and/or algorithms. Most of these methods will be discussed and illustrated using input from microarray gene expression data as a prime example for high-dimensional biological data. However, these approaches are often directly applicable to dataset types obtained from other high-throughput experimental technologies. For the lower level pre-processing methods, which are typically data type and platform-specific, separate sections will deal with different input data types, including microarray, protein-protein interaction and cellular pathway data.

### 3.1 Low Level Analysis: Pre-Processing, Normalisation and Quality Checking

In high-throughput biological studies different samples are often affected by a different experimental bias and quality of the corresponding biological material. To account for these technical and biological sources of noise, data pre-processing and quality analysis methods are essential first steps in a data processing



pipeline designed to extract new biologically meaningful insights from the input. This section will discuss these challenges and current methods to tackle them, using the example of microarray gene expression data as a representative input data type, and highlighting typical problems in high-dimensional biological data analysis. Though an exhaustive discussion of microarray pre-processing methods is beyond the scope of this thesis, a diverse selection of widely used approaches will be presented, illustrating that a straightforward “method of choice” for data pre-processing and normalisation does not always exist in biological data analysis. The discussion will also reveal that there is a wide room for improvement even for low level pre-processing methods, due to a multitude of error sources which are difficult to model realistically. A summary of the general types of higher level machine learning analyses, which are the main focus of this chapter, are shown in figure 3.1, whereas figure 3.2 provides an overview of the workflow for the low level analysis of microarray data.

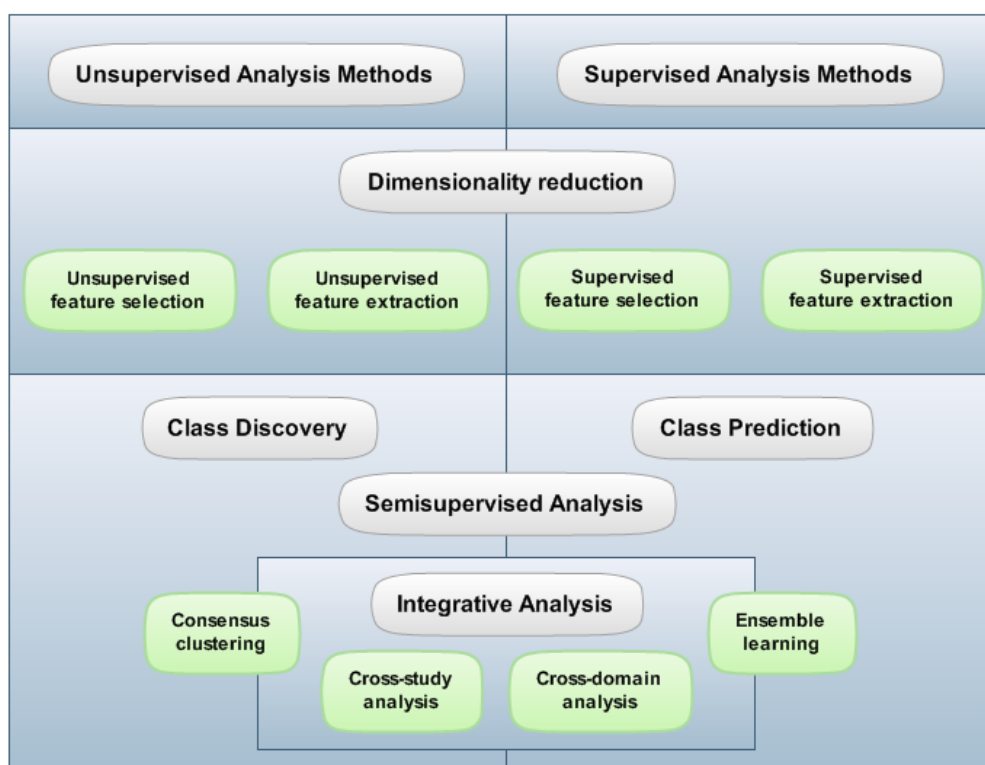


Figure 3.1: Overview of supervised and unsupervised higher level statistical learning methods considered in this survey, including integrative methods to combine algorithms and datasets within and across different domains in biology and computer science.

### 3.1.1 Microarray technology - Introduction

Microarrays have been one of the first miniaturised high-throughput experimental technologies in the biosciences, and in spite of recent advances in RNA sequencing technology, they still belong to the most widely used large-scale measurement devices in clinical and biological research. The wealth of information gathered in microarray studies and the variety of problems that have to be addressed in the analysis of the data, make microarrays a prime example for the opportunities and challenges of biological high-throughput technologies.

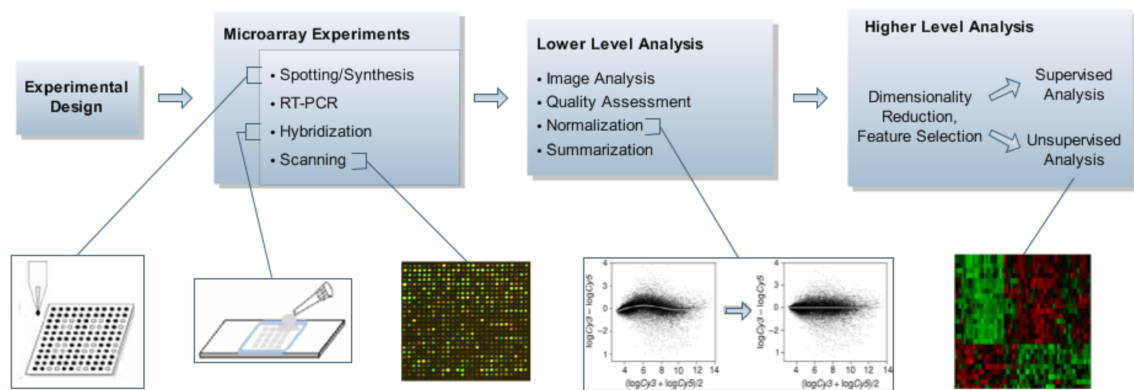


Figure 3.2: Overview of a typical microarray analysis workflow. This survey will focus on the steps following after the microarray image analysis, in particular higher level analysis methods.

With *DNA gene expression microarrays* (also known as *DNA chips*, *gene arrays* or *gene chips*) the activity levels of thousands of genes from a cell sample can be measured using an approximate quantification of the messenger RNA abundance for each gene. Thus, especially for organisms with a small genome, a microarray enables the measurement of the entire *transcriptome* of a cell, i.e. measuring the expression levels of all genes in a certain cell type under well-defined conditions. This technology has countless applications with high practical relevance in basic biology and biomedicine, including the comparison of different disease states (e.g. tumour types and progression states) in biological samples, the analysis of the effects of drugs, toxins or viral infections, the identification of new marker genes for disease monitoring and outcome prediction, and the discovery of new drug targets for developing novel therapies.

However, microarray studies also have several practical limitations. In contrast to more accurate methods for gene expression level determination like real-time PCR [62], they can often only provide relative expression levels, i.e. enable a relative comparison of expression levels in cells from different tissue types, alternative disease states or different environmental conditions. More importantly, current microarray studies are mostly characterised by small sample sizes and high noise levels, several outliers and systematic bias. Results obtained using different chip technologies and experimental procedures from different laboratories are often hardly comparable (Cross-Study normalisation will therefore be considered as a special data integration problem in this thesis, see section “Cross-Study normalisation” below). Even if different laboratories use the same microarray platform, high experimental reproducibility is mostly only achieved within a single laboratory after several months of practical training. Moreover, due to the high costs associated with microarray experiments and the problem of multiple testing [2, 3], the sample sizes are often too small in relation to the number of genes to robustly identify genes with significant differential expression across the sample groups. Therefore, before discussing higher level analysis methods in detail, the next sections of the survey first review the various typical sources of additive and multiplicative noise in microarray experiments, as well as several methods to adjust for bias and filter out noise.

Another major challenge in microarray data analysis, occurring especially when employing the technology to study cancer diseases and genetic disorders, is that these biological conditions often have multiple, combined causes and/or several alternative causes. In addition to information on non-linear relations between the features in transcriptomic data, genetic, epigenetic, proteomic and metabolomic information is often essential to understand the origins of a genetic disease (e.g. for epigenetic diseases like the Prader-Willi [63] and Angelman syndromes [64]). The risk for an individual to develop a certain cancer disease is typically influenced by a complex combination of hereditary risk factors, environmental influences and/or sponta-

neous mutations [65, 66]. This motivates the combination of microarray analysis with other biological data sources, clinical measurements and algorithmic and biological analysis techniques. [3mm]

Especially for basic biological research, integrative analysis techniques are required to obtain a more complete picture of the molecular processes in cells under biological conditions of interest. Therefore, the second part of this survey will not only focus on tailored higher level analysis techniques for microarray data, but specifically on methods that integrate additional data sources and multiple algorithms into the analysis.

### 3.1.2 Microarray data pre-processing

The vast quantities of raw image data obtained from the scanning process during a microarray experiment contain a large amount of uninformative data and do not provide an adequate input for statistical analysis. Several image analysis, filtering and normalisation steps, summarised under the notion *low-level analysis*, should be carried out prior to the final statistical analysis for data interpretation. Since the image analysis procedures, involving image addressing [67], segmentation [68] and background correction [69], require specialised algorithms and sometimes manual adjustments which are not within the scope of this thesis, only the computational methods for the subsequent *quality assessment*, *normalisation* and *summarisation* steps will be discussed in this section.

Though low-level analysis is a common requirement for all microarray experiments, there is no agreement in the scientific community on a single “method of choice” to accomplish this task. Instead, many competing approaches for each step of a microarray analysis have been published in the literature and new approaches are still emerging. Due to the different pros and cons of these methods a frequent recommendation is to test different popular approaches and compare the final results. In the following sections, the general procedure of a low-level analysis will be outlined and some examples given for different approaches to realise each single step in the pre-processing pipeline.

#### Quality Assessment

In order to prevent erroneous biological conclusions from a microarray analysis, low-quality spots and outliers should be filtered out from the data in a preliminary *quality assessment* step. Often already simple visualisations of the raw data using pseudo-colour image representations (also known as “false array plots”), histograms and box-plots allow the experimenter to recognize errors quickly. *MA-plots* (or M vs. A plots), in which the log ratio of the red and green intensities ( $M = \log_2(R/G)$ ) for two-colour microarrays is plotted against the average log intensity ( $A = \log_2(R \times G)/2$ ), can reveal an unwanted dependency of the log ratio on the average log intensity. An example MA-plot for simulated data and created using the software package *limma* [70] is shown in Figure 3.3 (in this case, no unwanted dependency between the M- and A-values can be observed). This plot can also be used as a basis for normalisation methods, making the data for different chips in a study more comparable (see *Normalisation* section below).

Additionally, many microarray software tools judge the quality of single spots using a combination of visual criteria, e.g. spot area, diameter and circularity, using adjustable cut-off values. Often spots are also flagged as unreliable if their pixel values vary too much according to a simple hypothesis test (used for example in the MAS 5.0 and QuantArray software).

McClure and Wit have summarised computationally inexpensive techniques to identify different types of quality problems in order to rectify them or to exclude corresponding data entries [71]. They distinguish

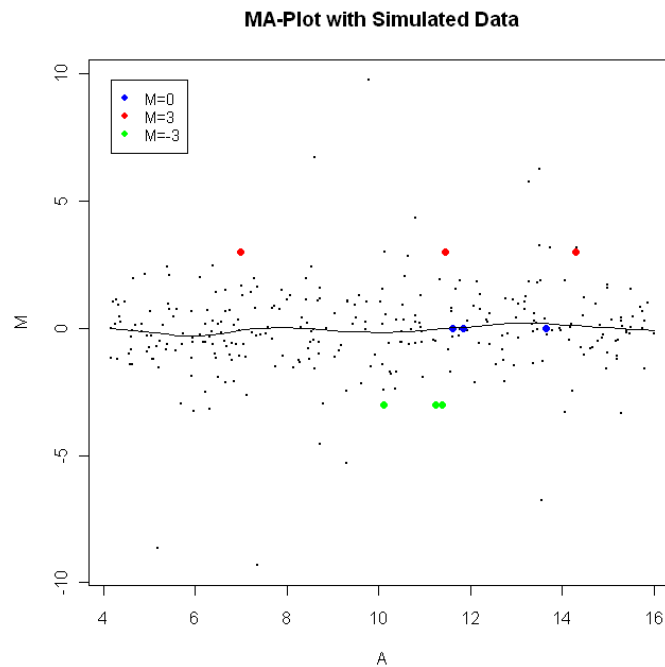


Figure 3.3: Example MA-plot for simulated data and created using the *limma* software package [70]. Microarray spots which do not correspond to the genes to be analysed, but to reference probes for normalisation, are highlighted by coloured points. The continuous line corresponds to a smoothed scatter plot line computed using locally-weighted polynomial regression (LOWESS technique), using a smoother span of 30%, i.e. the smoothness at each value takes into account 30% of all data points.

between four general problem types (simple clerical mistakes, array-wide hybridization problems, normalisation and mishandling problems) and suggest three main techniques to identify them (dimension reduction techniques including visualisation methods, false array plots and correlograms). If the genetic probes have not been arranged in a special order on the array, than a false array plot should normally not exhibit any specific patterns, otherwise such patterns can help to identify technical error sources. Similarly, correlograms, which visualise the correlation between expression level intensities at increasing level of separation between the probes on the array, can help to discover unwanted spatial correlation patterns between neighbouring probes on a DNA chip.

Although most quality assessment tools employ a direct statistical analysis of the data for single microarrays, experimental reproducibility should ideally be tested additionally as one of the most reliable validation measures, especially when establishing a new microarray laboratory. Per-gene replicates on the same chip are not sufficient for this purpose, since they cannot help to detect biases affecting a whole experiment. When repeating a microarray experiment, experienced laboratories reach over 95% reproducibility [72]. Additionally, cross-chip (and some times also cross-study) normalisation has to be applied, because not only single experiments but also single biological samples and even the whole experimental procedure in a laboratory might be affected by systematic biases. These biases mainly result from both technical and biological noise, e.g. due to varying laboratory conditions and the accuracy in carrying out the reverse transcription and amplification, the array hybridisation and the washing steps.

In summary, automatic computer-based quality assessment methods assist the experimenter in identifying a multitude of possible error sources, but the results for each single array should also be inspected manually using visualisations and statistics, because often only the experimenter can determine whether a sample with an outstanding signal intensity distribution represents an outlier or a special biological phenotype.

### Normalisation

A great variety of possible noise sources, which scale and shift the original signal intensities, can influence the microarray-based measurement of gene expression. The main purpose of normalisation is to filter out these effects in order to make different experiments comparable and obtain a reliable estimate of the relative abundance of mRNA for specific genes in the biological samples. Since the normalisation process could also remove important biological information, the goal is to maintain the biological variability in the data, while removing or reducing the variability resulting from the experimental procedure and the limitations of the measurement technology. However, single normalisation methods typically do not account for all known sources of noise, and even for specific, partial normalisation tasks no single approach has been accepted as a standard by the scientific community. Therefore, a suitable combination of normalisation methods should be chosen from the wide spectrum of available approaches in the literature.

The classical microarray normalisation methods use a selection of genes with nearly invariant expression levels across all tissues and experimental conditions (so-called *House-keeping genes (HKGs)*) as a control to adjust the intensity values on all arrays. However, this method proved to be ineffective, as over the years more and more special cases were identified, in which *HKG* expression levels varied significantly across different conditions. Nevertheless, attempts have been made to use other selections of control genes or to improve other normalisation methods by restricting the analysis to a specific selection of genes [73].

The more widespread *global normalisation* methods use information from the entire dataset in order to make different microarray experiments in the same laboratory comparable. These methods are mostly requiring the assumptions that the majority of genes on an array platform is not differentially expressed across the studied conditions, the number of data points is large and that there is an approximate relative symmetry between up- and down-regulation of genes. Global normalisation methods are only suitable, if no local print-tip block effects or intensity effects occurred during the microarray experiment [74]. The simplest global method for normalisation is *linear scaling*, where a baseline array is chosen and the intensities of the other arrays are scaled until all arrays have the same mean intensity as the baseline array. However, in most cases non-linear methods, e.g. using *scatter plot smoothers*, are more appropriate due to the non-linear nature of various types of noise in microarray experiments.

Many global normalisation methods can also be applied locally, i.e. on a physical subset of the data. *Local normalisation* is often used for groups of genetic probes which were deposited by the same spotting pin, when the array was printed, in order to correct for systematic spatial variation on the array.

*Signal-dependent normalisation* (or intensity-dependent normalisation) is the generic term for methods accounting for the intensity-dependence of gene expression ratios, which are typically larger in the low-intensity range. Using a MA-plot (see previous section on Quality Assessment) the distribution of differentially expressed genes across the entire intensity range can be made uniform by subtracting a local linear or polynomial regression fit from the data using the non-parametric methods *LOWESS* (= locally weighted scatter plot smoothing [75], see an example in the previous MA-plot in figure 3.3) and *LOESS* (= locally estimated scatter plot smoothing [76]). This approach can be applied both on the colour channels in a single two-colour array [77] and on the probe intensities from two distinct arrays. Again, applying these methods only locally often provides improved results, i.e. the corresponding “Subarray LOWESS” method typically outperforms global LOWESS.

In recent years, several algorithms have been developed to attain more stringent normalisation properties. For example, *Quantile normalisation* [69] and *Qspline normalisation* [78] attempt to make the empirical

distribution of intensity values similar for all chips using the assumption that the distribution of the real gene abundances is approximately the same in all samples. In quantile normalisation, for each expression level on each chip the corresponding quantile in the pooled distribution of probes on all chips is computed and the original value is transformed into that quantile's value on a given reference array. The more recent Qspline method also computes signal intensity quantiles, but both on the arrays' signal intensities and on so-called "target intensities", which can be derived either from another reference array or calculated as averages from multiple arrays. The quantiles are then used to fit B-splines, i.e. piecewise polynomial functions with minimal support with regard to chosen properties like degree, domain partition and smoothness, which have been shown to adequately estimate smooth relationships [79]. By using quantiles instead of the original values, the fitting problem is greatly simplified. This does not only increase the processing speed but also reduces the risk of over-fitting.

More recent microarray normalisation methods use mathematical error models consisting of a combination of additive error-terms, where each term attempts to model the contribution of one specific error type. Commonly, microarray error models assume that the measured intensity ( $y$ ) corresponds to the sum of a scaled version of the real intensity  $x$  (multiplicative noise) and an offset value (additive noise):

$$y_{ik} = a_{ik} + b_{ik}x_{ik} \quad (3.1)$$

where index  $i$  refers to the sample, i.e. the array, index  $k$  refers to the genetic probe,  $a$  is the offset and  $b$  a scaling normalisation factor (also known as "gain") [80].

This error model can be refined by splitting up the offset and gain into subcomponents corresponding to different types of error sources. More precisely, the offset can be written as the sum of a per-sample offset  $a_i$ , depending only on the array and not on the probes, and an additive noise component  $\epsilon_{ik}$  assumed to have a normal distribution ( $\epsilon_{ik} \sim N(0, b_i^2 s_1^2)$ ). The gain contains multiplicative sources of noise,  $\eta_{ik} \sim N(0, s_1^2)$ , as well as factors depending only on the sample (per-sample normalisation factor  $b_i$ ) or only the genetic probes (sequence-wise probe efficiency  $b_k$ ). On the whole, the following refined error model is obtained:

$$y_{ik} = a_i + \epsilon_{ik} + b_i b_k \exp(\eta_{ik}) x_{ik} \quad (3.2)$$

With this error model it has been shown that there is a dependence of the variance of the measured intensities on the expected intensity values [80, 81] (using the linearity of the expectation and  $E(\epsilon_{ik}) = 0$ ):

$$E(y_{ik}) = E(a_i) + E(\epsilon_{ik}) + E(b_i b_k \exp(\eta_{ik}) x_{ik}) \quad (3.3)$$

$$x_{ik} = (E(y_{ik}) - a_{ik}) / (b_i b_k E(\exp(\eta_{ik}))) \quad (3.4)$$

$$x_{ik}^2 = \frac{(E(y_{ik}) - a_{ik})^2}{b_i^2 b_k^2 E(\exp(\eta_{ik}))^2} \quad (3.5)$$

$$\text{Var}(y_{ik}) = E(y_{ik}^2) - (E(y_{ik}))^2 \quad (3.6)$$

Insert 3.3 into 3.6:

$$\text{Var}(y_{ik}) = E(\epsilon_{ik}^2) + b_i^2 b_k^2 x_{ik}^2 \text{Var}(\exp(\eta_{ik})) \quad (3.7)$$

Insert 3.5 into 3.7:

$$\text{Var}(y_{ik}) = E(\epsilon_{ik}^2) + b_i^2 b_k^2 \left( \frac{(E(y_{ik}) - a_{ik})^2}{b_i^2 b_k^2 E(\exp(\eta_{ik}))^2} \right) \text{Var}(\exp(\eta_{ik})) \quad (3.8)$$

$$\text{Var}(y_{ik}) = E(\epsilon_{ik}^2) + \frac{\text{Var}(\exp(\eta_{ik}))}{E(\exp(\eta_{ik}))^2} (E(y_{ik}) - a_{ik})^2 \quad (3.9)$$

Equation 3.9 reveals that the variance of the measured intensities is a quadratic function of the expected intensity value. This undesired effect can be compensated by a dedicated normalisation method, the *Variance-stabilizing transformation* (VSN), introduced independently by Huber *et al.* and Durbin *et al.* [80, 81]. The VSN is an integrative normalisation method that combines several low-level analysis steps within one procedure. It performs both background adaptation and normalisation, using the information of all chips to estimate background adaptation parameters, instead of considering each array separately. However, the main benefit of the VSN method is the capacity to filter out intensity-dependent variance. Depending on whether the error model includes a bias offset, additive or multiplicative noise or a combination of these terms, different variants of the VSN method can be derived. In the most general case of additive and multiplicative noise, as in the error model shown above, variance stabilization is achieved by the generalized logarithm (or glog-) transformation, a generalized form of the logarithm and the inverse hyperbolic sine function (arcsinh). This function makes the variance across replicates independent from the mean signal intensity:

$$\text{glog}_c(x) = \ln \left( \frac{x + \sqrt{x^2 + c^2}}{2} \right) \quad (3.10)$$

$$\text{vsn}(x_{ik}) = \text{glog} \left( \frac{x_{ik} - a_i}{b_i} \right) \quad (3.11)$$

where  $a_i$  is a background offset and  $b_i$  the scaling parameter for array  $i$ . In the special case of  $c = 4$ , the glog-function corresponds to the arcsinh-function, and for  $c = 0$  the natural logarithm function is obtained. Glog-transformed data are linear at the low intensity values (to compensate for the higher variance) and logarithm-like at high intensity values. The parameters can be fitted by means of a maximum-likelihood procedure.

More recently developed normalisation methods additionally use sequence-specific probe affinities (e.g. GC-RMA [82]) and account for non-specific hybridization of the target DNA to be analysed with the probe DNA that is immobilised on the microarray chip (e.g. BGX [83] and PUMA [84]). Some methods even attempt to model the hybridization process directly, e.g. by physicochemical models for a specific array type [85], but most of these models are not yet fully recognized by the scientific community.

Since an exhaustive discussion of normalisation methods is beyond the scope of this thesis, the approaches discussed above are only a small representative selection of the available approaches, highlighting the types

of problems that are addressed by most of the commonly used methods. A more comprehensive overview on widely accepted, classical normalisation methods can be found in a review article by John Quackenbush [86].

### Summarisation

If replicate genetic probes are present on the chip platform used in a microarray study, these replicates can be used to filter out noise and obtain more robust point estimates. For this purpose, the normalised signal intensities of the replicates representing one genetic probe have to be *summarised* to a single value. Even for this seemingly simple task of combining multiple probe intensities to a single expression value, a variety of alternative methods exist. However, these methods are often already part of common integrative approaches, which provide summarisations that match well with the included background adaptation and normalisation methods.

For single-array replicates, the most common summarisation approaches simply compute the average or median of the logarithm (or vice-versa) for all probe intensities. Additionally, the maximum and minimum values are often excluded to avoid influences of outliers on the final expression value. Alternatively, instead of simple averaging *Tukey's Biweight Algorithm* [87] can be used to obtain more robust averages, by first determining the median and then its absolute distance to each data point, in order to assign weights for the contribution of each intensity value to the average (far-off outliers obtain low weights). The process can be applied iteratively, but typically a 1-step Tukey Biweight is considered as sufficient. Indeed, most of the current popular summarisation methods use weighted sums of the probe set values like in Tukey's Biweight algorithm and only the methods to estimate the weights differ, e.g. *Li and Wong* [88] suggest weighted sum conditional least squares estimates and *Lazardis et al.* [89] propose a non-parametric method functionally related to leave-one-out cross-validation, which estimates and weights the performance of each single probe intensity value estimate by comparing it against an estimate derived from the other replicate probes. For multi-array replicates, typically a linear model fit (optionally using M-estimation [90]) is used for summarisation. A commonly used alternative is the *median polish* method proposed by John Tukey [91], which fits an additive model using an iterative algorithm to increase model robustness.

In brief, summarisation methods typically employ robust averaging techniques, using different approaches to achieve high robustness with regard to outliers.

### 3.1.3 Protein interaction data pre-processing

Molecular interaction data and specifically protein-protein interactions, corresponding to physical binding reactions in which both interaction partners are proteins, belong to the most frequently used data sources in computational systems biology analyses and integrative bioinformatics methods. In this section, data pre-processing methods required to build a protein-protein interaction network using publicly available data are reviewed, focussing on their opportunities and limitations in integrative analysis approaches for high-dimensional biological datasets.

The pre-processing and filtering of protein interaction data is typically less complex and time-consuming than for other high-throughput data sources like microarray data. In contrast to the multitude of low-level image analysis and normalisation steps required in a typical microarray study, pairwise interactions only need to be filtered and categorized using simple criteria like the experimental method(s) employed to verify



the interaction and different types of confidence scores. However, regarding the reliability of the data, the assembly of protein interaction networks is affected by similar limitations as other biological datasets obtained from high-throughput technologies. Public interaction databases, including BioGRID [92], MIPS [93], DIP [94], MINT [95], HPRD [96] and IntAct [97], are known to contain a significant amount of false positives and false negatives. These erroneous data entries and other limitations result from a variety of problems in the experimental validation of protein-protein interactions, in particular:

- Some interactions only take place in the proteins' native compartments, which might not be the compartments assessed in the detection experiment.
- Several experimental methods are biased to detect only specific types of interactions, a *coverage bias* that reflects different strengths and weaknesses of different methods [98]. For example, transcription factors cannot be studied with the Yeast-2-Hybrid (Y2H) technique, because the corresponding fusion hybrids could activate the transcription of a reporter gene even without an interaction. Similarly, methods using purified protein complexes tend to underestimate the number of interactions for proteins involved in transport and sensing [98].
- Common experimental techniques like Tandem Affinity Purification (TAP) or Co-Immunoprecipitation (CoIP) cannot detect direct binary interactions between proteins, but only co-occurrences within the same complex
- The application of some methods (e.g. Y2H) might inadvertently lead to conformational changes in the proteins of interest, and thus prevent a correct detection (increasing the number of false negatives).
- Only few techniques, like the Yeast-2-Hybrid (Y2H) method, also detect unstable and transient interactions, but do not always enable the experimenter to distinguish these from other more stable interactions.
- Interactions which are predicted as being possible, do not necessarily occur in the real physiological setting.

In summary, the current data sources provide only a simplified representation of the real networks of protein interactions in living cells, mostly ignoring the differences between different tissues, diverse cellular compartments and transient vs. stable interactions, and containing several type 1 and type 2 errors.

To reduce the number of false positive interactions when assembling the data for a large-scale protein-protein interaction network, the first pre-processing step after collecting the unfiltered data for the species of interest, is typically a simple filtering, taking into account the experimental methods used to verify the interactions. Generally, interactions that were verified by multiple independent experiments and different technologies can be considered as more reliable than those confirmed by only a single experiment. However, the overlap between the datasets from different experiments has been shown to be very small in many cases [98]. A further important filtering criterion is the type of experiment: Interactions that were only verified by methods detecting co-occurrences within the same protein complex, e.g. TAP and CoIP, should in most cases be excluded from the analysis, unless the investigator wishes to build a network representing general "functional associations" between the proteins rather than direct physical interactions. To facilitate these filtering tasks, several public databases provide their molecular interaction data in the HUPO PSI-MI data exchange format [99], including standardized names for the experiments (e.g. *two-hybrid* experiments have the PSI-MI code "MI:0018"). As part of the data preparation for this thesis, a PSI-MI parser was written in the scripting language Python; however, for various databases not supporting this standard, separate dedicated parsers had to be implemented additionally.

Improved filtering methods rely on *confidence scores* in addition to the information on the experimental methods. These scores can also be used to assign weights to the interactions instead of filtering low-confidence edges out. In graph-based representations of the data, where nodes correspond to proteins and edges to interactions, these interaction weights become edge weights, which can be used for distance computations in the network, e.g. to be used in algorithms to identify dense communities of nodes in the network. A multitude of validation methods and confidence assignment schemes have been presented in the literature. They are using several independent confidence measures, including

- the functional similarity between interacting proteins, measured by the semantic similarity between corresponding functional terms in the Gene Ontology [100] biological process graph [101–103]
- the gene neighbourhood method [104,105], predicting proteins as functionally related and more likely to interact in the species of interest, if their genes are found to assemble in putative operons and in a conserved gene order in different orthologs (homologous genes in another species)
- the analysis of gene fusion data, considering the proteins for genes whose orthologs are fused in another organism as more likely to interact [106–108]
- the identification of orthologs which are known to interact in other species can point to an interaction of their homologous proteins in the species of interest, if the interaction is conserved across many species (these conserved interactions are also known as “interologs”) [109]
- the comparison of phylogenetic profiles, i.e. fingerprints for the co-occurrence of genes/proteins across different species stored in binary vectors, which can predict genes with similar function and proteins more likely to interact from the similarity or complementarity of these binary profile vectors [110,111]
- co-expression of genes found in microarray, RNA sequencing or qPCR data, increasing the likelihood of the corresponding proteins being interaction partners or functionally associated [112]
- correlation with localization, since proteins will only interact if they are co-localized in the same cellular compartment [113]
- correlated gene mutations between interacting protein families, which can point to conserved binding sites for interactions [114]
- topology-based scores, analysing the local topology in the network surrounding the interaction of interest, e.g. using the number of non-shared interaction partners [115] or the Czekanowski-Dice similarity and its iterative extensions [116,117]
- text-mining based scores, e.g. taking into account the publications in which an interaction is reported (care must be taken when using automatic methods, e.g. searching for phrases like “A binds to B” or “A interacts with B”, because the context might not always be a supportive statement, e.g. “...these results do not support the conclusion that *A binds to B*”) [118]

Typical approaches combine several of these scoring approaches into ensemble probabilistic scores (see [119] for an example). However, the information required to compute these scores is not always available for all interactions, especially when analysing non-human species; hence, in many cases only unweighted networks are built or only topology-based weights are computed. Depending on the distance measure used in the analysis of the protein-protein interaction (PPI) network, e.g. the discrete shortest path distance or continuous kernel-based or random walk distances, a weighted network might not always be required as

input, though weights that accurately reflect the significance of edges are likely to provide more sensitive results in statistical analyses.

Importantly, when building a network from individual interactions, the final protein-protein interaction network obtained after a filtering procedure does not necessarily consist of a single connected component, which is required as input by some integrative analysis methods. However, in most cases, the output contains one large connected component (containing several thousand nodes) and only few and small additional components (containing less than a few dozen nodes), so that the small components can be removed without losing a significant amount of information.

In order to increase the number of interactions which are correctly covered by an assembled PPI network, the true positives, some studies have also included *in silico* predicted protein interactions obtained from bioinformatics methods taking into account both structural information [118, 120] and the data sources for inference of functional associations mentioned above. The prediction methods have been evaluated using large reference sets of known interactions, and several published methods reach high test set accuracies [121]. However, the reference sets used for model training are incomplete and sometimes biased (low-abundance proteins are typically neglected), and even if only high-confidence predictions are added to a PPI network, the coverage is usually increased at the expense of a higher number of false positives. Therefore, the choice of the methodology for constructing PPI networks, and thus the choice of the coverage/accuracy trade-off, should depend on the purpose of the higher level analysis method: For predictive tasks relying more on high coverage than local accuracy, the inclusion of predicted interactions might improve the performance, but for methods which aim at the biological interpretation of local structures in the network, only experimentally verified interactions should be used in the input.

Since PPI networks are scale-free, i.e. their node degree distribution follows a power-law [122], the scale-free property enables the estimation of the percentage of covered protein interactions using a known estimate of the percentage of covered proteins from the whole proteome of a studied species. These estimates can help to guide the decision on whether methods to increase the coverage (see above) should be applied, or whether the current coverage is sufficiently high.

Example integrative bioinformatics methods using protein interaction networks as data sources will be discussed in the higher level analysis part of this survey (see below) and in chapters 6 and 7.

### 3.1.4 Gene/protein name standardisation

One of the most common data pre-processing tasks in the integrative analysis of functional genomics data, is the standardisation of gene and protein names, i.e. the mapping of different identifier formats onto a single, standardised format. This task is complicated by a multitude of naming conventions, which differ across diverse biological disciplines and partly even contain ambiguities. These naming ambiguities include cross- and intra-species ambiguity (e.g. the name “CAT” stands for different genes in cow, chicken, fly, human, and other species), and ambiguities with general English words (e.g. the mouse gene “hair loss” or the fly gene “lie”) and medical terms (e.g. the mouse gene “diabetes”, see [123] for more details and examples). Moreover, further difficulties arise from several spelling variations used in the literature, which do not adhere to any standards (e.g. “ABC-1” or “ABC\_1” instead of “ABC1”).

In high-throughput data analysis, the situation is even more difficult, because the molecular identifier formats do not only differ across the diverse branches of biology, but even across single experimental platforms. For example, genetic probes on a microarray chip are typically named using platform-specific identifiers.

Often some of these probes correspond to unknown genes with missing annotations and cannot be mapped onto any standard gene identifier. Furthermore, in many cases genes are represented by multiple probes on a chip; hence, it might be necessary to summarise the expression values for these probes for one gene into a single value per sample (see *Summarisation* section above). These issues are not limited to genetic data, but similar problems occur in the analysis of protein expression data.

The scientific community has recognised the problem of different and complex naming conventions, and for most of the recent experimental platforms, mappings of the platform-specific identifiers to unambiguous, standardised identifiers like the ENTREZ [124] or ENSEMBL [28] gene name format are directly provided with the analysis software. However, the situation becomes more complex if the experimenter aims at integrating data from very diverse and partly unstructured data sources, e.g. combining experimental data with cellular pathway and process data, molecular interaction networks and information from the literature. In this case, public gene and protein annotation databases typically only contain the necessary mapping information for a subset of the identifiers occurring in the input data.

To address this problem, several gene/protein name conversion tools and web-services, which integrate data from multiple public databases and recognise simple spelling variants, have been published in recent years. These include the DAVID functional annotation web-service and database [29], the CNIO IDconverter web-application [125], the MatchMiner database and web-tool [126] and the R/bioconductor software package “biomaRt” [127,128], among others. For some applications, e.g. when considering data for a special species or using articles extracted from the bioscientific literature within an integrative analysis method, even these dedicated gene/protein name conversion services cannot provide sufficiently large dictionaries to obtain enough correct mappings using both exact and approximate string matching. For these situations, text-mining based systems for an automatic creation of synonym dictionaries have been proposed [129–132]. These approaches require both algorithms to address the task of finding gene/proteins mentions within full-text articles, known as *named entity recognition* (NER), and to normalise the resulting gene/protein lists. In contrast to the gene/protein name conversion web-services, these methods do not only use approximate string matchings but also apply several string transformations. In some cases, they are also capable of taking contextual information into consideration [133]. As part of the BioCreative competition [132,134], a comparative evaluation of a multitude of corresponding techniques has been conducted for different model organisms, with the highest ranked systems reaching a balanced 80% precision / recall or better. The current approaches include machine learning methods [135–137], extensions of string matching using automatically extracted dictionaries [138], rule based approaches [139] and integrative approaches combining multiple methods together [140].

In the future, increased efforts to ensure the use of standardised identifiers for new experimental platforms and to improve the annotation of the bioscientific literature (e.g. annotating the PubMed database with MeSH keywords [141]) are likely to reduce the impact of the gene/protein normalisation problem significantly.

## 3.2 Higher Level Analysis - Introduction

When all required pre-processing procedures have been applied to the data from a high-throughput biological experiment, the resulting input data for all subsequent analysis steps is typically a continuous-valued matrix with columns corresponding to the samples and rows corresponding to genes or proteins (or genetic probes, in the case of microarray gene expression data). Using this input, and optionally additional

biological data for an integrative analysis, the higher level analysis methods discussed in the following sections mainly aim at the identification of important features (feature selection), the discovery of informative patterns and structures in the data (clustering), or the creation of predictive models (classification and regression) for the biological interpretation of the data. Moreover, integrative extensions and modifications of these analysis types using additional biological data and algorithms from graph/network analysis, optimisation and literature mining, can help to better understand the molecular processes of interest.

Depending on the biological question behind the experiment, the selection of the general type of analysis method is often straightforward. In most cases *unsupervised learning techniques* should be applied when patterns and structures like natural groupings of features (genes, proteins, etc.) or samples are to be identified in the data, and *supervised learning* techniques should be used, when labelled training data is available and individual samples or genes are to be classified into one of several given biological categories. Typical examples for the output generated by unsupervised learning methods are clusterings of samples to identify tumour subtypes or clustering of features (genes, proteins or gene/protein groups) to identify molecules with similar changes in the measured abundance across different biological conditions. Examples for supervised learning results are the predicted memberships of new samples to certain sample groups (e.g. “tumour” or “healthy”) or the predicted membership of genes to groups of functionally related genes. Another supervised analysis task is the *identification and selection of differentially expressed genes/proteins* to find molecules whose abundance differs significantly across different sample groups and which might therefore be functionally related with the biological conditions of interest. If these conditions are disease-related and, ideally, a large-scale experimental validation confirms the utility of certain molecules in the dataset as biomarkers, this type of analysis can support the development of new assays for disease monitoring and diagnosis.

Apart from these classical analysis types, new integrative methods which take additional biological knowledge into account, e.g. to analyse the data in the context of cellular pathways and interaction networks, will be considered in the last part of this survey.

### 3.3 Dimensionality reduction and feature selection

Dimensionality reduction and feature selection approaches are particularly useful for the analysis of datasets obtained from high-throughput experiments like microarray gene expression profiling. For these datasets the number of features (the matrix rows corresponding to genes or proteins) typically exceeds the number of samples (the matrix columns corresponding to different biological samples) by approx. two orders of magnitude, and a majority of the features is uninformative with regard to the biological question of interest. Thus, in addition to the pre-processing methods discussed in the previous section, dimensionality reduction methods are often also applied to identify or extract informative features and decrease the number of features. Moreover, dimensionality reduction can also profitably be combined with integrative analysis techniques, e.g. by grouping functionally related genes or proteins together using additional biological information before applying the reduction (see section on “Integrating Cellular Pathway Data” below and the *TSSP* approach in chapter 7).

Here, two of the main types of dimensionality reduction methods will be reviewed in separate sections: Unsupervised approaches, including both feature-preserving methods and feature extraction methods, and supervised approaches, again including methods retaining the original feature definitions (also known as feature selection methods) and supervised feature extraction methods. Supervised feature selection in

particular, is a fundamental task in biological data analysis, as it enables the identification of diagnostic biomarkers (genes, proteins or metabolites) corresponding to predictor attributes in the data, whose abundance values enable a discrimination between different biological conditions of the samples. For this reason, supervised selection methods will occupy the main part of this section, whereas unsupervised dimensionality reduction will be discussed first, but in less detail. A general overview of the feature selection methodologies that will be the main focus in the following sections is given in figure 3.4.

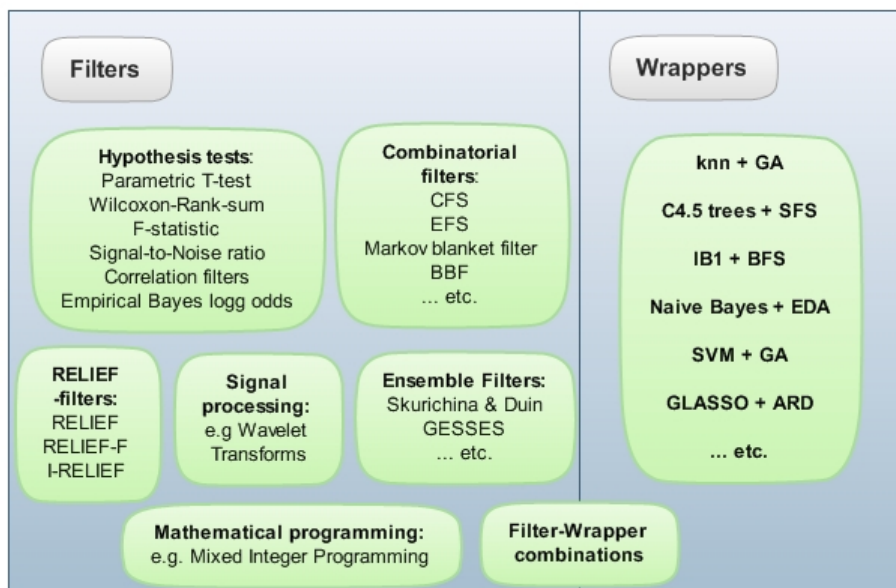


Figure 3.4: Overview of generic feature selection methodologies and example algorithms, consisting of univariate and combinatorial filters (left) and wrapper approaches (right).

### 3.3.1 Unsupervised filtering

Unsupervised dimensionality reduction approaches are often considered as pre-processing methods rather than higher level analysis techniques. However, in the case of high-dimensional datasets, these methods can already enable a first interpretation of the data using low-dimensional visualisations, often revealing biologically meaningful groupings and other structures and patterns in the data.

Nevertheless, the main motivation behind these methods is to alleviate computational and statistical problems in the analysis of the data, summarised under the notorious expression *curse of dimensionality* [1]. This term describes data analysis problems related to the fact that the mean Euclidean distance between points uniformly distributed in an  $n$ -dimensional hypercube as well as the mean distance from the centre to the closest data point increase exponentially with the dimension  $n$ . Thus, if one considers a local, supervised learning method like the  $k$ -Nearest neighbour approach applied to a data set with a uniform distribution of data points, then for a higher dimensionality of the data set, the expectation is that exponentially more training data points are required in order to find the same number of observations within a certain distance of the query point [142]. Moreover, in higher dimensions, extrapolation will be required more often than interpolation to make predictions, since uniformly distributed data points are expected to be closer to the boundaries than to the centre of the input space. Although in microarray experiments the distribution of data points in the feature space is often significantly different from the uniform distribution, this does not always alleviate the dimensionality problem. Often some data points are clustered together locally, which

simplifies the identification of groupings, but this advantage is compensated by a substantial degree of noise masking the real intensity values.

Moreover, due to the high number of genes/proteins in most expression datasets, some of them might appear to be significantly differentially expressed in a subsequent supervised feature selection by mere chance, if the hypothesis test used to select the features is not adjusted for the number of tests (an issue known as the “multiple testing problem” [2,3]). Dimensionality reduction can alleviate this problem by either computing a small number of derived features (feature extraction), covering the majority of the variance in the data, or by removing features considered as uninformative by a filtering criterion (unsupervised filtering).

Unsupervised filtering can, for example, remove microarray genes which display a low variance and/or low expression values across the samples in a dataset and are therefore regarded as uninformative. Early microarray filtering approaches eliminated genes with a small number of two-fold differences to the mean expression value [143], a small difference between the maximum and minimum expression value, or a small variance across the samples [144]. However, for some functionally important regulator genes, small expression level changes might still have significant biological effects, hence, these genes should not be filtered out. Accordingly, recent analysis methods try to circumvent this filtering step or replace it by more advanced techniques. These include a recently published parameter-free filtering method by Tritchler *et al.* (COVSUM [145]), which avoids the removal of regulator genes with small variance in expression values, by filtering genes depending on their covariance with other genes, rather than only their variance. This method, which uses the sum over the absolute values in rows of the covariance matrix as the filtering criterion, and replaces the manual threshold selection by a two-group clustering method, has been shown to perform well for higher level analysis tasks on both simulated and real-world data.

A further alternative to simple filtering methods, is to select the most informative features in terms of their contribution to the overall variance in the data, using information from feature extraction methods like Principal Component Analysis (PCA). For example the *sparse PCA* method [146], which uses sparse vectors as weights for a linear combination of the original variables, can help to remove genes with small information content, by filtering out the genes with corresponding zero-entries in the first sparse principal component vector. Other dimensionality reduction methods relying fully on feature extraction will be discussed in detail in the section on unsupervised learning algorithms (see below).

Since class labels are available for most microarray datasets, more advanced *supervised* filtering methods are often applied during a feature selection analysis (see next section). Thus, in current analysis approaches, unsupervised filtering is mainly used as a pre-processing method for high level unsupervised analysis methods like clustering and network analysis and often completely circumvented for supervised analysis tasks. Therefore, in the microarray data analysis software ArrayMining, from the integrative framework developed as part of this PhD project, the sparse PCA and the COVSUM method have been integrated into the unsupervised analysis modules, whereas other dimensionality reduction techniques are used for the supervised analysis modules.

### 3.3.2 Supervised feature selection

Using only unsupervised filtering methods, the number of features in high-dimensional biological data can typically not be reduced sufficiently to avoid statistical problems associated with the “curse of dimensionality”, and to obtain a compact and interpretable set of features as input for predictive model building. Therefore, if labelled training data is available, supervised feature selection methods are invaluable both

for the interpretation of the data and to find the most informative features in classification and regression models. In this section, a brief introduction on the importance of feature selection for improving the interpretability and complexity of predictive models in terms of finding an optimal bias/variance-trade-off will be given, and then the two main types of feature selection methods will be discussed in detail: *Filters*, relying on direct statistical measures of the informativeness of features, and *wrappers*, employing a search algorithm and a statistical learning method to score feature subsets by evaluating predictive models built on them.

Although a multitude of genes/proteins can be significantly differentially expressed across different biological conditions, predictive models containing only a small set of highly significant features rather than all statistically significant features are often considered as preferable, because these models are easier to interpret and the risk for model overfitting is reduced. Overfitted prediction models typically have a very low bias and training error, but tend to display a higher variance resulting in a high generalization error and poor test set performance. The objective of creating parsimonious models, also matches with the concept of “Occam’s razor” [147], according to which simpler explanations of a phenomenon are always preferable to more complex models, if there is no additional information favouring a complex model. Applying a gene selection algorithm to obtain a reduced subset of informative genes decreases the complexity of a prediction model built upon these features and will also reduce the generalisation error, if the simplified model provides a better trade-off between the bias and variance contribution to the error (also known as the *bias-variance trade-off* (see Fig. 3.5 for an illustration) [6].

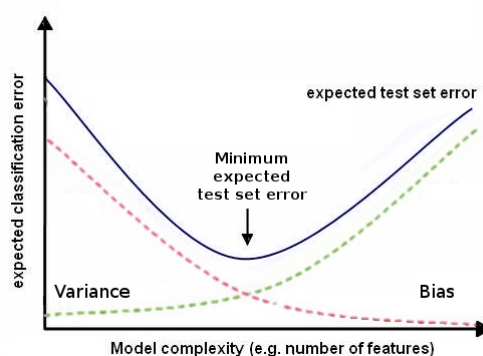


Figure 3.5: Illustration of the bias-variance trade-off in statistical learning. Supervised feature selection can help to reduce model complexity to obtain an optimal balance between bias and variance that minimises the expected test error [6].

Although for the sake of structural clarity, feature selection and classification methods will be discussed in different sections, from an algorithmic perspective, several studies suggest that applying feature and model selection jointly provides better prediction results [148]. Therefore, after first reviewing supervised filter methods, which are independent from prediction methods, wrapper-based methods combining feature selection and prediction will be presented in a dedicated section.

### Filter methods

Earlier microarray studies often made use of filter techniques, which try to identify informative subsets of features using simple empirical statistical measures and hypothesis tests, without directly evaluating their predictive power.



**Univariate filters:** The most traditional feature selection methods are parametric and non-parametric univariate filters, including the *parametric student's t-test*, the *non-parametric Wilcoxon-Rank-sum statistic*, the *F-statistic*, the *signal-to-noise-ratio*, the *fold change* and the *correlation* to the response variable (using different association measures like the Pearson correlation, Spearman's rho, Kendall's tau, mutual information, the normalised compression distance [149] or combinations of different measures). A more recently introduced test statistic tailored specifically for microarray gene expression analysis is the *empirical Bayes log odds' test statistic* [150], which uses a shrinkage technique to shift the gene-wise sample variances towards common values, reducing the effects of outliers (a moderated variant of this statistic additionally obviates the need for estimating certain meta-parameters of the method [151]).

Most of these hypothesis tests used in filter methods require simplified assumptions, e.g. a parametric student's t-test relies on the assumption that the data is generated by sampling from normal distributions for the features with common variance and tests the null hypothesis that the means are equal.

Importantly, when filtering a large number of individual genes using these methods, the significance criterion needs to be adjusted for multiple testing [2] in order to restrict the expected number of false positives. Corresponding methods requiring the assumption that the statistical tests are independent from each other include the classical *Bonferroni* method [152], where the size of the allowable error  $\alpha$  is simply divided by the number of comparisons, and further developments like the *Bonferroni-Holmes* [153] approach and the more recently introduced *Benjamini-Hochberg* method [3], which enables a direct control of the expected false discovery rate (FDR). Moreover, if the tests are assumed to be interdependent, the *Benjamini-Hochberg-Yekutieli* procedure [154] provides an alternative means to control the expected FDR.

**Margin-based feature selection:** Using simple univariate filters in combination with multiple testing adjustments can already provide an experimenter with a wealth of statistically significant and biologically meaningful information. However, especially when analysing complex genetic diseases, functional relations and interactions between multiple genes/proteins often influence the observed phenotype of the samples, and univariate methods cannot account for these interdependencies between different features. More recently, several *margin*-based and combinatorial filters have therefore been developed. The margin is a confidence measure for a classifier with regard to its decision on a single sample by computing the distance from the corresponding data point to a decision boundary [155]. Margin-based feature selection methods try to weight each feature to receive a maximal margin, and corresponding weighting methods have been shown to be more robust against dependencies between attributes than the ranking scores of classical univariate filters [156]. Other filters, which directly score the combined information from multiple features on the outcome variable, or explicitly remove redundant features from the data, will be discussed in the following section. Importantly, however, the capacity to consider global information or model feature dependencies typically comes at the expense of a reduced speed and scalability in comparison to classical univariate methods.

Among the margin-based filters, an example for one of the most popular methods in microarray analysis is the RELIEF method [157] which iteratively adjusts feature weights according to the features' capacity to distinguish between the distances of randomly selected samples to their nearest neighbour with a different class label ("miss"-sample) and the nearest neighbour with the same class label ("hit"-sample). Since the classical RELIEF method is limited to two-class data without missing values, and is strongly affected by noise, several extensions have been proposed (alphabetically named RELIEF-A, RELIEF-B, etc. up to RELIEF-F [158]) which attempt to overcome these restrictions. The most recent extension, I-RELIEF, additionally addresses the problem that in the original RELIEF algorithm the nearest hits and misses for a

sample are defined in the original feature space and not in the new weighted feature space [159].

In summary, margin-based feature selection methods are particularly useful in cases, when strong dependencies are observed between different attributes, but the direct consideration of these dependencies using combinatorial filters or wrappers (see next sections) would lead to a combinatorial explosion.

**Combinatorial filters - (2) penalization and removal of redundancy:** An alternative approach for taking feature-dependencies into consideration is to penalize redundant features, while searching through the space of feature subsets rather than considering individual features independently. Hall's *Correlation based feature selection* (CFS) [160, 161] was one of the first attempts to exploit this idea, by identifying subsets of features with a high average correlation to the response variable but low correlations amongst each other. This concept is formalised by the following feature subset score:

$$CFS(S) = \frac{k\overline{c_{rf}}}{\sqrt{k+k(k-1)\overline{c_{ff}}}} \quad (3.12)$$

where  $S$  is the selected subset with  $k$  features,  $\overline{c_{rf}}$  is the average feature-class correlation and  $\overline{c_{ff}}$  the average correlation between feature pairs in  $S$ . While the denominator reduces the score for correlated features to eliminate redundant variables (minimising redundancy), the numerator promotes features with high correlation to the class variable in order to retain them as discriminative predictors (maximising relevance). For discrete features the *symmetrical uncertainty* can be used as a correlation measure, whereas for continuous features usually the Pearson correlation is chosen as association measure. The CFS scoring function can in principle be used in combination with any search algorithm to explore the space of feature subsets; however, the original CFS publication proposed a fast greedy best-first search strategy as the method of choice [161].

A similar approach, explicitly removing redundant features, was introduced by Ding and Peng [4] as their *minimum redundancy - maximum relevance* (MRMR) selection framework. They defined different measures for the relevance and redundancy of features using the mutual information (for discrete attributes) or the F-test (for continuous attributes). For example, in the discrete case, the minimum redundancy is defined as the minimum average mutual information across all pairs of features in a subset, and the maximum relevance as the maximum average mutual information between the features and the outcome variable (i.e. the known biological conditions for the samples, encoded as an ordinal variable). Using a greedy search procedure and MRMR selection in combination with an SVM and a Naive Bayes classifier, high average LOOCV accuracies were obtained on two binary-class and three multi-class microarray cancer datasets.

**Combinatorial filters - (3) information theoretic measures:** After the introduction of the CFS method, other research groups aimed at providing improved combinatorial filters by employing measures from information theory in the feature subset scoring function. For example, the *balanced information gain* ( $B_g$ ) [162] was proposed as an alternative to the CFS score to estimate the contribution of a feature for class separation:

$$B_g(f_m) = \frac{I(C|f_m)}{\log_2 \kappa} \quad (3.13)$$

where  $f_m \in F$  is a numerical feature,  $C$  the class variable,  $I(C|f_m)$  the *information gain* (also known as *Kullback–Leibler divergence* or *relative entropy*) of  $f_m$  with respect to  $C$ . The parameter  $\kappa$  is a penalty on the bias of the information gain in the case of nominal features with many possible values (or numerical features with high “discretization cardinality”, i.e. a high number of value ranges obtained when applying an adaptive discretization procedure). Similar to the CFS approach, the authors used a forward selection

algorithm to explore the search space of feature subsets, naming their approach *efficient feature selection* (EFS). When evaluating the feature subsets obtained with EFS using a random forest based learning algorithm (see *Class Prediction* section) on two microarray benchmark datasets for sample classification, EFS provided significantly better accuracies than CFS on both datasets, while the number of used features was similar.

Another fast filter method using an information theoretic measure was proposed by Zhang and Deng [163], who estimated the goodness of a set of features as discriminative predictors by means of the *Bhattacharyya distance*. This distance measure provides an upper bound of the Bayes minimum error probability [164] and can therefore be used to derive informative criteria for feature selection to build predictive models. Specifically, for two multivariate Gaussian distributions, their distance can be quantified as follows:

$$\frac{1}{8}(M_2 - M_1)^T \left[ \frac{\Sigma_1 + \Sigma_2}{2} \right]^{-1} (M_2 - M_1) + \frac{1}{2} \ln \frac{|\frac{\Sigma_1 + \Sigma_2}{2}|}{\sqrt{|\Sigma_1| |\Sigma_2|}} \quad (3.14)$$

While the first term scores the class separability according to the difference between the class means  $M_1$  and  $M_2$ , the second term provides the separability derived from the difference between the class covariance matrices  $\Sigma_1$  and  $\Sigma_2$ . In contrast to the widespread assumption that feature selection and classification should be considered in combination as a single optimisation problem to obtain the best predictive performance, according to which wrapper-based methods (see next section) would be expected to outperform filter methods, Zhang and Deng argue that the Bayes error is theoretically the best criterion to evaluate the effectiveness of a feature set for classification and depends only on the feature space, not on the classifier. Their method, the *Bayes Based error Filter* (BBF), takes the Bayes error indirectly into account for feature selection by controlling its upper bound using the Bhattacharyya distance.

A further distance measure derived from information theory and used in many feature selection methods is the *mutual information* (MI). The MI is a more general association measure than the Pearson correlation, accounting also for higher-order dependencies between the selected features and the outcome variable, and similar to the Bhattacharyya distance, it is related to the Bayes error. More specifically, Hellman and Raviv [165] showed that an upper bound on the Bayes error  $E(f)$  for a feature  $f$  is obtained by:

$$E(f) \leq \frac{1}{2}(H(C|f)) = \frac{1}{2}(H(C) - I(C, f)) \quad (3.15)$$

where  $I(C, f)$  is the MI between the class variable  $C$  and the feature  $f$ . Since this bound is minimised when the MI is maximised,  $I(C, f)$  is a valid approximation of the Bayes error. Another Bayes error bound using the MI had been proven earlier by Fano [166]. A practical drawback of the MI is that it requires discretized input variables or alternatively, a procedure for estimation of the entropy when using continuous variables. When applying a discretization method, even if the assumption of an underlying group of nominal variables for the features is valid, the risk of losing important biological information is high. However, discretization procedures have also been employed successfully to remove noise in microarray data [167].

The MI measure can also be used to compute a normalised, symmetric variant of the MI, the *symmetric uncertainty* (SU) [168], which is frequently used in microarray analysis:

$$SU(X, Y) = \frac{2.0 \times MI}{H(Y) + H(X)} \quad (3.16)$$

where  $H(X)$  is the Shannon entropy of variable  $X$  and MI the mutual information of  $X$  and  $Y$ .

An even more general association measure than the MI and SU, is the *normalised compression distance* (NCD) [149], which is derived from the theory of Kolmogorov complexity [169] and in practice computed from the lengths of compressed data files for single inputs and pairwise input concatenations. The NCD of two objects (e.g. discretized gene or sample vectors)  $x$  and  $y$  is defined as:

$$NCD(X, Y) = \frac{C(XY) - \min(C(X), C(Y))}{\max(C(X), C(Y))} \quad (3.17)$$

where  $C(X)$  is the length of a compressed version of variable  $X$  and  $XY$  is the concatenation of  $X$  and  $Y$ . Since the NCD only provides very rough approximations of the theoretically optimal normalised information distance (obtained when the compressor reaches the Kolmogorov complexity of the data), the estimated distances can often be very inaccurate, especially for small input vectors. However, if non-linear inter-dependencies between variables are to be captured, combining the NCD with other distance measures can help to find improved predictors [170].

Another information theoretic attempt towards feature selection, which does not only rely on a special distance measure but on a formal framework for defining optimal attribute subset selection, is the *Markov blanket filter* introduced by Koller and Sahami [171]. In agreement with other combinatorial filters like CFS, the goal is to find feature subsets with maximised relevance with respect to an outcome variable and minimised redundancy among the features. However, by modelling dependencies between features as a Bayesian network, a new graph-based interpretation of redundancy, the *Markov blanket*, is obtained, providing an alternative method to remove redundant information from the feature space. A Markov blanket for a node  $v$  in a Bayesian network (where the nodes correspond to the features and the edges to conditional dependencies) is the set of nodes composed of  $v$ 's parents, children and its children's parents. Figure 3.6 shows an example graph, where the nodes corresponding to the Markov blanket of node  $v$  have been highlighted in blue colour.

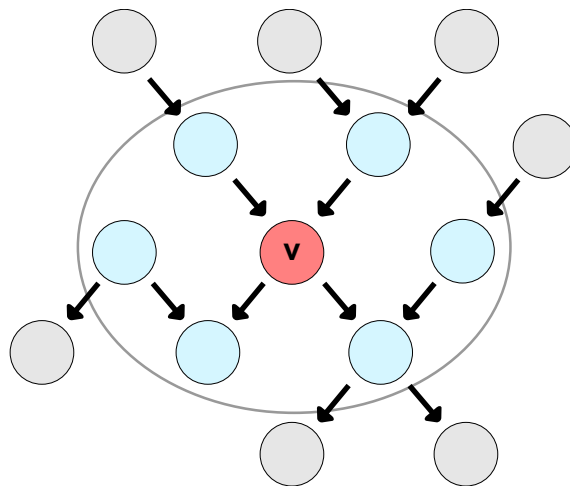


Figure 3.6: Example illustration of a Markov blanket in a sub-graph of a Bayesian network. The node  $v$ 's Markov blanket is highlighted by blue colour and surrounded by a grey circle.

The Markov blanket is the only knowledge needed to predict a node's behaviour; hence, a node which has a Markov blanket can be removed from the network without losing any information. Thus, the idea behind Markov blanket filtering is to start with the full set of features and iteratively remove features for which

a Markov blanket is found within a greedy backward-elimination algorithm. Since there might not be a full Markov blanket for every feature containing redundant information, a heuristic is applied to estimate how close a candidate set  $M_i$  is to a Markov blanket and features that are approximately subsumed by the information content of such a set are also removed. Candidate sets  $M_i$  for a feature  $f_i$  are the  $k$  features with the largest magnitude of correlation to  $f_i$  and the “Markov blanket likeness” is scored using the expected cross-entropy of  $f_i$  and  $M_i$ . The main benefit of the Markov blanket approach is that it can remove redundant features while still being more computationally efficient than most wrapper methods (see below). However, the approach also has significantly higher computational demands than univariate filters.

**Signal processing based methods:** Although most feature selection methods apply association measures on the original, untransformed data, the *normalised compression distance* (see above) is not the only example using transformed data to assess the similarity/distance between different variables. Subramani *et al.* showed that mathematical transforms from the field of signal processing, which are frequently used in image and video processing, can also provide benefits in the feature selection domain [172]. When analysing genes with the *Haar Wavelet power spectrum* of the gene vectors, they observed significant differences between the spectra for different diagnostic outcome classes. Building upon these results, they developed a new class separation measure and computationally simple and fast methods for gene selection and clustering. Wavelet transforms have several advantages over other transforms in this context, as they provide a lossless signal transformation with results that have good properties of localization both in time and frequency. Especially the capacity of these transforms to spatially adapt to varying frequency behaviour can be exploited for class distinction tasks in data analysis. The authors use the Haar wavelet because it consists of very simple low- and high-pass filters and can be computed efficiently. A technical drawback of the Haar wavelet transform is that it only accepts a number of input points corresponding to an integer power of 2. To address this limitation, zeros can be appended at the right end of the input data, extending the lengths of the gene rows to the next largest number  $x = 2^n$  (a technique known as “zero-padding”). Using this approach, Subramani *et al.* modelled the values of each gene across multiple samples (i.e. the input matrix rows) as one-dimensional signals and applied the 1D-Haar transform on them to compute a local wavelet power spectrum at different levels of detail of the signal decomposition (see [172] for details). Genes with a high difference in the average spectra across the subsets are selected as predictors for classification. When comparing the best-ranked genes on real-world data against those obtained using classical selection methods, the results revealed a high similarity. Therefore, the main benefit of this wavelet-based feature selection lies in the very efficient computation of the feature ranks.

A further wavelet-based selection approach [173] used only a subset of orthogonal wavelet approximation coefficients, in order to reduce the dimensionality of the input microarray data. In combination with a genetic algorithm to explore the space of these compressed features, promising results were obtained using a linear discriminant classifier on four benchmark datasets.

Recently, Nanni and Lumini extended previous methods by considering orthogonal wavelet detail coefficients from multiple wavelet types (Haar-wavelets, Daubechies, Symmlets and Coiflets of different orders, among others) [174]. For each set of coefficients, a support vector machine was trained, and feature selection was applied on the classifiers using Sequential Forward Floating Selection [175]. Using different cross-validation schemes, areas under the ROC curve above 0.9 were obtained for four microarray datasets when combining the selected SVMs into an ensemble.

**Ensemble feature selection:** Many approaches for microarray feature selection combine multiple algorithms to exploit their different strengths, i.e. employ ensemble learning techniques. Although feature

selection typically attempts to reduce the number of features used in a model, increasing model interpretability and reducing model complexity, some researchers argue that in large-scale datasets often many of the unselected features contain discriminative information that should be taken into account. Therefore, in addition to classical techniques using the intersection set or a majority vote combination of results from filter methods, various specialised ensemble approaches have been developed.

For example, Skurichina and Duin proposed to use ensemble learning with classifiers constructed on sequentially selected sets of features [176]. The general idea is to first determine an optimal feature subset with respect to a given scoring function, then a second-best subset on the remaining set of features, and iteratively repeat this procedure until all features have been assigned to one of several ranked subsets. The ranked list of feature subsets is then used as input for an ensemble prediction method. This generic approach enables the experimenter to identify the most discriminative features, while at the same time taking into consideration other less informative attributes. The authors conducted experiments with three types of subset selection mechanisms (simple forward selection algorithms, Principal Component Analysis (PCA) and random feature selection) on several real-world data sets, and observed that the mean classification error of an ensemble of Linear Discriminant Analysis (LDA) classifiers was significantly reduced by additionally considering low-rank feature subsets.

A more complex evolutionary search approach towards ensemble feature selection was adopted by Deutsch [177] in the *GESSES* (genetic evolution of sub-sets of expressed sequences) algorithm. This method constructs an initial gene pool using a filter and applies an evolutionary algorithm using a statistical replication operator (or alternatively deterministic evolution) and simulated annealing to evolve an ensemble of different gene subspaces to provide an optimal set of k-nearest neighbour predictors (k-NN with  $k = 1$ ) in terms of a score closely related to the LOOCV-accuracy. *GESSES* provided competitive classification results on four microarray cancer benchmark datasets, including a multi-class prediction problem, but the ensemble models tend to be complex.

In summary, filters include many state-of-the-art feature selection methods, and especially combinatorial filters, ensemble filters and signal-processing based methods provide informative attribute selections at low computational costs, whose predictive power can often compete with wrapper-based methods, discussed in the next section.

### Wrappers and embedded methods

Filter-based feature selection methods are computationally efficient and provide attribute rankings, which are useful for data interpretation and dimensionality reduction prior to the application of other higher-level analysis methods. However, for the specific task of building predictive models for classification and regression analysis, the filter-based scoring functions often only provide crude estimates of the combined utility of multiple features.

As an alternative, *wrappers* directly combine the feature evaluation process with the chosen prediction method by ranking feature subsets using the prediction results of the model built upon them. Wrapper methods can account for effects of inter-correlations and redundancies among the feature variables and thus remove redundant features from the selection that would unnecessarily be selected by a univariate filter. Even in comparison with combinatorial filters, wrappers tend to select feature subsets with superior predictive performance, because the filter-based scoring of feature subsets can only provide a rough estimate of the relative scores to be expected in a direct predictive evaluation using the learning method of choice and

an external test set. However, the benefits of wrappers for feature selection typically come at the expense of higher computational costs, due to the expansion of the search space when considering combinations of features, and the necessity to train and evaluate a predictive model for every subset evaluation. When considering both algorithmic *effectiveness*, in terms of maximising the predictive accuracy, and *efficiency*, in terms of minimising runtime, comparative evaluations suggest that filter methods tend to provide a better trade-off between both goals [178]. In practice, the runtime problems associated with wrappers can however be alleviated by employing fast search algorithms or using simple or very efficient statistical learning methods.

Ideally, the search algorithm's balance between *exploitation* and *exploration* of the search space should provide a quick convergence, but at the same time avoid entrapment in local minima. For microarray datasets, fast search heuristics are used in most cases, since exhaustive search is already infeasible on standard computers when the number of features is in the range of a few hundreds (the number of possible feature subsets for  $n$  features is  $2^n - 1$  and  $n$  is typically in the range of thousands to a few ten thousands). Therefore, filter methods are still widely used for microarray analysis, and for wrappers both the search algorithm and the classifier are typically chosen to be as simple and efficient as possible to reduce the computational costs.

**Wrappers using efficient learners:** One of the first prominent examples for wrappers employed in gene expression data analysis was the method by Inza for gene subset selection based on a LOOCV scheme and four machine learning algorithms (IB1, Naive-Bayes, C4.5 and CN2), evaluated on three benchmark microarray cancer data sets [179]. The author used a simple sequential forward search (SFS) algorithm to explore the feature subset space. When applying the supervised learning algorithms for classification, prediction accuracies above 85% were achieved in all cases using less than seven genes as features. However, the reported CPU times for the analysis (e.g. 203,053 seconds in the worst case) are often considerably higher than usual runtimes required when using simpler filter methods.

An earlier attempt to use wrappers in microarray feature selection employed a k-nearest neighbour algorithm ( $k = 3$ ) in combination with a standard generational genetic algorithm (GA) [180]. This method classified 33 out of 34 test samples in a leukaemia data set [181] correctly, using a training set of 38 samples. However, no cross-validation was applied and 50 genes were included in the prediction, after ranking features according to the frequency of their selection in the GA. The strategy to compensate for a computationally expensive evolutionary search algorithm by employing a simple classifier was also adopted in an alternative approach combining the fast Naive-Bayes classification algorithm with an Estimation of Distribution Algorithm (EDA) [182]. Although EDAs are stochastic in nature like other evolutionary algorithms, a robust selection of genes across different initializations was obtained in this study.

Instead of using simple and fast machine learning approaches like k-NN and Naive Bayes, more complex techniques including regularised classification methods [183] can also be employed, if the algorithm or implementation is particularly efficient. For example, Roth used a very efficient version of the *Least absolute shrinkage and selection operator* (LASSO) [184] by Osborne *et al.* [185] and additionally extended it to work with general loss functions ("Generalized LASSO") [186]. The main benefits of the models with  $l_1$ -regularization obtained from this technique, are that they are mostly very sparse, easy to interpret and provide posterior probabilities instead of only binary class labels as output. The method was tested on a well-known leukaemia benchmark dataset using a Monte Carlo cross-validation procedure (200 randomly chosen 80%/20% training/test set splits) reaching a comparatively low average error rate (0.025) for a low average number of selected genes (24.4).

**Wrappers using efficient search methodologies:** An alternative strategy to alleviate the problem of high runtimes in wrapper-based feature selection is to use more efficient search space exploration methods, like evolutionary algorithms, instead of simplifying the scoring function. In most of the corresponding approaches, solutions are represented by a binary bit-string with length equal to the number of features/genes in the dataset, encoding selected features by 1, and others by 0. For example, a parallel genetic algorithm (PGA) [187] has been proposed, using a two-criteria fitness function including the test set accuracy of a correlation-based classifier and a penalty term for complex models proportional to the size of the selected feature subset. Although this classifier is only applicable to binary class problems, the possibility to parallelize the computation on separate CPUs enables further improvements in terms of runtime efficiency.

A genetic algorithm (GA) has also been used by Peterson and Thaut [148]; however, they applied a simple GA in combination with support vector machines (SVM) with a  $2^{nd}$  order polynomial kernel or the radial basis function (RBF). In their wrapper approach, the classification accuracy was evaluated using a stratified 10-fold cross-validation scheme on a Lymphoma microarray dataset, resulting in 30 test sets to evaluate each combination of parameter settings they tested. Overall, their feature selection method performed favourably in comparison to using a full feature set and random feature subsets, but was not compared against other selection schemes.

GA-based feature selection has been further extended by Ooi and Tan [188], who presented an approach optimising both the features and the feature subset size automatically, using a maximum likelihood classification (MLHD) method as the scoring function. In contrast to previous GA-based selection approaches, solutions have a more compact string representation, in which the first element  $R$  denotes the size of the selected subset and the remaining entries the indices of selected genes up to a given maximum size  $Rmax$  ( $Rmax$  can be larger than  $R$ , in which case only the first  $R$  entries are considered for classification). Using extensive evaluations on real-world data, recommendations were derived for the choice of the selection and crossover operator and other GA-parameters. The authors presented cross-validation and test set prediction results on multi-class microarray benchmark datasets according to which their method outperformed previous competitive approaches.

Finally, apart from evolutionary algorithms, fast greedy search approaches can be employed to decrease the runtime, including both top-down approaches (eliminating one feature at a time) and bottom-up approaches (adding one feature at a time to the subset selection). One of the most prominent examples for this type of approach is the combination of linear SVMs with a top-down recursive feature elimination (RFE) selection procedure [189], outperforming classical filters on several microarray datasets.

**Embedded methods:** Even with efficient scoring functions and search methodologies, wrapper-based attribute selection is still computationally more expensive than other approaches. However, some prediction methods enable an implicit selection of features by using internal model parameters for the selection, a strategy termed as “embedded selection”. These methods are specific to the prediction algorithm and couple the feature selection tightly with the model generation, similar to wrapper approaches, but the runtime is reduced drastically by avoiding the separate scoring of entire feature subsets and the search space exploration.

The most prominent example for embedded feature selection is the decision-tree based *random forest* (RF) [190, 191] classification and regression method, assessing the importance of features during the model construction phase. In the RF approach this can be achieved by computing the *Gini index* [192], a measure for the inequality of the class distribution across child nodes in a decision tree. In the domain of microarray analysis the RF method has often been applied successfully [193, 194], although predictor correlations can



lead to spurious signals in some cases [195].

Other embedded methods select features by using the weights assigned to attributes in linear prediction models as feature scores. For example, such techniques can be applied to ridge regression [196] and regularized logistic regression [197].

**Wrapper/filter hybrids:** Due to the benefits of wrappers in terms of predictive power, and the advantages of filter methods in terms of efficiency, various research groups have presented hybrids of these approaches. A prominent example is the two-stage hybrid proposed by Xing, Jordan and Karp [198]. In the first stage, *unconditional univariate mixture modelling* was applied to fit a two-component Gaussian mixture model using an EM algorithm to estimate an underlying binary state for each gene, and to filter genes according to their mixture overlap probability (a redundancy-based filtering approach). The remaining features were ranked and filtered again based on a mutual information selection scheme, and finally Markov blanket filtering (see section on filters above) was used in the final pre-filtering step. The selected features were then embedded into three different prediction methods (a Gaussian quadratic classifier, logistic regression and k-NN using the Pearson correlation as distance metric) and evaluated on gene expression cancer data [181] using both cross-validation and a training/test set partition. Although best error rates between 0% and 2.9% were obtained, considering gene sets sizes between 2 to 100, a low number of 34 test samples and the consideration of only one dataset limit the informative value of this comparison with other approaches in the literature.

To improve upon the simple correlation and separation measures used as filters in the first stage of hybrid approaches, C. H. Yang *et al.* proposed to combine an information theoretic measure related to the mutual information, the *information gain* (IG) with a wrapper selection method. After the IG is used to pre-select features, a simple generational GA with a 2-point crossover operator in combination with a k-nearest neighbour classifier and LOOCV is employed for the final selection task. On 8 of 11 microarray cancer benchmark datasets, the IG-GA method achieved the highest test set accuracy in comparison to ten other commonly used algorithms (no comparison using cross-validation was carried out).

Similarly, Akadi *et al.* proposed the use of a combinatorial filter, MRMR (Minimum Redundancy-Maximum Relevance, see section on filter methods), as an extension of univariate filters in combination with a GA-based wrapper [199]. By combining this methodology with the Naive Bayes method and SVMs as scoring functions in the GA, LOOCV accuracies and gene set sizes obtained on five microarray cancer datasets were superior to those reached by the independent application of the filter and wrapper methods.

Recently, hybrid selection methodologies have also employed ensemble selection approaches in both the filter and wrapper algorithms. Leung and Yong presented a *multiple-filter-multiple-wrapper* (MFMW) method, combining 3 filters (signal-to-noise ratio, t-statistic and Pearson correlation) and three classifiers (weighted voting, k-NN and SVM) [200]. The ensemble of filters was obtained by computing the union list of selected genes, while the ensemble of wrappers resulted from selecting gene sets with minimal number of wrong or undecided predictions according to a unanimous voting scheme for classification. The authors showed that MFMW outperforms corresponding SFSW techniques in all cases on six benchmark datasets in terms of the average LOOCV accuracy.

In summary, wrappers and embedded methods typically have the capacity to provide feature subset selections with superior prediction performance (given the chosen classification method) than filter-based approaches. However, in terms of cost-efficiency and with a limited runtime in most practical scenarios, filters are often the method of choice [178], and especially if combinatorial filters or filter ensembles are

used, high performance can still be achieved at relatively short runtimes. Moreover, to fine-tune the balance between performance and runtime, filters and wrappers can be combined flexibly in hybrid approaches, pre-processing the data with filters and then exploiting the efficiency of wrappers. In the future, as multi-core CPUs become more affordable, parallelized wrappers are likely to make the direct application of wrappers on unfiltered datasets more practically feasible, allowing experimenters to avoid potential performance bottlenecks of simple pre-processing filters.

### 3.4 Class Discovery (Unsupervised Machine Learning)

Apart from identifying informative genes, proteins or metabolites in large biological datasets using feature selection methods, another frequently occurring task for machine learning methods is to address the question whether biologically meaningful groupings of similar samples or features occur within the data. Unsupervised learning methods aim at the identification of such interpretable but non-trivial groupings and structures in the data. In contrast to supervised analysis techniques for predictive model generation, unsupervised methods do not require any training data with class labels or numerical outcome variables, but are also incapable of using this information, if it is available. However, clustering methods, which represent the most prominent class of unsupervised learning algorithms, often provide an intuitive means for classification, e.g. by labelling new samples according to their distances to the cluster centroids.

In microarray analysis, both the genes and the samples can be clustered depending on the goal behind the experiment. Clustering of sample vectors is used to identify natural groupings of similar samples (e.g. groups of biological samples representing different tumour subtypes), whereas clustering of feature vectors can help to discover genes/proteins with similar expression patterns, which might point to functional associations.

In this section, apart from classical partition-based and hierarchical clustering techniques, more recent unsupervised methods will be reviewed, which abandon the assumption that the data can be grouped into distinct, non-overlapping classes. These approaches model internal structures in the data as probabilistic *mixtures* or overlapping *processes* and assign probabilities for the membership in different classes to each instance. Importantly, the multitude of existing clustering methods is also complemented by several approaches to analyse the validity of clusters. These techniques are reviewed in detail here, because often a multitude of informative patterns can be identified in the data using different clustering methods and parameters, and validity measures enable an evaluation of the significance of the identified structures. However, no gold standard for the validation of clustering result exists, hence, this survey will also provide guidance on how to compare and combine multiple clustering methods and validity measures to obtain more robust and reliable results. Figure 3.7 shows an overview of different categories of clustering approaches that will be discussed in the next paragraphs, as well as some of the representative algorithms.

#### Clustering approaches and related methods

In the literature, three general categories of clustering algorithms are typically distinguished [6]:

- *Combinatorial algorithms* are methods operating directly on the observed data, i.e. without using derived latent variables or probabilistic models. These algorithms belong to the most frequently used techniques; however, they are not designed to model overlapping groupings and require the specification of a pre-defined number of clusters.

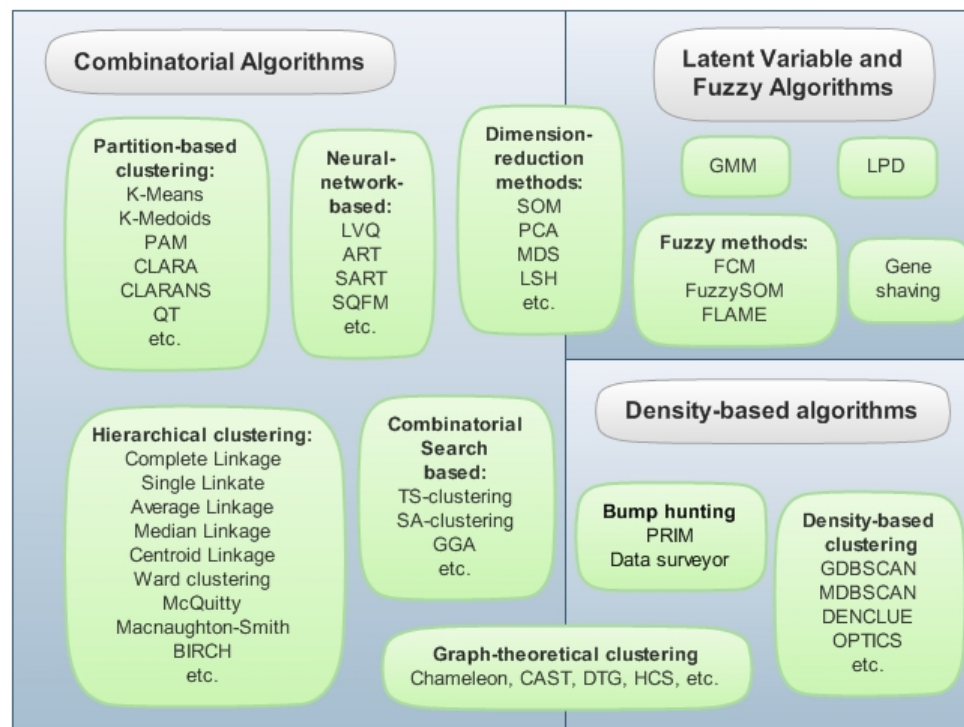


Figure 3.7: Overview of unsupervised learning methods (non-comprehensive)

- *Mixture modelling* approaches assume that the data points are samples drawn from a certain probability distribution. Their density is described by a parametric mixture model, which can for example be fitted by a maximum likelihood approach.
- *Mode seekers* (or *bump hunters*) are methods which try to structure the data by identifying distinct regions of high data density, separated by low density regions.

The last type of methods, density-based approaches, will not be considered in detail here, because although some microarray clustering approaches rely mainly on the analysis of data density [201], the results are often difficult to interpret in high-dimensional spaces and most of these methods are not yet widely used by microarray experimenters.

Apart from the algorithmic categories above, another line of distinction can be drawn between *partition-based clustering methods*, searching for data partitions which optimise a certain merit function, and *hierarchical clustering methods*, iteratively dividing or merging clusters in a hierarchy of clustering levels, by starting with a single large cluster for the entire data set (top-down approach) or many small clusters for each data point (bottom-up approach) [202]. Importantly, many clustering techniques are at the same time *dimensionality reduction methods*, since mapping the data into a lower-dimensional space can facilitate both the group separation task and the interpretation for the user (see also the discussion of the “curse of dimensionality” in the section on feature selection). More recently, links between graph/network analysis and clustering techniques have been discovered and exploited in *graph-theoretical clustering* methods, which will also be discussed in the method overview in the following section.

**Combinatorial algorithms:** One of the earliest and simplest combinatorial, partition-based clustering methods is the *k-Means* algorithm, which in spite of its simplicity has been shown to be highly effective in many practical applications. Starting with a given assignment of the data points to a pre-defined number of

$k$  clusters (e.g. a random assignment), a two-step iterative procedure calculates the cluster centroids (i.e. the averages of the cluster member coordinates) for the current assignment, re-assigns the data points to their nearest cluster centroid, and then iteratively repeats this procedure until convergence or a maximum number of iterations is reached. The objective function minimised by k-Means is the sum of squared distances between data points and their associated cluster centres.

Different variants of this general k-Means clustering scheme have been developed by Hartigan and Wong [203] (often regarded as the “standard” k-Means algorithm), MacQueen [204], Forgy [205] and Lloyd [206]. In spite of the success of these methods, an important general limitation of all k-Means variants is the strong dependency of the output on the initialisation. If random assignments are used, the procedure should be repeated several times with different initialisations to avoid entrapment in local minima of the objective function. If the number of clusters  $k$  is unknown, cluster validity measures (see section below) can be used to identify a suitable value for this parameter. However, there are many different ways to define optimality for cluster separations, and since there might be multiple biologically meaningful structures in the data, it is recommendable to consider several different solutions.

*K-medoids* is a robust variant of k-Means using *medoids* instead of centroids as an alternative representation of the cluster centre. The medoid for a cluster is the data point closest to all other (current) members of the cluster with respect to a given similarity measure. After starting with an initial cluster assignment, the algorithm iteratively calculates the medoids for the current assignment and then re-assigns observations to the cluster of their closest medoid. Using medoids instead of centroids as cluster centres reduces the influence of outliers on the final cluster assignments and therefore tends to provide more robust clustering results. Moreover, any (dis-)similarity matrix will suffice as input for the algorithm, and the Euclidean distance does not necessarily need to be used as a distance metric. However, a significant practical drawback for high-dimensional data is that the runtime for computing the cluster medoids is quadratic in the number of data points.

A further robust alternative to the k-Means algorithm is the *Partitioning around medoids* (PAM) clustering method [207]. In contrast to k-Means and similar to k-Medoids, PAM minimises a sum of within-cluster dissimilarities instead of a sum of squared Euclidean distances, using medoids as robust cluster centre representations. Accordingly, a proximity matrix is required as input, which can be computed from the original data using a distance measure of choice. The PAM algorithm clusters the data in two iteratively repeated phases: In the “Build”-phase,  $k$  initial medoids are selected, typically by choosing  $k$  centrally located data points. In the subsequent “Swap”-phase, the sum of within-cluster dissimilarities is reduced iteratively, by considering all possible pairings of a medoid with a non-medoid data point and swapping their roles if this reduces value of the objective function. This swapping process continues until no further reduction is possible. Similar to the K-Medoids approach, a quadratic runtime complexity can render the application of PAM infeasible for very large data sets, but extensions of PAM like CLARA [207] and CLARANS [208] using sampling techniques with improved efficiency have been developed to address this problem.

Apart from these k-Means and k-Medoids derived algorithms, the group of partition-based clustering methods also contains approaches involving dimensionality reduction of the data. A prominent example which is frequently used for microarray analysis is the *Self-Organizing Map* (SOM) approach by Kohonen [209]. SOMs provide easily interpretable low-dimensional visualisations of the clustering results, however, the selection of suitable parameters is often difficult. The SOM procedure maps the data points into a low-dimensional topological map according to their closeness to overlaid prototype grid points in the map.

More precisely, a SOM can be generated using the following online pseudocode algorithm:

1. choose a random data point  $x_i$
2. find the prototype  $m_j$  on the grid that is closest to  $x_i$
3. for all neighbour prototypes  $m_k$  of  $m_j$  (neighbours = points within user-defined Euclidean distance threshold  $r$ ): Move  $m_k$  into the direction of  $x_i$ , so that the magnitude of the movement is proportional to the distance between  $m_k$  and  $x_i$  and to a learning rate parameter  $\alpha$  ( $m_k += \alpha(x_i - m_k)$ )
4. reduce the learning rate  $\alpha$  and the distance threshold by a small amount and repeat the procedure iteratively

If a small enough value is chosen for threshold parameter  $r$ , such that every neighbourhood only contains one data point, this algorithm corresponds to an online version of k-Means with an additional low-dimensionality constraint. This inherent low-dimensionality constraint in SOM can however also limit the applicability of the method, since significant errors might be introduced due to the dimension reduction. When calculating this “reconstruction-error” for k-Means- and SOM-models after dimensionality reduction, the error is generally smaller for k-Means, but as long as the difference to k-Means is not large, SOMs can still be used effectively.

**Mixture modelling and fuzzy approaches:** Apart from these classical clustering algorithms, mixture modelling methods have been developed, which are often inspired by the closely related original k-Means method. One of the most prominent examples is the *Gaussian Mixtures* approach derived from the concept of Gaussian mixture models (GMM), which has already been employed for several other purposes including density estimation. In the context of unsupervised learning, a Gaussian mixture can be understood as a “smoothed” variant of k-Means clustering, where every class (i.e. every biological condition) is modelled by a multivariate Gaussian function and the samples are assumed to be drawn from a mixture of these distributions. Importantly, each data point can be generated by choosing one of these functions with a given probability, hence, observations are considered to belong to a specific class only with a certain probability. Mixture modelling based clustering extends the k-Means algorithm by replacing the two iteratively repeated centroid-computation and membership-assignment steps by the two steps of an EM-algorithm: The expectation step (E-step) computes the expected values for the so-called “responsibilities” of all data points (i.e. estimates of the conditionals  $p(z_i|x_j)$ , for the mixture component  $z_i$  and observation  $x_j$ , given a model with  $M$  components and  $N$  observations, with  $i \in M$  and  $j \in N$ ) using the current parameter estimates for the distributions. These responsibilities are then provided as input in the maximisation step (M-step) to update the distribution parameters, which enter the next iteration until convergence is observed.

An alternative possibility to account for contradicting evidences for the assignment of observations to different clusters is *Fuzzy clustering*. Instead of assigning the samples to crisp sets, a partial membership to several categories is estimated and expressed by membership functions taking values between 1 (full membership) and 0 (no membership). The *Fuzzy C-Means* (FCM) algorithm [210] is a typical representative of these fuzzy clustering approaches, minimising the following objective function:

$$J(U, V) = \sum_{i=1}^C \sum_{j=1}^N u_{ik}^m d_{ik}^2 \quad (3.18)$$

where  $d_{ik}^2 = (x_k - v_i)^T (x_k - v_i)$  corresponds to the squared Euclidean distances between the data points  $x_k$  and the cluster centres  $v_i$  (stored in matrix  $V$ ),  $u_{ik}$  are the membership degrees (stored in matrix  $U$ ),  $C$  is

the number of clusters and  $N$  the size of the data set. The so-called “fuzzifier”  $m$  is a user-defined trade-off parameter with a value greater than 1, determining the balance between crispness and fuzziness (also known as “sharpness”). The smaller the value of  $m$  (i.e. the closer  $m$  is to 1), the closer the membership values  $u_{ik}$  will be to 0 or 1, and the greater the value of  $m$ , the more similar the membership values become (in practice  $m \in [1, 2.5]$  is often used [211]). Similar to k-Means, the optimisation algorithm is using an iterative two-step procedure, minimising the objective function.

**Hierarchical algorithms:** If the experimenter assumes that the data might have an underlying hierarchical structure, which is often the case for transcriptome and proteome datasets due to the regulatory effects of transcription factors and signalling proteins, *hierarchical clustering* approaches can be used as an alternative to partition-based methods. The main benefits of hierarchical clustering schemes are that they do not require a pre-specified number of clusters as input parameter and provide easily interpretable tree visualisations (dendrograms) of the clustering results. A dendrogram is a rooted tree, where the root represents the entire data set, the inner nodes correspond to the cluster of all their children and the leaves stand for single data objects.

In microarray analysis, a hierarchical clustering is often applied to both the samples and features, and the resulting groupings of sample and feature vectors in the expression matrix are visualised in a *heat map*, coloured according to expression value ranges (see an example in figure 3.8, containing a dendrogram for the genes (left vertical axis) and samples (top horizontal axis)).

However, the applicability of these methods is limited by the assumption that the data has a hierarchical, tree-like structure. Since only  $n-1$  parameters are required to describe the cluster dendrogram for a data set with  $n$  observations, a problem arises from the potential loss of information that can occur when transforming the original dissimilarity matrix (requiring  $n(n-1)/2$  parameters for its description) into a dendrogram. In order to verify the validity of the assumption that the data has a tree-like structure and only a negligible information loss occurs due to the transformation into a dendrogram, the “tree-likeness” of the pairwise distances in the dataset can be estimated with the *cophenetic correlation coefficient* (CCC). The CCC is the correlation between the original dissimilarities of observations in the distance matrix and the “cophenetic dissimilarities” in the final tree, i.e. the inter-group dissimilarities determined by searching the smallest common ancestor of two data objects in the dendrogram. If a CCC close to 1 is obtained, the calculated dendrograms can safely be assumed to provide a good representation of the original input data.

In order to build a clustering dendrogram, two alternative strategies can be adopted. The *top-down* (or *divisive*) approach starts with a single cluster containing the entire dataset and iteratively divides clusters into the subclusters with the largest between-cluster dissimilarity. By contrast, the more common *bottom-up* (or *agglomerative*) approach is initialised with a cluster for every single object and iteratively combines clusters with the smallest between-cluster dissimilarity. Several measures for the between-cluster dissimilarity have been proposed in the literature, e.g. the minimum distance between two members from different classes (*Single Linkage* (SL)), the corresponding maximum distance (*Complete Linkage* (CL)) and the average distance (*Average Linkage* (AL)).

AL-based clusterings tend to be superior to SL- and CL-based results, providing more compact clusters and avoiding concatenations, but at the expense of higher runtimes. Further alternative agglomerative clustering methods include *Ward’s minimum variance method* [212], which tries to find spherical and compact clusters (often similar to the CL-clusters) and McQuitty’s *Similarity Analysis by Reciprocal Pairs* [213], which merges clusters with highest average similarity until the similarity measure between every pair of clusters is less than a predefined cut-off. In order to choose the best clustering method for a specific dataset, the

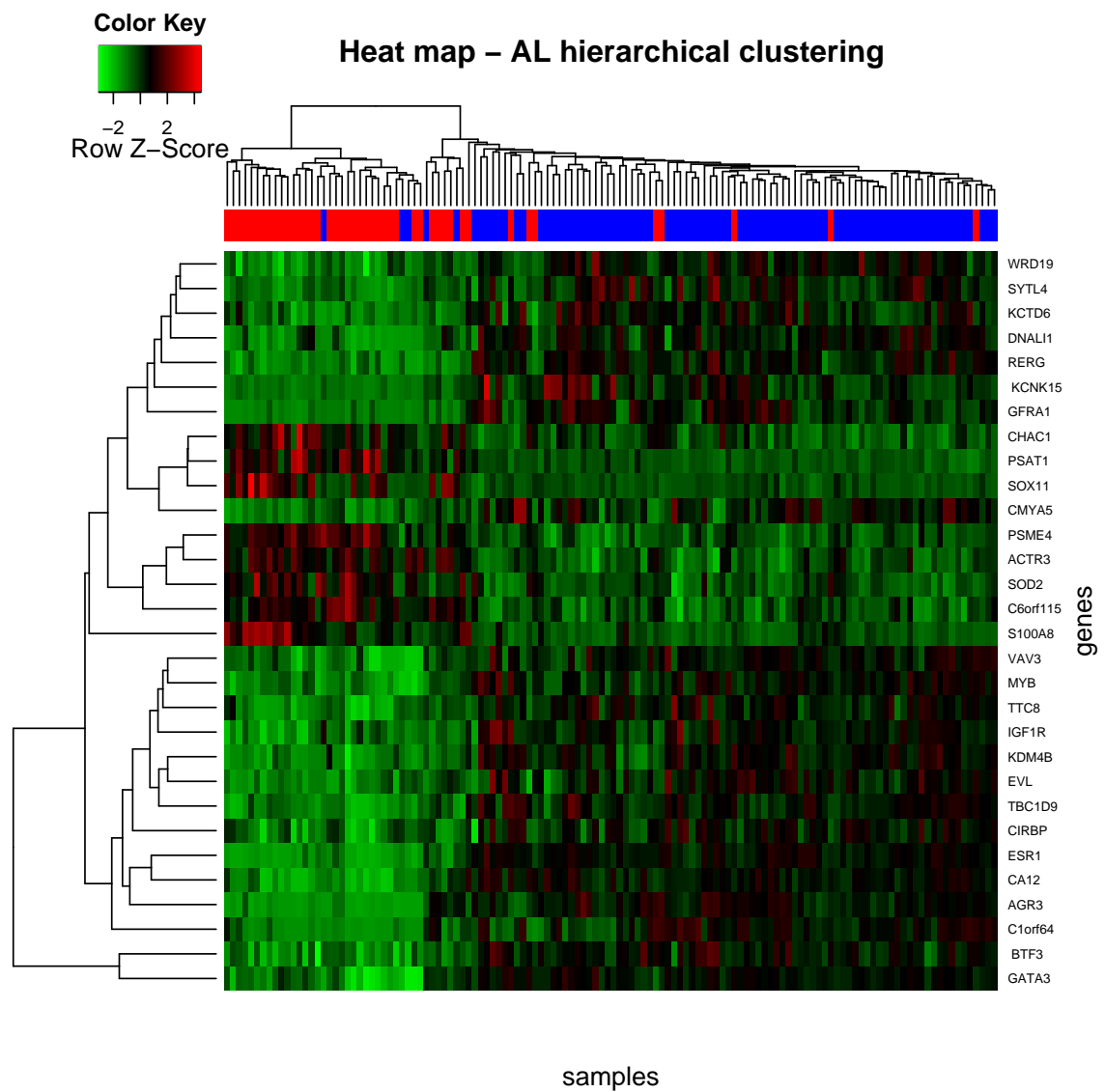


Figure 3.8: Example heat map created for 128 breast cancer microarray samples and 30 genes analysed in collaboration with the Queens Medical Centre in Nottingham [19] (see chapter 8 for details). Both the samples and the genes are clustered using average linkage hierarchical clustering with a Euclidean distance metric. The sample columns highlighted in blue and red correspond to two different clinically relevant tumour subtypes (blue = luminal group, red = non-luminal group; since the genes were selected in a supervised manner, the clustering was only applied for visualisation purposes).

experimenter can compute the cophenetic correlation coefficient to compare the different approaches.

For divisive clustering, which is less frequently applied than agglomerative techniques, a separation rule to split parent clusters into child clusters can be obtained using an iterative k-Means or k-Medoids clustering with  $k = 2$ . An interesting, parameter-free alternative is the *Macnaughton-Smith algorithm* [214]. This approach places the observation with the largest dissimilarity to all other items in the initial cluster ( $G$ ) into a new cluster ( $H$ ), and iteratively moves the members from cluster  $G$  with the highest difference between the average dissimilarity to  $G$  and to  $H$  ( $\text{dissim}(G) - \text{dissim}(H)$ ) to cluster  $H$ , until this difference becomes negative. This procedure is repeated by selecting the cluster with the largest diameter and dividing it analogously into two sub-clusters until all elements are assigned to a different cluster, and a complete dendrogram has been built.

A more recently developed, related method is the *Cluster Affinity Search Technique (CAST)* [215], which improves upon previous greedy search based hierarchical clustering methods by allowing the algorithm to alter cluster assignments even when there are no more unassigned observations left after an initial greedy cluster construction. Specifically, in a special “REMOVE-step”, cluster members are moved back to the set of unassigned objects, if their average similarity to the other cluster members falls below a given threshold. This technique decreases the risk that the objective function, the sum of pairwise dissimilarities within the clusters, converges to a local minimum.

However, in contrast to the fuzzy clustering extensions of partition-based clustering methods, classical hierarchical methods cannot model uncertainty in the data. Therefore, more recently, hierarchical clustering has been extended to provide a probabilistic model of the data using a Bayesian approach [216, 217]. This *Bayesian agglomerative clustering* method does not only account for the uncertainty in the data, but also frees the user from the choice of a distance metric and the selection of parameter values, by using Bayesian hypothesis testing to decide on cluster merges. The algorithm runs in polynomial time, consisting only of a one-pass bottom up merging procedure, and has provided new biological insights on microarray data for the plant model organism *Arabidopsis thaliana* [217].

**Ensemble/hybrid approaches and other techniques:** Similar to the feature selection domain, where hybrid and ensemble approaches have outperformed many of the traditional algorithms, the ideas of combining the outputs from multiple methods (ensemble/consensus approach) or sharing information between different algorithms running in parallel or sequentially (hybrid approach) have also been applied to the clustering problem.

The simplest ensemble clustering techniques, apply the same clustering algorithm multiple times with different parameters and aggregate the results. For example, the recently proposed *MULTI-K* algorithm [218] applies the k-Means method several times, sampling the number of clusters from a uniform distribution between a chosen minimum and maximum number of clusters. The clustering results are modelled as graph, where the nodes represent observations and the edges between them receive weights, increasing by 1, each time the corresponding nodes are clustered together. After completing this weighting procedure, all positive edge weights are iteratively reduced by a unit weight until they are zero (i.e. the nodes are disconnected), so that the graph is progressively decomposed into smaller connected components. By plotting the number of reverse weight reduction steps (y-axis) against the number of divided sub-graphs (x-axis), resulting in a so-called “cut-plot”, robust cluster structures can be identified as long horizontal regions in this plot, and the y-axis value for the longest region is the estimated number of clusters. When comparing these clusters with those obtained by traditional clustering methods using the adjusted rand index with known outcome classes (see section on validity measures below), MULTI-K tends to provide superior results [218].



The most common ensemble clustering methods count and compare the cluster associations of sample pairs across the base clusterings to calculate a *similarity matrix* or graph (some approaches also use the synonyms *agreement* [12], *consensus* [11] or *co-association matrix* [219]). The final clustering is then obtained by applying a function that integrates the information from the matrix into a single ensemble clustering result (e.g. using fuzzy k-means or hierarchical clustering). In contrast to this pairwise sample similarity based approach, other research groups have tried to integrate multiple clustering results by analysing sample-cluster relations. They introduced the *binary cluster-association matrix*, in which the rows represent the samples and the columns the clusters in different clustering solutions (given a fixed number of clusters), setting the matrix entries to 1, if a sample was assigned to the corresponding cluster, and 0 otherwise [220, 221]. Although good consensus clustering results have been obtained both using similarity matrices and cluster-association matrices, relations between the clusters are ignored in both approaches. Therefore, more recently, the *link-based cluster ensembles* (LCE) method has been developed as an extension of the above ensemble clustering techniques. LCE summarises the information from multiple clustering results and additionally identifies and preserves associations between the clusters [222, 223]. Using the binary cluster-association matrix (BM, see above), which contains crisp associations between samples and clusters, a refined cluster-association matrix (RM) is computed by taking into account the similarity between clusters. These similarities are used to refine the binary assignment of samples to single clusters (1 = “known association”, 0 = “unknown association”) into probabilistic real-valued assignments to multiple associated clusters. While all entries in the original BM equal to 1 are preserved, the zero entries are replaced by the similarity between the corresponding clusters and the cluster with value 1. This *cluster similarity* is estimated using a new link-based algorithm operating on a graph representation of the BM (vertices = clusters, edge weights = number of shared samples / number of sample in both clusters) and the Connected Triple method [224].

Apart from combining multiple clustering results using consensus techniques, increased robustness can also be achieved by combining different search methodologies within the algorithm, e.g. using hybrids of multiple optimisation techniques. For example, Marinakis *et al.* have presented a clustering approach, merging an evolutionary algorithm (Honey Bees Mating Optimization Algorithm, HBMO) with the Greedy Randomized Adaptive Search Procedure (GRASP) [225]. This method is also an example for the combination of unsupervised feature selection (using HBMO) with clustering (using GRASP).

In summary, depending on the nature of the data, partition-based, hierarchical and mixture modelling approaches towards clustering are all capable of providing new biological insights in microarray data analysis. More recent extensions employing hybrid or consensus techniques, exploit the benefits of different types of clustering algorithms, data representations and dimensionality reduction methods in order to find cluster structures which are more robust with regard to noise. Although some clustering methods can also automatically determine the optimal numbers of clusters or processes, the computation of validity indices is still required for validation purposes (see next section).

#### **Cluster validity / Selection of the number of clusters**

For clustering methods relying on a fixed number of clusters or processes, regardless of whether cluster assignments are crisp or whether the clusters can overlap, the selection of the number of these groupings is a key parameter of the algorithm. Using many clusters could result in artificially separated data points with similar properties, whereas creating few clusters might force the algorithm to group very dissimilar

objects together. In both cases, more biologically meaningful, inherent structures in the data could remain concealed due to an inadequate parameter selection. Thus, the outcome for different parameter choices should be compared and validated objectively.

For this purpose, several cluster validity indices have been proposed. These measures and the main ideas behind them can be grouped into three categories [226]:

- *internal* validity measures using cluster *compactness/division*: quantifying cluster *homogeneity* and *separation*
- *internal* validity measures using cluster *robustness*: assessing cluster *stability and reliability* based on p-values
- *external* validity measures: assessing the *agreement with a reference partition*, requiring external data as a “ground truth”

Since these validity measures could theoretically be used both for parameter optimisation and for the validation and comparison of the final clustering with other methods, the experimenter has to make sure not to use the same or related measures for both of these purposes.

**Compactness/division-based measures:** The homogeneity- and separation-based methods score the similarity of objects within a cluster and/or the dissimilarity of objects across clusters. Depending on the similarity matrix provided as input, the most common validity measures combine similarities of single pairs of observations into an overall similarity score for cluster pairs by just averaging pairwise similarities. A less computationally expensive alternative for high-dimensional data is to only consider the average similarity of the cluster centroid/medoid to the remaining cluster members. Moreover, after having computed a homogeneity score for single clusters and/or a separation score for a single cluster-pair, the corresponding measures for a complete clustering result can analogously be obtained by averaging over all clusters (or respectively, all cluster pairs).

One of the most frequently used validity measures combining both within- and between-cluster dissimilarity is the average *silhouette width* [227]. For an object  $i$  assigned to a cluster  $A$ , the silhouette width  $s(i)$  is computed using the average dissimilarity of  $i$  to all other objects of  $A$  (termed  $a(i)$ ) and the average dissimilarity of  $i$  to all objects of the nearest neighbour cluster (termed  $b(i)$ ):

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \in [-1, 1] \quad (3.19)$$

An observation  $i$  with a score  $s(i)$  close to 1 has a high between-cluster dissimilarity in relation to the within-cluster dissimilarity and would be considered as well assigned, whereas a value for  $s(i)$  close to -1 is a strong indication for a wrong assignment. Accordingly, the higher the average silhouette width  $\bar{S}$  across all objects, the higher the confidence for the overall clustering result. An estimate of the “optimal” number of clusters  $k$ , in terms of homogeneity and separation) can therefore be obtained by choosing the  $k$  that maximises  $\bar{S}$ .

A similar measure, the *Dunn index* [228], computes the ratio between smallest inter- and largest intra-cluster distance. Formally, for a given clustering, where  $c_i$  represents cluster  $i$ , the Dunn index  $D$  is defined as follows:

$$D = \min_{1 \leq i \leq n} \left\{ \min_{1 \leq j \leq n, i \neq j} \left\{ \frac{d(c_i, c_j)}{\max_{1 \leq k \leq n} d'(c_k)} \right\} \right\} \quad (3.20)$$

where  $d(c_i, c_j)$  is the *inter-cluster* distance between clusters  $c_i$  and  $c_j$ ,  $d'(c_k)$  is the *intra-cluster* distance of cluster  $c_k$  and  $n$  is the number of clusters.

Another popular compactness-based validity index is the *C-index* [229] which can be computed as follows:

$$C = \frac{S - S_{min}}{S_{max} - S_{min}} \quad (3.21)$$

where  $S$  is the sum of distances over all pairs of objects from the same cluster. If the number of these pairs is  $l$ , then  $S_{min}$  is the sum of the  $l$  smallest distances, considering *all* pairs of objects (including objects assigned to different clusters), and accordingly,  $S_{max}$  is the sum of the  $l$  largest distances across all pairs. Since a good cluster should be as compact as possible, and the compactness of a cluster is represented by  $S$ , with the worst-case being  $S = S_{max}$ , the smaller the average of  $C$  across all clusters, the better the clustering result.

Finally, one of the statistically most effective internal cluster validity indices is the *Calinski-Harabasz* index. This index ( $G$ ) performed best among 30 indices evaluated on synthetic datasets in a well-known study by Milligan and Cooper [230] and is defined as follows:

$$G = \frac{B}{c-1} / \frac{W}{n-c} \quad (3.22)$$

where  $B$  is the sum of the squares of the distances between the cluster centroids and the mean of all objects in all clusters (“between-cluster sum of squares”),  $W$  is the sum of the squares of the distance between all objects and the class centroid to which the object belongs (“within-cluster sum of squares”),  $n$  is the total number of features and  $c$  is the total number of clusters. Again, this validity measure relies on maximising the between-cluster distances and minimising the within-cluster distances (i.e. the larger the score, the better the clustering result), and also accounts for the number of clusters  $c$ , so that the score does not improve with increasing  $c$ .

A fundamentally different approach towards cluster validity assessment is employed by the *Gap statistic* [142], proposed by Tibshirani *et al.*, which makes use of a reference distribution as a null model. The idea is to compute the change in cluster dispersion, measured as the pooled within-cluster sum of squares around the cluster means in terms of the Euclidean distance, for different numbers of clusters  $k$ , and compare it to that expected under an appropriate null distribution. The parameter  $k$  is then chosen such that the difference between the observed value and the null model, called the *gap*, is maximal. A high gap indicates a clustering result with high significance (for a theoretical motivation, see [142]). Moreover, a simulation study for a uniform reference distribution showed that the Gap statistic outperforms many other well-known validation methods from the literature.

**Robustness-based measures:** Stability-based validity measures evaluate how robust a clustering result is against noise or removal of data. These measures are computed by creating new artificial datasets, removing single columns from the original data, or introducing artificial noise into the data, and comparing the new clustering results with the original result using a cross-classification table. For example, the *average proportion of non-overlap* (APN) is the average percentage of observations which are placed in diverse clusters

in both results, the *average distance* (AD) measures the average distance between samples assigned to the same cluster for both results, and similarly, the average distance between means (ADM) represents the average distance between cluster centroids. Another stability measure is the *figure of merit* (FOM), which is given by the average intra-cluster variance of a randomly removed observation, according to a clustering with the remaining samples. Thus, by choosing the clustering parameters such that the APN, AD, ADM and FOM measures are minimised, a more stable and reliable clustering result can be obtained (see [231] and [232] for more details and implementations of these validity indices).

**Agreement with a reference partition:** Another group of methods, which are often also referred to as “co-clusteredness” indices, can only be used when a reliable, external reference partition of the samples is available. In this case, a new clustering result can be evaluated by a measure for the similarity between the clustering output ( $C$ ) and the reference partition ( $P$ ) as a “ground truth”. These methods are for example used to evaluate new clustering methods on data sets for which a good clustering result is already available (e.g. obtained from another algorithm), or to check whether a categorisation of the samples by human experts (e.g. clinical tumour grades for cancer patients) matches to natural groupings within the experimental data, identified by a clustering method.

In order to compare clustering partitions, a  $n \times n$  binary matrix  $C$  can be built ( $n$  = number of instances), setting  $C_{ij} = 1$  if instance  $i$  and  $j$  belong to the same cluster and  $C_{ij} = 0$  otherwise (and analogously, matrix  $P_{ij}$  can be obtained from the reference partition). By counting the number of true positive matches TP for these matrices (where  $C_{ij} = P_{ij} = 1$ ) and respectively, the true negatives (TN,  $C_{ij} = P_{ij} = 0$ ), false positives (FP,  $C_{ij} = 1, P_{ij} = 0$ ) and false negatives (FN,  $C_{ij} = 0, P_{ij} = 1$ ), a variety of similarity and dissimilarity estimates can be extracted from this data:

**Similarity indices:**

$$\text{Rand statistic: } RI = \frac{TP + TN}{TP + FP + TN + FN} \quad (3.23)$$

$$\text{Jaccard coefficient: } JC = \frac{TP}{TP + FP + FN} \quad (3.24)$$

$$\text{Folkes and Mallows index: } FM = \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}} \quad (3.25)$$

**Dissimilarity indices:**

$$\text{Minkowski measure: } MI = \sqrt{\frac{FP + FN}{TP + FN}} \quad (3.26)$$

Moreover, the *corrected* or *adjusted Rand index* (ARI) [233] normalises the original rand statistic (see above) so that its maximum is one and its expected value is zero when random partitioning is used (this statistic can be seen as a special form of Matthew’s correlation coefficient for pairwise assignments of observations):

$$\text{Adjusted Rand Index: } ARI = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TN + FN) \cdot (TP + FN) \cdot (TN + FP)}} \quad (3.27)$$

Since the ARI accounts for both specificity and sensitivity, and normalises the results for better comparability, it is often regarded as the method of choice.

If no reference clustering results are available, single outcomes can also be evaluated using estimates of the probability that the clusters were formed by chance, derived from the analysis of additional external data. For example, given the known membership of genes in the functional categories of the MIPS database [93], Tavazoie *et al.* proposed a p-value significance score for clusters using the hypergeometric distribution [234]. This p-value corresponds to the probability of observing  $k$  or more genes belonging to the same functional category in the same cluster ( $f$  is the total number of genes in a functional category and  $g$  the total number of genes across all categories):

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{f}{i} \binom{g-f}{n-i}}{\binom{g}{n}} \quad (3.28)$$

A limitation of this method is that many genetic probes in microarray studies have not yet been functionally annotated and predictions of the membership to a functional category, e.g. using sequence homology, might often not be reliable enough.

In practice, the decision on a suitable number of clusters should be made by considering the combined information from different validity measures to account for the diverse possibilities to measure the quality of clustering results. Moreover, an experimenter might also want to inspect the results for different numbers of clusters, since multiple biologically meaningful patterns might occur in the data and only be identified, when interpreting the data using different stable partitions. A more detailed overview on different cluster validity measures can be found in a review by Boutin and Hascoët [235].

### 3.5 Class Prediction (Supervised Machine Learning)

Many large-scale transcriptomics, proteomics and metabolomics datasets do not just consist of measurements of the abundance of certain molecules (mRNA, proteins and metabolites), but also contain additional annotation data for both the samples and the attributes. In particular, class labels representing the biological conditions of the samples (e.g. categories like “normal tissue” and “tumour tissue”) or numerical data (e.g. survival times) are available in a majority of cases. In addition to the unsupervised dimensionality reduction and clustering methods considered in the previous sections, the information from these additional dependent variables enables the application of *supervised analysis* methods. These techniques include supervised feature selection methods (see section on feature selection above), but more importantly, classification and regression methods, which enable the experimenter to find a predictive function that relates the input features to the given outcome variable(s). Such predictive models can be useful both for the interpretation of the data, explaining how certain features (molecules) affect the outcome (the biological conditions), but also to predict the outcome for new, unlabelled samples. These predictions can be of great practical use in many bioscientific and biomedical applications, including clinical diagnosis and prognosis, especially when the predicted attribute value cannot be measured directly or the measurement would be too difficult or expensive to be repeated many times (see chapter 2 on biomarker discovery). Moreover, if the predictive model is not too complex for human interpretation, model inspection might help to improve the understanding of a biological process (e.g. a studied disease) and to find new ways of influencing it (e.g. by identifying new drug targets).

In general, supervised analysis methods can be grouped into regression techniques to predict numerical outputs, e.g. the expected survival time for a tumour patient, and classification methods to estimate categorical outputs, e.g. the assignment of a patient's sample to a specific disease class. When analysing gene expression data, prediction approaches also differ depending on whether time-series data or relative expression levels of different tissue types are analysed. Moreover, similar to clustering approaches, prediction methods can be applied both to the columns (representing samples) and rows (representing genes, proteins or metabolites) of a data matrix. These two problems are structurally very dissimilar, due to the large difference between the number of samples and the number of features in a high-throughput experiment. Sample classification requires feature selection and is affected by the multiple testing problem, while gene function classification is complicated by small numbers of features (samples) and large numbers of observations (genes), many of which are not differentially expressed under the studied conditions. This section will focus on microarray sample classification and regression and only provide a general introduction on different types of approaches, discussing representative algorithms (see figure 3.9 for an overview).

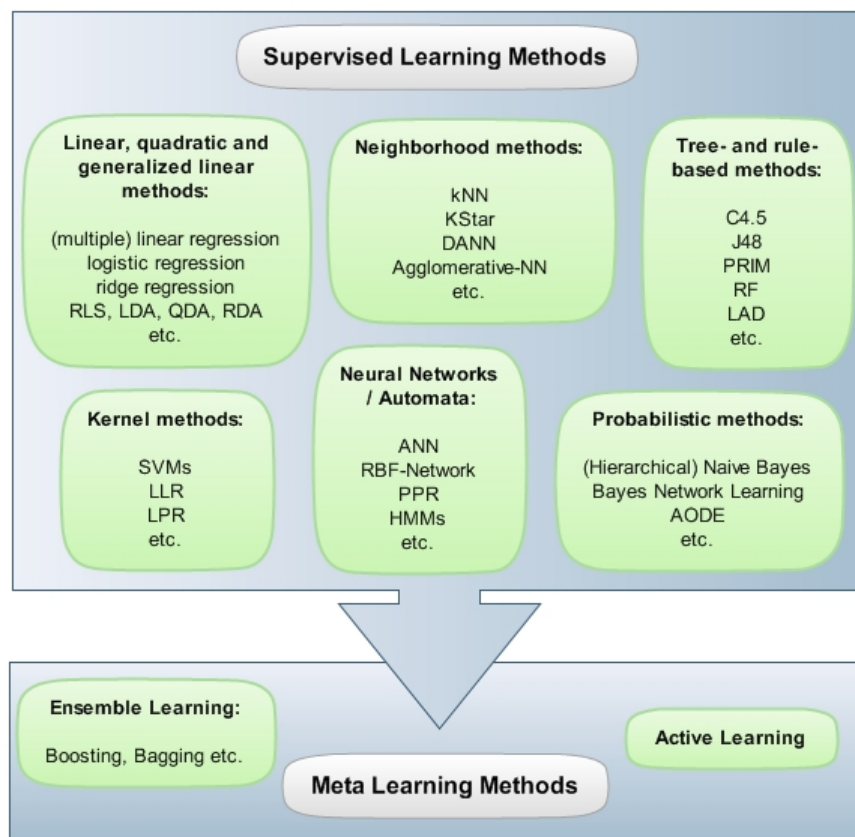


Figure 3.9: Overview of supervised learning methods and meta learning techniques discussed in this survey (non-comprehensive)

### Sample classification and regression

When high-throughput technologies like microarrays were first introduced, applying supervised analyses to this type of data had first been considered as infeasible due to the small sample sizes, the noise and the multiple testing problem in gene selection. Instead, the results from unsupervised clustering of samples were used in order to identify subclasses and estimate to which of these groups a new sample most likely

belongs [236]. However, though unsupervised analysis provides a useful instrument for class discovery, the assignment of a sample to a class according to its distances to the class centroids often led to unsatisfactory prediction results. Thus, when larger data sets became available and more powerful methods for dimension reduction and combinatorial feature selection were applied, research groups started to use well-known supervised learning techniques, which had already been applied successfully in other scientific fields. The following paragraphs will discuss the benefits and limitations of these generic machine learning methods, as well as extensions specifically tailored for microarray gene expression analysis.

**Linear methods and extensions:** Although simple prediction methods using linear functions as decision boundaries like linear regression and Fisher’s Linear Discriminant Analysis [237] are not adequate for the analysis of high-dimensional and noisy datasets, several variants of these approaches provide competitive prediction results obtained from human-interpretable models.

One possibility to make simple linear classifiers applicable to microarray data is to use regularization techniques. In particular, the *regularized least-squares* (RLS) approach has been very successful in microarray sample classification. RLS minimises the Residual Sum of Squares (RSS) in combination with a penalty term limiting model complexity:

$$\min_w = \frac{1}{l} \sum_{i=1}^l l(y_i - w \cdot x_i)^2 + \lambda(|w|^2 - \alpha) \quad (3.29)$$

where  $l$  is the size of the training set,  $x_i \in \mathbb{R}^n$  are the inputs,  $y_i \in \{0,1\}$  the outputs,  $\alpha \in \mathbb{R}$  and  $\lambda$  is a free regularization parameter that can be chosen using nested cross-validation. The major benefit of RLS is that solving a classification problem by minimising the expression in equation 3.29 only requires solving a linear system of order equal to the number of features or the number of training samples [238]. Another special advantage is the possibility to measure the leave-one-out (LOO) error by only training a single machine learning model. Thus, the computational cost is low and the obtained linear models are relatively simple to interpret, although they are typically not as sparse as those produced by other regularized learners. Ancona *et al.* compared the accuracy and number of selected genes for RLS with support vector machines (SVMs) as a state-of-the-art technique for microarray cancer classification [238] (see section on kernel methods below). The signal-to-noise filter and a recursive feature elimination strategy (RFE, see section on feature selection methods) were used to determine relevant attribute sets. Evaluating the LOOCV accuracy for these methods on three microarray data sets (for leukaemia, colon cancer and a dataset with multiple cancer types) the authors concluded that in spite of the simplicity of RLS models, a similar generalisation capability to SVMs is obtained. This might however not be the case for other microarray data sets, in which non-linear dependencies play a more important role.

Regularized extensions also exist for the most well-known classification approach using linear decision boundaries, Fisher’s Linear Discriminant Analysis (LDA). One of these extensions explicitly dedicated to gene array data is *Prediction Analysis for Microarrays* (PAM, not to be confused with the “Partitioning around medoids” clustering method with the same abbreviation) by Tibshirani *et al.* [239]. PAM, also known as the *nearest shrunken centroid* method, addresses the “curse of dimensionality” problem by directly combining feature selection and classification within the algorithm. This is achieved by shrinking the class centroids towards the centroid of the entire dataset to reduce the effect of outliers, and applying a soft thresholding to set many of the class centroids’ vector components to 0. Thus, some features (genes) for which all centroids have a 0-entry are completely removed from the class separation problem without applying a separate feature selection method. In order to choose the amount of shrinkage, which can be un-

derstood as a bias/variance trade-off parameter (see supervised feature selection section), a cross-validation procedure is applied and the misclassification error is minimised. A major benefit of the PAM approach is that the influence of noisy genes is reduced by the shrinkage procedure and that the implicit gene selection is completely automated and interlinked with the prediction algorithm. Drawbacks of the PAM approach are the restrictive model assumptions, e.g. the covariance matrix is assumed to be diagonal, and the results greatly depend on the correct estimation of the regularization parameters.

A similar method which was also derived from a regularized version of LDA is the *shrunk centroids regularized discriminant analysis* (SCRDA), proposed more recently by Guo *et al.* [240]. Instead of assuming the covariance matrix to be diagonal, the matrix entries are estimated in a more general manner using a regularization that eliminates features from the classification problem. However, SCRDA cannot really achieve variable selection, since even genes that do not contribute to classification are still involved in the construction of the decision rule, as shown by another group [241]. Nevertheless, the authors demonstrated that SCRDA performs slightly better than PAM on several benchmark data sets. However, in this extension of LDA the decision boundaries can be non-linear and more difficult to interpret.

**Neighbourhood methods:** Neighbourhood-based prediction methods, which classify samples according to their proximity to known training samples (usually using a distance metric like the Euclidean distance), are often very successful in microarray analysis, in particular when being used in combination with other techniques. Simple and fast approaches like the k-nearest neighbours (k-NN) method, which assigns new unlabelled samples to the majority class of the k closest training samples in Euclidean space, have achieved good classification accuracies in combination with search metaheuristic like genetic algorithms using wrapper-based feature selection [180] (see feature selection section), and in combination with different noise-reduction filters [242].

The first extensions of the k-NN method provided runtime and memory allocation improvements by taking advantage of the redundancy in some datasets, clustering nearby data points together and storing these clusters instead of the data points (agglomerative nearest neighbour method [243]). Moreover, several extensions used other distance functions, e.g. the KStar (or K\*) method employs an entropy-based distance measure [244]. However, these methods and the original k-NN suffer considerably from the “curse of dimensionality”, providing biased predictions in high dimensions and assuming class conditional probabilities to be locally constant. Therefore an adaptive form of the k-NN method was introduced [245], which applies a linear discriminant analysis locally to estimate an alternative, adaptive metric for the neighbourhood computation. This *Discriminant adaptive nearest neighbour* (DANN) classification method first computes local decision boundaries around the class centroids (similar to Fisher’s LDA), and then adapts the neighbourhoods, by shrinking them in directions orthogonal to the decision boundaries, and elongating them parallel to the decision boundaries. In contrast to classical spherical neighbourhoods, the resulting ellipsoidal neighbourhoods are less likely to overlap with the local decision boundaries. The authors presented various simulated and real-world example datasets on which DANN reaches superior performance both in comparison with the k-NN method and LDA. Since a multitude of methods are available to compute linear decision boundaries, a variant of the DANN method was later introduced by Domeniconi *et al.*, who computed locally adaptive distance functions using SVM decision boundaries [246].

A further idea to adjust the neighbourhood computation uses *semi-supervised learning* [247], i.e. exploiting information from additionally available unlabelled data. Driessens *et al.* presented a corresponding classifier, called YATSI (Yet Another Two Stage Idea), which takes unlabelled data into account and re-weights



the k-NN algorithm using both labelled and unlabelled data [248]. Comparative evaluations on real-world data suggested that this technique significantly outperforms the original k-NN method, if sufficient unlabelled data is available.

More recently, a new method, ROC-kNN, was proposed to optimise the distance function for k-NN with regard to the classifiers area under the Receiver Operating Characteristic (AUROC) [249]. Specifically, a new distance function is derived by a procedure for re-weighting features, which first computes the ROC curve for a series of point pairs, obtained from a threshold value for a single variable  $x_i$  and the corresponding outcome class  $y_i$ , and then uses the AUROC as a feature weight. To select the threshold values for a feature required to generate the corresponding ROC curve, only the feature values on the interval between the two observations whose distance is calculated are used, widening this interval only if it covers less 50% of all values for the feature in the dataset. When comparing ROC-kNN with other approaches on 12 datasets, the method was always superior to the classical k-NN algorithm, and on some datasets also outperformed state-of-the-art non k-NN classifiers. However, if the sample size is small like in typical microarray studies, the computation of ROC-based weights is less reliable and support vector machines still tend to provide better performance.

In summary, when using nearest neighbourhood techniques as stand-alone methods, they tend to perform well on high-dimensional and noisy data only when adjusting the neighbourhood or distance function definition to the input data. However, even simple variants of the k-NN method can be of great practical use when being combined with other prediction and filtering methods, as well as search space exploration metaheuristics in wrapper-based feature selection approaches.

**Kernel methods:** Generic kernel-based machine learning methods, and in particular standard implementations of support vector machines (SVMs) with a linear kernel, belong to the microarray classification methods achieving the highest accuracies on many datasets. The main benefit of these methods lies in the so-called “kernel trick”, which makes observations linearly separable by mapping them from the original space  $O$  into another, higher-dimensional inner-product space  $S$ , using a kernel function to describe the inner products in space  $S$ . A linear classification problem in the space  $S$  will then correspond to the original, non-linear classification problem in  $O$ . As long as a linear kernel is used, the models can also be interpreted easily by inspecting the feature weights, although an additional feature selection is often required to reduce the number of features. Nevertheless, various research groups have tried to optimise SVM-based prediction for the specific task of analysing high-dimensional, noisy data with small sample sizes.

A prominent example for these extensions was the introduction of the *maximum entropy kernel* (ME) [250]. Instead of employing a pre-defined kernel function, the ME kernel approach uses more adaptive functions for distance computations, similar to some of the techniques discussed in the section on neighbourhood-based prediction methods. More specifically, the kernel is generated in an entropy maximisation process using the input sample distance matrices as constraints. Using any type of distance or similarity data as input, the procedure expands the distances between most sample pairs in the feature space, except for the most similar samples, which are held together by the constraints. In the resulting feature space, heterogeneous datasets tend to become less entangled, and the problem of finding a discriminant boundary is greatly simplified. This new SVM approach was compared to SVMs with other kernels (linear, polynomial, RBF, and two other distance-based kernels) on three gene array benchmark datasets using the matrix of Euclidean distances as input for the kernel generation. The genes to be used for classification were selected by a two-sample t-statistic, embedded into each cycle of a LOOCV procedure. Average accuracies above 87.5% were reached on all datasets, even when introducing high levels of noise into a dataset. Comparing the results to

other kernels for feature selections with 8 to 296 attributes, the ME kernel reached better results in the great majority of cases. On the whole, the results suggest that the adaptive kernel can reduce the effects of noise and thus provide a better separation between the sample classes,

In addition to improvements with regard to the used kernel functions, recent research efforts have also been devoted to finding more suitable definitions of the penalty term [251] and the loss function [252] for SVMs. Moreover, special advantages of SVM-based models in the analysis of small sample high-dimensional data have been exploited by combining SVMs with the *active learning* paradigm [253]. In contrast to classical machine learning techniques, which use all the training data at once and build a predictive model in a single pass, active learning attempts to make an informed selection of the labelled instances to train the classifier and can iteratively update an existing model, when new data becomes available. This approach can help to circumvent problems arising from class imbalances in the training data, and a limited availability of labelled data. The procedure proposed by Liu *et al.* [253] initially chooses a random training sample from both classes in a binary classification problem and builds a preliminary classifier. While there are still unlabelled samples, the current classifier is applied to each of them and the  $m$  samples which are most informative for the classifier are selected, labelled and a new classifier is trained on the updated set of labelled samples. While this procedure is repeated iteratively, new data obtained from a recent experiment can be added dynamically to the sample pool. Applying this SVM-based active learning algorithm on three microarray data sets (colon cancer, lung cancer and prostate cancer), areas under the ROC curves above 0.81 were obtained for all data sets ( $m = 1$  or  $5$ ), whereas the AUCs were below 0.5 when using passive learning. Moreover, the results showed that active learning could reduce the *cost* for classification, in terms of the required number of training samples. For example, in the case of the lung cancer data, only 31 labelled examples were needed to find 96% of the total positives. The observation that active learning approaches work particularly well with margin-based classifiers like SVMs has also been confirmed in other studies [254, 255].

In addition to the successful application of support vector classification and regression techniques in many domains of high-dimensional data analysis, other kernel-based prediction approaches like Gaussian process (GP) regression [256] and Kernel Fisher discriminant analysis (KFDA) [257] have recently been considered as alternatives, but in comparison to SVM approaches currently only limited information on their utility for microarray data analysis is available in the literature.

Overall, kernel-based approaches include several state-of-the-art methods in terms of predictive performance and runtime efficiency, however, especially for the non-linear approaches, the direct interpretation and extraction of biological knowledge from the models is often difficult and prevents a wider acceptance in the scientific community.

**Neural network based methods:** Artificial neural networks (ANNs), inspired by the mechanisms of real biological networks, have a long tradition in machine learning [258] and are widely used in many domains of bioinformatics, due to their capacity to learn non-linear decision boundaries and their flexibility to deal with noisy data. The classical multilayer perceptron (MLP) feed-forward approach for these graph-based models consist of layers of nodes (representing neurons) connected by directed, weighted edges (representing possible paths to forward signals), transmitting the input data from an *input layer* of nodes through a user-defined number of *hidden layers* to the *output layer*, providing the prediction results of the model. To obtain these predictions, the data are processed at each node, typically by computing a nonlinear edge-weighted sum of the incoming data and processing it by a pre-defined *activation function* (usually a sigmoid function like the hyperbolic tangent). A model can be trained by adjusting initial edge weights using a back-

propagation algorithm [6] to minimise a given cost function (e.g. the mean-squared error between the target values and the network's output). Since ANNs require at least one hidden layer to provide non-linear models, this methodology has been criticized for providing models that are difficult to interpret and easily overtrained, due to their large number of parameters and the lack of general rules to choose the number of hidden nodes [259]. However, state-of-the-art performance has been reached with ANNs in many fields of bioinformatics (e.g. to predict properties of proteins like residue contacts [260], disulfide connectivity [261], or the topology of transmembrane domains [262]), and several extension to the standard MLP feed-forward approach have been proposed for high-dimensional, noisy biological data.

Although ANNs are capable of feature selection to a certain extent, e.g. by reducing the weight of edges with small contribution to the predictive power of the system to zero, the large number of uninformative features in typical high-throughput experimental data poses a major problem for these network-based models. Therefore, the first approaches using ANNs for supervised microarray analysis have applied different pre-filtering approaches before training a network model. For example, Khan *et al.*, who aimed at classifying microarray samples for small, round blue-cell tumours (SRBCTs) into four diagnostic categories, first filtered out genes with small intensity levels across the samples, and then extracted 10 features corresponding to the 10 dominant principal components from a projection of the genes using PCA eigenvectors [263]. Linear ANN models were then calibrated using these 10 features within a 3-fold cross-validation procedure, providing 100% accuracy in all cases.

More recent approaches for microarray sample classification employ hidden layers to obtain non-linear classifiers, e.g. the method by Lancashire *et al.* [264, 265] contained a single hidden layer with between 2 to 5 nodes. Instead of applying a separate feature selection method, an ANN model was trained for each single feature over 50 randomly selected sample subsets. Average mean squared error (MSE) values were calculated over the predictions of the 50 models for separate test sets to rank the features. Then, the best out of the total  $n$  features was selected to generate the final model, and sequentially, each of the  $n-1$  other features were added to this model, creating  $n-1$  two-feature models and applying the above training and performance evaluation procedure again. This process was repeated iteratively until no significant improvement was obtained from adding further features to the model. Using this procedure to classify breast cancer samples from a separate validation set into different clinical categories, a set of 9 genes, capable of predicting distant metastases with 98% accuracy, was obtained. On a second independent dataset of 295 samples, the model achieved 63% accuracy, indicating that the model also possesses a certain level of predictive power in a cross-study classification setting, where external samples are affected by a study-specific bias.

To automatically improve the architecture of neural networks, several evolutionary computation based methods have been proposed, e.g. genetic programming has been employed to optimise the ANN architecture for modelling and detecting gene-gene interactions in human disease studies [266, 267]. The resulting models had a better predictive performance and were superior in detecting gene-gene interactions when non-functional polymorphisms are present in the data.

Recently, several efforts have been made to increase the interpretability of ANN models. Some research groups have presented methods to extract simple "if-then-else" classification rules from neural networks, which outperform classical decision trees on real-world data [268, 269]. Specifically, *fuzzy-neuro* networks were proposed, which enable fuzzy rule-based classification using ANNs and can account for uncertainty in the data. Most of these approaches use fuzzy sets to transform the continuous input data into linguistic terms, apply an ANN and extract decision rules from it [270–272]. For example, a classical feed-forward

network structure has been combined with a special hidden layer with the same number of nodes than the input layer, in which the input is transformed into 3 linguistic groups (“small”, “medium”, “large”) using a Gaussian membership function [273]. The linguistic features with the highest membership values are assigned to a dedicated class (+1), whereas all other remaining features receive a different class label (-1). Weighted links connect these features from the special hidden layer with the output layer, containing one node per biological condition in the dataset and applying a sigmoid activation function on the incoming data. Thus, by training the weights of these links using back-propagation, an attribute selection is performed that selects the most informative linguistic features with regard to the class separation problem. From a model trained with this procedure, two types of decision rules can be extracted from the top  $n$  features with largest weights in the network: Disjunctive decision rules for each class (termed “simple OR rules”), or decision rules considering the order of important linguistic features and the class order (termed “layered rules”). In both cases, a small number  $n$  is chosen (e.g.  $n = 9$ ) and each of the  $n$  top-ranked features is used in the rules. Applying this fuzzy-neuro learning approach to a colon cancer dataset, cross-validated accuracies above 90% were obtained in comparison to SVM models and conventional ANN approaches, all reaching average accuracies below 82% with similar numbers of features.

These recent advances in neural network based classification of high-dimensional biological data suggest that hybrids of the original feed-forward ANN approach with rule-based methods can generate interpretable models comparable with the state-of-the-art in terms of accuracy.

**Bayesian learning methods:** Due to the high uncertainty in single gene or protein expression values in large-scale array datasets, probabilistic machine learning methods are a natural choice for the analysis of this data. Most of these methods apply Bayes’ theorem, i.e. they calculate the posterior probability of a hypothesis  $H$  given an evidence  $E$  ( $p(H|E)$ ) from the prior probabilities of  $E$  and  $H$  and the likelihood of  $E$  given  $H$ :

$$P(H|E) = \frac{P(H)P(E|H)}{P(E)} \quad (3.30)$$

This approach enables the experimenter to include prior knowledge into the estimation of predictive functions. The simplest Bayesian classifier, Naive Bayes, assumes that all features contribute independently to the prediction of the outcome classes and estimates model parameters like the class priors from the relative frequencies observed on the training set (corresponding to a maximum likelihood estimation). The trained probability model can then be used to classify new samples, by assigning them to the class (hypothesis) with the highest estimated conditional probability. Although Naive Bayes (NB) classifiers tend to be outperformed by many other machine learning methods like SVMs and RFs on high-dimensional noisy data, they are very efficient and work particularly well in wrapper-based feature selection approaches (see feature selection section). Moreover, NB does not contain any tuneable parameters or involve any model selection.

In order to address the limitations of the feature-independence assumption behind NB, the *Averaged One-Dependence Estimators* (AODE) approach was developed, using a weaker independence assumption than NB. Although this method often significantly outperforms NB, the computational costs only increase linearly with the number of samples [274]. However, for high-dimensional datasets the application of AODE can still be infeasible, since the runtime also has a quadratic dependence on the number of features.

An alternative approach to make the NB method applicable to high-dimensional data was therefore proposed by Bressan and Vitria using the new feature extraction method *class-conditional independent component analysis* (CC-ICA) [275]. CC-ICA enables the application of NB in a derived feature space obtained

from an independent component analysis (ICA), in which the class-conditional independence assumption is fulfilled. This pre-processing approach improved the accuracy of the original NB algorithm on many real-world datasets, however on microarray datasets with a small sample size per class, the ICA transformation cannot be applied. To solve this problem, Fan *et al.* introduced a *partition-conditional independent component analysis* (PC-ICA) [276], which represents a compromise between ICA and CC-ICA for feature extraction, splitting the samples into different partitions in a manner that enables the application of ICA-based feature extraction within each partition (where partitions can consist of samples from multiple classes). To obtain these partitions, a hierarchical clustering method can be applied, e.g. average linkage hierarchical clustering. Only if enough samples per class are available to apply ICA directly on each class, no partitioning will be applied, and PC-ICA becomes equivalent to CC-ICA. When evaluating the performance of PC-ICA based NB classification on two microarray cancer datasets (leukaemia and lung cancer), PC-ICA overall provided higher average accuracies than classical ICA.

Other Bayesian classification approaches avoid the independence assumption in NB completely, and instead take the dependency between features directly into account. A recent example is the *network-based sparse Bayesian classifier* (NBSBC), which models the dependency between features in a graph representation, where nodes correspond to features and edges connect features that should be either both included or excluded in the final prediction model [277]. The NBSBC approach implements an approximate Bayesian inference using the *expectation propagation algorithm* [183, 278] and was evaluated on four real-world classification problems in comparison against two other classifiers, which also incorporate information on feature dependencies from a network into the analysis - the network-based SVM (NBSVM) [279] and the Graph Lasso (GL) [280]. On three of the four datasets, and in particular on microarray data, NBSBC reached the highest predictive performance and always outperformed its predecessor, the sparse Bayesian classifier (SBC), which results from NBSBC when ignoring the feature dependencies.

On the whole, Bayesian classification methods have particular strengths in the analysis of noisy biological data, but since an exact Bayesian inference is currently not feasible on high-dimensional data, approximate inference methods have to be used.

**Rule- and tree-based methods:** A major drawback of some of the most accurate prediction methods for high-dimensional biological data, affecting both SVMs with non-linear kernels and classical Artificial Neural Networks (ANNs), is the complexity of the generated models, which often impedes human interpretation. In clinical applications of machine learning, simple decision trees (e.g. C4.5 [281] and CART [282]) are therefore often preferred in spite of their weaker performance on many datasets, because these models can easily be visualised and interpreted as combinations of simple “if-then-else” decision rules (often they are referred to as *white-box* models, as opposed to *black-box* models like SVMs, ANNs, etc.). The rationale is that a model should be useful in terms of helping the experimenter to understand the data, and robust in terms of providing simple rules that are applicable also to data from other experimental platforms.

However, classical decision trees tend to have a poor predictive performance in domains with high noise levels, large numbers of uninformative features, inconsistencies and uncertainty [283]. Features can typically not be weighted but are either fully included into the model as tree nodes or completely excluded, and the only complexity reduction consists in pruning the tree after completing the tree building procedure. Moreover, the tree nodes, corresponding to splitting rules obtained by choosing an attribute and a split point, are normally added using a simple greedy search procedure. In order to obtain a better predictive performance while preserving high model interpretability, various new rule-based prediction methods have been developed in recent years.

A very simple, but robust and effective rule learning approach for microarray sample classification is the *top-scoring pair(s)* (TSP) classifier by Geman *et al.* [284]. This method only compares the expression levels within a single pair of genes, but applies an exhaustive search through all pairs of genes to identify the most discriminative feature pairs. The corresponding scoring function  $s(i, j)$  simply computes frequency-based probability estimate for observing lower expression levels in a gene  $i$  in relation to a gene  $j$  in a sample class  $c = 1$ , and the opposite relation in sample class  $c = 2$ :

$$p_{ij} = P(X_i < X_j | c = 1, 2) \quad (3.31)$$

$$s(i, j) = |p_{ij}(1) - p_{ij}(2)| \quad (3.32)$$

TSPs provide very simple decision rules of the form “if gene A has a significantly higher expression level than gene B, assign the corresponding sample to class 1, otherwise to class 2”. Since these rules rely on relative comparisons of expression values rather than on absolute mRNA abundances and fitted threshold values, they are very robust with regard to study-specific bias. Moreover, the method is parameter-free and in spite of the quadratic runtime during the exhaustive search for TSPs, the computation is feasible on typical microarray data, due to the simple scoring function. To evaluate the statistical significance of the TSP-scores, a simple non-parametric permutation analysis can be performed. However, when testing the approach on real-world data, the assumption that a single gene pair, or a small number of TSPs can already provide accurate sample classification results is not fulfilled on all datasets, although in many cases, state-of-the-art accuracy has been achieved (e.g. on leukaemia, prostate and breast cancer microarray data, average LOOCV accuracies above 79% were reached).

Some of the limitations of the original TSP approach can be overcome with simple extensions, e.g. enabling TSP to handle multi-class problems, or to create hierarchical and ordered combinations of multiple TSP decision rules [285], as well as weighted rules [286]. The method can also be used as a feature selection approach in combination with other machine learning techniques like SVMs [287]. Moreover, a new variant of the TSP approach, integrating information from cellular pathways into the analysis, was developed as part of this doctoral project, and will be presented in chapter 7 of this thesis.

More recently, Li *et al.* introduced an alternative prediction model using gene pairs and simple linear models to explain the relation between the gene expression values in a sample class [288]. More specifically, for two genes  $g_1$  and  $g_2$ , the authors fit a linear regression model, but only in the first group of samples (class 1). In class 1, the expression values of  $g_1$  can then be predicted from  $g_2$  using the model, whereas when applying this model in class 2, if the relation between  $g_1$  and  $g_2$  has changed, a large deviation (or *bias*) between predicted and observed values will be obtained. Accordingly, the two sample types can be distinguished by computing the difference of the predicted biases. By contrast, the TSP approach classifies samples according to the difference of frequency counts for the relations of expression values in the sample classes (see above). As further extensions, Lie *et al.* combine rules to majority-vote ensembles and use a GA to speed up the search process for the top-scoring gene pairs. Experimental results using LOOCV on leukaemia, lung, breast and colon cancer data showed that the method reaches similar average accuracies (between 90 and 100%) like other state-of-the-art approaches, using only a small number of genes.

Since rule-based learning models have particular benefits in terms of interpretability, the study of new approaches in this category, including the evolutionary machine learning system *BioHEL* (see chapter 4 for details) and the self-devised *Top-scoring pathway pairs method* (see chapter 7), has been one of the major goals of this doctoral project.

**Ensemble learning methods:**

The great variety of successful prediction methods for microarray sample classification discussed in the previous sections suggests that a multitude of diverse methods can deal effectively with high-dimensional, noisy data with small sample sizes. Thus, by combining the models obtained from these learning algorithms into a unified model, their different strengths might be better exploited.

The theoretical benefits of ensemble learning, in particular its variance-reducing effects, have already been described in detail elsewhere [8]. Briefly, ensemble learning methods can be applied successfully if the base classifiers that are to be combined into a single model, provide a higher accuracy than a random class assignment and are diverse, so that they can complement each other to provide better predictions than the individual base classifiers [289]. However, a common drawback of ensemble learning is the high complexity of the resulting model, especially if the number of base classifiers and their individual complexity is high. This problem can be alleviated by post-processing the base classifier models using simple statistics, e.g. by computing averaged ranks of the features included in the base classifiers to obtain a robust and informative ranking of features. Moreover, the number of base classifiers can be adjusted as a trade-off parameter to obtain an optimal balance between the ensemble model's bias and variance (see the "bias-and-variance trade-off" described in the "supervised feature selection" section). Thus, in spite of potential drawbacks in terms of model complexity, considering different ensemble prediction approaches on the basis of adequate single-algorithm classifiers is a useful technique to better exploit the information content in biological datasets and reduce problems associated with high variance and the curse of dimensionality.

Most ensemble learning approaches are using *model averaging*, i.e. the base models are applied independently on the target samples and the ensemble prediction is obtained by computing an (optionally weighted) average across the base model predictions. For regression problems, this is achieved by computing the mean prediction for each sample, whereas for classification problems, a majority voting scheme can be adopted. More advanced ensemble learning techniques attempt to explicitly promote model diversity and/or improve the assignment of weights to the base classifiers, e.g. by considering their predictive accuracy and diversity in comparison to other base classifiers.

Three of the most wide-spread, generic ensemble learning techniques are bagging, boosting and stacking. *Bagging* stands short for *bootstrap aggregating*, a procedure introducing diversity into the base classifiers by using a resampling technique known as *bootstrapping* [290, 291]. Bootstrapping applies random sampling with replacement on an original dataset, to obtain equally sized resampled versions of this dataset. These *bootstrap resamples* of the original data can be used to create a more robust classifier by training independent prediction models on them and applying model averaging (see above) to obtain an ensemble classifier. This technique has been applied successfully in many scientific and industrial problem domains, including the analysis of high-dimensional biological data. For example, bagged ensembles of SVMs have been shown to provide more stable and equal or better sample classification accuracies than single SVMs on leukaemia and colon cancer microarray data [292].

Importantly, randomisation and resampling techniques can be applied both to the samples and to the features of a dataset. Due to the large number of features for typical microarray datasets, a commonly used technique is to only apply resampling on the samples and random selection on the features to obtain diverse input data for ensemble learning. One of the most well-known algorithms employing this approach is the *random forest* (RF) method introduced by Leo Breimann [190, 191]. The RF approach trains unpruned classification and regression trees (see section "Rule- and tree-based methods" above) using bagging to determine the input samples and a random candidate feature selection for each node in the tree. The di-

verse base classifiers obtained from this procedure are then combined to a model averaging ensemble. The RF algorithm belongs to the few machine learning techniques with known convergence proofs for the generalisation error [191], and is one of the most popular machine learning approaches next to SVMs. On large-scale bioscientific datasets, special RF variants often outperform other state-of-the-art approaches, e.g. for microarray sample classification, Amaratunga *et al.* replaced the random sampling of genes by a sampling procedure accounting for the significance of differential expression [293], Zhang *et al.* introduced a deterministic variant of random forests [294, 295], and to reduce the complexity of the models, Zhang and Wang proposed a method to find the smallest sub-forest that achieves the same prediction accuracy as a given large random forest model [296].

However, resampling and random selection techniques are not always the most effective methods to introduce diversity into the base classifiers of an ensemble learning approach. *Boosting* algorithms represent a further class of ensemble approaches, in which weighted base classifiers are iteratively added to a combined model and at the same time weights are assigned to the samples, such that misclassified samples receive higher weights than correctly classified samples. These sample weights are then used in subsequent iterations to add base classifiers that focus on improving the predictions for the misclassified samples, hence, diversity is introduced into the base classifiers by forcing them to have different prediction strengths on different groups of samples. Depending on the procedure to calculate the sample and classifier weights, a multitude of different boosting techniques have been developed, including the classical AdaBoost approach [297], a variant using a logistic regression cost functional (LogitBoost) [298], and a linear programming approach to combine optimally weighted base classifiers (LPBoost) [299]. However, Dettling and Bühlmann showed that these classical boosting methods are often not robust enough for noisy, high-dimensional microarray data, and developed a boosting variant in conjunction with decision trees, which provided increased performance on publicly available gene array data sets [300].

An alternative and very effective generic approach to learn weights for base classifiers is to use a meta-learning technique, i.e. to apply a higher-level machine learning algorithm in order to estimate the optimal weights. This approach, known as *stacking*, *stacked generalisation* or *blending* [301], uses the same cross-validation schemes that are also employed for model selection or for single-model parameter fitting, but applies them to fit the base classifier weights for optimally combining models into an additive ensemble. Stacking techniques have recently become very popular methods in the machine learning community, due to their success in one of the largest data mining competitions, the Netflix prize [302], where the two best-performing methods both used stacking. However, in bioscientific research, stacking methods have not yet been widely employed, since they tend to have a poor runtime performance on large-scale datasets and the ensemble models they generate are often difficult to interpret.

On the whole, ensemble learning methods provide an effective means to improve the results of single-algorithm classifiers. In many cases, potential drawbacks in terms of low model interpretability can be alleviated by using techniques like the search for smallest sub-forests in random forest models, or by computing statistics on the occurrence of features in different base classifier models. Due to the increasing number of available algorithms and data sources for the same biological problems, ensemble techniques capable of aggregating diverse information are likely to become even more relevant and valuable in the future.



## 3.6 Data Integration 1: Cross-Study Analysis

In the assessment of new microarray prediction methods, such as those discussed in the previous sections, a typical strategy to circumvent problems with small sample sizes and experimental bias in a single study is to train and test the method on a variety of different data sets. However, if the purpose of the study is to analyse the data rather than to compare algorithms, this approach does not really solve the small sample size problem, because microarray studies using different platforms or carried out in different laboratories are typically not directly comparable, even if the same cell type is analysed under the same conditions. Thus, without cross-platform normalisation, the predictor can often only be trained and tested successfully on data from a single platform with a precisely defined experimental procedure.

To alleviate the problems associated with small sample sizes, three basic approaches exist:

- using improvements in microarray technology, noise filtering and analysis techniques to obtain more robust models
- combining gene array data with external biological data (e.g. biological databases or clinical indices)
- integrating microarray data from different laboratories and platforms by *cross-study normalisation*

While the first two points are discussed in other dedicated sections of this chapter and make important contributions to the solution of the problem, the integration of similar microarray data from different studies is certainly the most promising approach to cope with the small sample size problem.

The first cross-platform integration techniques used simple transformations of the raw expression data, e.g. by *median rank scores* or *quantile discretization* [303]. The median rank scores approach replaces the gene expression values of a *target* dataset by the median expression values of genes from a *reference* study whose positions in a sorted vector correspond to the expression value rank in the target dataset. Although this transformation will to a certain degree result in a loss of information, the data distributions become comparable and the datasets can be combined.

In the second transformation approach, quantile discretization, each data set is discretized into the same number of bins using the quantiles of the array expression values as cut-points. The central bins are merged and all expression values are replaced by integers corresponding to the bin they have been assigned to. Again valuable information might be lost, but the transformed data sets have the same value ranges and the discretized data enables the application of machine learning methods which are not compatible with continuous data.

Both methods were applied to integrate example data sets and the transformed data for three pairs of studies (breast cancer, prostate cancer and acute myeloid leukaemia) were used to train SVM classifiers. The approaches both achieved cross-validation accuracies above 85% and outperformed models trained on the respective single data sets. The authors also observed positive effects of the integrative analysis on the gene selection results. Important differentially expressed genes, which are missed in either of the single-platform analyses are identified by the combined analysis.

A more recent method for integrative array analysis is the *XPN cross-platform normalisation method* by Shabalin *et al.* [14]. The approach applies linked gene and sample clustering on the studied data sets and is based on a block-linear error model. Specifically, every expression value  $x_{gsp}$  for a gene  $g$  in sample  $s$  of study  $p$  is assumed to be a scaled and shifted block mean with additional noise:

$$x_{gsp} = A_{\alpha^*(g), \beta_p^*(s), p} \cdot b_{gp} + c_{gp} + \sigma_{gp} \epsilon_{gp}$$

where  $A_{ijp}$  are the block means and the functions  $\alpha^*(g)$  and  $\beta_p^*(s)$  define groups of linked genes or samples, respectively. The model also contains platform- and gene-specific sensitivity ( $b_{gp}$ ) and offset ( $c_{gp}$ ) parameters and independent Gaussian noise variables ( $\epsilon_{gp}$ ).

In the XPN procedure, after pre-processing the data and selecting the set of common genes from two studies, the data are sample standardised and gene median centred (MC) to remove systematic differences (the result of this simple MC-normalisation is often already directly used in practice as a simple cross-study normalisation method). Next, the ordered sample vectors from the different studies are combined to a single matrix and K-means clustering is applied both to the rows and columns using random initial centroids. Multiple clusterings are computed to enable a simple form of model averaging. The gene clusters obtained in both studies can be linked together using the unique gene identifiers and summarised by an assignment function  $\alpha : G \rightarrow 1, \dots, K$ ; where  $G$  is the set of genes and  $K$  the total number of gene clusters. For the column clusters two assignment functions are needed:  $\beta_p : 1, \dots, n_p \rightarrow 1, \dots, L$ ; where  $p \in \{1, 2\}$  is the index of the study,  $n_p$  the number of samples in study  $p$  and  $L$  the number of sample clusters. Based on these mapping functions  $\alpha(g)$  and  $\beta_p(s)$  the unknown parameters for the block linear model can be obtained using maximum likelihood estimation. The required selection of the number of row and column clusters is made *a priori*, e.g. by using cluster validity indices. As indicated before, in order to increase robustness, the clustering procedure is repeated (at least 30 times) with varying random initial centroids and the average of the estimated expression values is the final result. Importantly, if no clustering solution is found that combines samples from different studies into a cluster, the procedure will terminate and the datasets cannot be combined.

The XPN algorithm was tested by the authors on three breast cancer datasets, considering each pair of datasets, and compared to the following alternative cross-study normalisation techniques: The MC-normalisation (see above), the *Empirical Bayes method* (EB) [304] and the *Distance Weighted Discrimination* (DWD) method [305, 306]. The EB method uses a model similar to that of the XPN approach:

$$x_{gsp} = \alpha_g + \gamma_{gp} + \delta_{gp} \sigma_g \epsilon_{gsp} \quad \epsilon_{gsp} \sim N(0, 1)$$

where the platform specific parameters  $\gamma_{gp}$  and  $\delta_{gp}$  are approximated by an empirical Bayes estimation and the remaining parameter values are computed using gene-wise ordinary least squares (OLS). The DWD method identifies a direction vector in which the projected samples from the two input datasets can be easily discriminated, and translates the samples along this direction until the corresponding sample groups have a large overlap.

The authors of the XPN approach propose validation methods for cross-platform data integration, which are used in their comparative evaluation. These methods aim both at identifying under-correction errors (the studies still have systematic differences) and over-correction errors (biological information was lost) and include measures for the centre and spread, the average Euclidean distance to the nearest array in another platform, the correlation between the data matrices before and after normalisation, the global integrative correlation [17], the correlation of t-statistics measuring the association between expression values and the outcome variable for each study, the prediction results for cross-platform sample classification and the preservation of significant genes in feature selection. On the breast cancer test datasets, the validation measures suggested that the XPN approach is most successful in avoiding over- and under-correction. Most

importantly, the cross-platform classification results, where XPN achieved the smallest cross-validated error using a PAM-classifier, indicate that XPN removes systematic differences between the datasets while at the same time preserving biological information.

Importantly, the cross-study normalisation methods discussed above can be used to extract new information from already existing publicly available microarray data-bases and thus provide a significant added scientific value. Moreover, combining datasets across studies does not only enable a more robust statistical analysis but also a more reliable cross-validation of the results.

However, cross-platform integration methods are also limited by the fact that the genetic probes and their identifiers in different microarray datasets often differ significantly, e.g. the set of shared genes might be small or it might not be possible to map all probes onto unique standardised gene identifiers. If the main goal of a microarray study is to identify biomarkers, the small sample size problem can alternatively be alleviated by just comparing the feature selection results on different data sets and searching common genes among the top-ranked features. Only genes with high ranks on several independent data sets are likely to be good candidates for prognostic biomarkers.

In summary, although microarray technology is likely to become cheaper in the future and the average number of samples per study will increase, combining evidence from multiple data sets does not only alleviate the small sample size problem but is also an effective means to exploit the synergies of already existing datasets.

### 3.7 Data Integration 2: Integrating Cellular Pathway Data

Apart from the possibility to integrate data from similar microarray studies, a multitude of opportunities exist to combine other types of biological information with large-scale gene and protein expression data. One of the most frequently used external knowledge sources are functional annotations which map genes and proteins onto cellular pathways and processes, complexes, chromosomal regions, or other biologically meaningful definitions of sets of functionally related genes/proteins. In particular, the possibility to map the attributes in a large-scale dataset onto cellular pathway definitions can provide several benefits for biological data analysis, e.g.:

- The data can be interpreted on the level of pathway deregulations, providing a more general “bird’s eye view” of the biological activity in the samples and enabling the experimenter to identify systemic changes across the studied biological conditions.
- The robustness of statistical analyses can be increased significantly, because the information from multiple noisy measurements on single genes/proteins can be aggregated into more robust *pathway expression fingerprints*, analysing global changes in cellular pathways rather than only small changes in single genes/proteins. Moreover, the dimensionality of the data is reduced, when considering aggregated feature sets rather than single features individually, alleviating statistical problems associated with the “curse of dimensionality” (see section on feature selection).
- The pathway definitions contain biological information on functional associations and similarities between genes and proteins which are often not extractable from the microarray data alone. Combining expression data and annotation data can therefore provide new biological insights, revealing deregulations of gene/protein sets related to specific functional processes (e.g. inflammation processes)

under certain biological conditions (e.g. a cancer disease).

The first microarray analysis methods using cellular pathway mappings and other definitions of functionally similar gene/protein sets, analysed the enrichment of these sets in differentially expressed genes/proteins, and are known as *singular enrichment analysis* (SEA) methods [307] or *over-representation analysis* (ORA) techniques. These approaches first select a set of informative features by applying a feature selection method on the data and a user-defined significance score threshold (e.g.  $q\text{-value} < 0.05$ ), and then test the enrichment of the corresponding genes or proteins among the selected features using a statistical test, e.g. the Kolmogorov-Smirnov test (for a ranked list of features) or the one-sided Fisher exact test (for unordered lists of features). However, the final results depend on the quality of the selection and the choice of the significance threshold, and several genes with small expression value changes, which only reveal a significant deregulation pattern when being considered together with other functionally similar genes, might be neglected.

For this reason, the SEA methodology was followed by a new generic approach known as *gene set enrichment analysis* (GSEA) [308]. GSEA uses all features in microarray instead of applying a threshold-based pre-selection, i.e. no information is discarded and no arbitrary parameter selection influences the analysis. The enrichment of pre-defined gene/protein sets, obtained from databases like Gene Ontology (GO), KEGG, BioCarta, Reactome etc., in the microarray data can be scored using a multitude of parametric and non-parametric statistical tests, using the raw experimental data directly to compute significance scores. The methods in this category include non-parametric Kolmogorov-Smirnov-based approaches like GSEA, CapMap and GeneTrail [309], and parametric methods like PAGE [310], GAGE [311], FatiScan [312], ErmineJ [313], MEGO [314] and ADGO [315].

Although these GSEA methods tend to outperform classical SEA approaches in terms of sensitivity and coverage of the biological information in the data, they are still affected by various limitations. Capturing the multitude of functional roles of genes and proteins in gene/protein set definitions (i.e. using a membership-function with discrete binary values, “member” or “non-member”) is often only possible to a limited extent, especially when only considering non-overlapping datasets. The similarity between genes and proteins is often better expressed using continuous similarity scores, and a great variety of similarity measures have been proposed for this in the past. Therefore, more recently, new types of enrichment approaches have been developed, termed as *modular enrichment analysis* (MEA) [307], which try to capture more complex continuous similarity information and non-linear dependencies between genes and proteins stored in networks and graphs. These methods take into account the functional interrelations between genes and proteins [316, 317] or combine the information from multiple types of annotation data (GO terms, KEGG pathways, protein domains, etc.) [29]. Instead of using pre-defined gene/proteins sets, modules or clusters are identified in large-scale data sources, and the raw experimental data is used directly on-the-fly in the computation of significance scores, instead of first extracting subsets of interesting genes/proteins in a pre-processing step (avoiding performance bottlenecks, similar to the previous improvement of GSEA over classical SEA).

In summary, these enrichment analysis methods can provide the user with new information on which cellular pathways, processes and complexes are activated or de-activated under certain biological conditions, and are therefore of great practical use for the biological interpretation of the data. However, the full potential of integrating pathway information into the analysis of high-dimensional biological datasets has not been fully exploited. For example, recent studies have shown that by summarising the expression values in gene sets to robust *meta-gene* fingerprints representing entire pathways, powerful predictors for the supervised analysis

of the data can be obtained [318]. This pathway-based integrative classification of microarray data can help to improve the robustness of the analysis, but there are still limitations in terms of the interpretability, the predictive accuracy and the applicability of models across different array platforms. Therefore, as part of this thesis, a new pathway-based classification algorithm, TSPP, was developed to alleviate some of these problems [23] (see chapter 7 for a detailed description of this approach), as well as a new approach to extend cellular pathway definitions based on molecular interaction data (see the description of the PathExpand approach in chapter 6).

### 3.8 Data Integration 3: Integrating Molecular Interaction Data

The previous section has discussed methods to exploit functional annotation data for the analysis of large-scale transcriptomics and proteomics datasets. However, in spite of the fast growth of public functional annotation databases, the annotations for many genes and proteins are still missing or insufficient. As an alternative data source, a multitude of large-scale experimental datasets are freely available on the web, containing implicit functional information which is not covered in the annotation databases. Among these data sources, one of the most important types is molecular interaction data, including protein-protein and protein-DNA interactions, gene regulatory interactions and metabolic interactions, which are commonly represented as networks (with nodes corresponding to molecules and edges corresponding to interactions). Moreover, in addition to physical molecular interactions, networks of functionally similar genes or proteins can also be constructed from other data sources for the inference of functional associations, e.g. gene co-expression data and synthetic lethality experiments, among others (see approaches mentioned in the section “Protein interaction data pre-processing”).

Both for weighted and unweighted interaction networks, a multitude of analysis methods exist to exploit the information content for the biological interpretation of other experimental data mapped onto the network. These include:

1. approaches for the identification of dense network *modules*, *clusters* or *communities* [319–324],
2. methods for the topological analysis of networks [325–327],
3. supervised analysis approaches for using information from the network as predictors [279, 328, 329].

These methodologies will be discussed in more detail in the following paragraphs.

**1)** The first type of methods, detecting communities of densely interconnected nodes, exploit the knowledge that in biological networks molecules often tend to work together in modules, e.g. protein complexes represent functional units, in which single proteins perform lower level functions within the same cellular process. Algorithms for identifying corresponding communities include methods scoring the edge-connectivity of nodes within a putative community against the connectivity with the rest of the network [320] (similar to the “within-cluster sum-of-squares” and “between-cluster sum-of-squares” comparison often used in unsupervised clustering) or using other *modularity* scores to identify dense subgraphs. Often advanced search space exploration methods are used to maximise these scores [330], including fast greedy methods [331], random walk simulations [332], nature-inspired optimisation approaches [333, 334] and mathematical programming [335]. Other graph-theoretic approaches are the MCODE method [336], using vertex weighting by computing a measure of the local neighbourhood density, and the Markov Cluster

(MCL) algorithm [337], simulating random walks by deterministic mathematical operations on stochastic matrices to identify densely clustered groups of nodes.

Interestingly, the problem of finding dense communities of nodes in a network is closely related to the task of unsupervised clustering, and recent approaches therefore combine classical clustering techniques with the identification of network modules [338, 339]. This also means that similar strategies can be used to overcome common problems, e.g. the problem of overlapping clusters/communities can be addressed with fuzzy clustering/community detection algorithms [340]. More recently, in this context the clustering of network *edges* rather than nodes has been proposed to address the issue of overlapping clusters [341].

A frequently occurring problem in these types of analyses is the size heterogeneity of communities identified on many real networks [342] (with many very small communities, and few very large communities), and the difficult interpretation of these modules, if they are not enriched in certain functional annotations. To circumvent some of these limitations in network community identification, as part of this thesis a new method to find functional gene set associations using a molecular interaction network was developed, which exploits network distance information rather than clustering nodes or edges (see the description of the EnrichNet software [21] in chapter 6).

2) A fundamentally different approach for identifying outstanding and potentially biologically meaningful features in networks is obtained by investigating their topological structures. Both global and local topological properties of a network can be used for a wide range of analysis purposes. *Global* analysis techniques often exploit the knowledge that many types of real-world biological networks display a scale-free topology [122], with a degree distribution following a power-law. This property can be used in the analysis of high-dimensional experimental data, e.g. for the creation of better gene co-expression networks [343]. Moreover, a global topological analysis can reveal whether a network displays assortative mixing [344], i.e. whether nodes with the same annotation tend to be closer together in the network than nodes with different annotations, or whether the network has hierarchical structures like a hub-and-spoke topology, i.e. few high-degree nodes (regulatory genes or signalling proteins) interlink many small-degree nodes. *Local* topological properties can provide several useful details to characterise single genes/proteins and node communities, e.g. by computing different centrality measures (degree, eigenvector centrality, betweenness, closeness), measures of the nodes' tendency to form clusters (clustering coefficient, transitivity), and the distances between them (shortest path length, random walk distance and kernel distances). For details on these descriptors, see [345] and chapter 6 in this thesis.

All these types of topological information are typically not considered in a classical gene set analysis and in microarray feature selection. However, an informative post-filtering of differentially expressed genes from a microarray study could be obtained by identifying genes with outstanding topological properties (see section "Integrating Cellular Pathway Data" above). For this purpose, a new method for the network-topological analysis and comparison of gene sets was developed during this doctoral project (TopoGSA [20], see chapter 6).

3) A further alternative approach to integrate network information into microarray analysis is to use network properties for the selection or definition of features in supervised sample classification. For example, instead of using single genes/proteins or pre-defined gene/protein sets as predictors, differentially expressed sub-networks can be extracted for sample classification, and have been shown to achieve higher robustness in cross-study classification than conventional approaches [328, 329]. Apart from the use of network-based

predictors, the edge information in a network also provides a means to account for corresponding dependency structures in the data, e.g. pairs of neighbouring genes can be considered as partly redundant, correlated, or functionally related features, and this information can be incorporated into the penalty terms in regularised classifiers [279]. However, in terms of the overall accuracy, these methods often do not always provide the high sensitivity and specificity required for clinical diagnosis, and should therefore not be used as the only source of information but combined with other clinical and experimental data.

Overall, when comparing molecular interaction networks derived from experimental data with cellular pathway definitions (see previous section) as a potential source of additional information for microarray data analysis, the interaction networks tend to provide a more holistic view of the complexity and multitude of interactions in living cells than the pathway definitions, which simplify the representation of processes for human interpretation by modelling different processes as independent modules. At the same time, this comparison also highlights the main drawback of interaction networks, consisting in the limited interpretability of the often very complex structures found in networks using graph clustering and community detection algorithms. The resulting gene/protein sets do not always have homogeneous functional annotations, their size can be too small or large to be used as gene set predictors in sample classification, and the problem of missing reliable “gold standard” reference partitions for validation purposes (see section on unsupervised class discovery above) raises questions regarding the reliability and significance of some of these network clustering results.

Thus, in order to obtain a better trade-off between the high interpretability of cellular pathway based gene set definitions and the higher precision and information coverage in molecular interaction networks, a method to redefine pathways using interaction data was developed as part of this doctoral project and tested on experimental data (see the section on the PathExpand software [22] in chapter 6).

In summary, a wide range of opportunities exist to extend statistical analysis and machine learning based analysis of functional genomics data by including external biological data within new integrative analysis techniques. Although several algorithms have already been developed for this purpose in the past, there is still much room for improvement in terms of robustness, interpretability and accuracy. In the following chapters, several classical and integrative analysis approaches will be compared using the framework developed in this doctoral project, and new integrative approaches in the framework will be introduced and discussed in detail.

## Chapter 4

# Comparison of Standard Machine Learning Techniques and Integrative Extensions

### Chapter abstract

To complement the overview of different feature selection, prediction and clustering techniques for high-dimensional data analysis presented in the literature review in chapter 3 with more details on the performance of different types of approaches, this chapter provides a comparative evaluation of machine learning methods on microarray gene expression data.

The first section compares a representative choice of attribute selection methods (including a univariate filter, a combinatorial filter, an embedded selection method and a filter/wrapper combination) and classification approaches (a kernel-based SVM approach, a tree-based random forest classifier, the nearest-centroid based PAM classifier and the in-house rule-based classifier BioHEL). The methods are evaluated on real-world microarray cancer datasets, including two benchmark datasets from the literature (prostate cancer and diffuse large B-cell lymphoma) and a breast cancer dataset from the co-operating Queen's Medical Centre in Nottingham using a two-level external cross-validation procedure [25]. The results show that rule-based methods with high model interpretability can achieve similar average accuracies as complex kernel-based classification approaches.

The second part of the chapter will compare and evaluate clustering methods (partition-based and hierarchical approaches, as well as a consensus clustering method) using multiple internal and external validity methods. Moreover, as a new integrative approach, a combination of gene set analysis and clustering methods will be presented, demonstrating that on average higher adjusted rand indices with known outcome labels are obtained when interlinking gene set analysis based dimensionality reduction and consensus clustering.

Importantly, this chapter uses material from a recently submitted paper and the compared methods are all included in the integrative framework for high-dimensional data analysis implemented in this doctoral project. They can be accessed online on the ArrayMining.net web-application [18].



### 4.0.1 Supervised analysis - experimental protocol

The analysis pipeline to compare both feature selection and prediction methods for microarray sample classification consists of three basic steps: Data pre-processing, supervised analysis and statistical post-analysis of the results. An illustration of the whole experimental procedure is shown in figure 4.1.

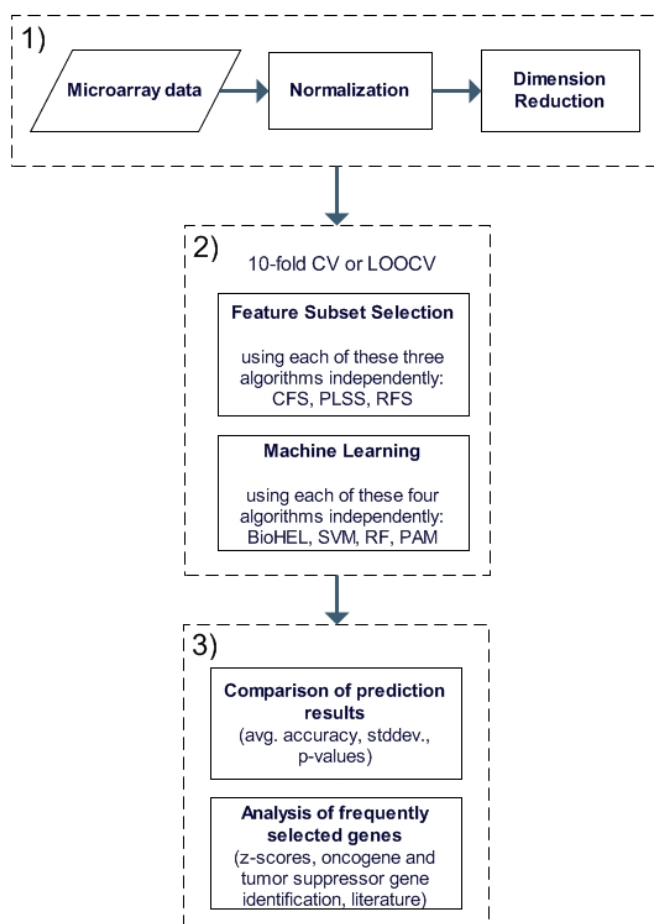


Figure 4.1: Flowchart illustrating the experimental procedure for evaluating feature selection and classification methods, consisting of three stages: 1) Pre-processing; 2) Supervised analysis; 3) Post-analysis. The wrapper-based SVM-RFE approach is not included in the flowchart, because feature subset selection and machine learning are linked together in this method.

In the first stage, the microarray datasets are normalised and an initial dimensionality reduction is performed (see section “Datasets and pre-processing” below). Next, an external cross-validation scheme is applied, i.e. in each cycle of the cross-validation, first a feature selection method is applied on the current training data and then a machine learning method on the resulting subset of features. This procedure is employed using both 10-fold external cross-validation (CV) and leave-one-out CV (LOOCV) [25] and all possible combinations of feature selection algorithms and classification algorithms. Specifically, the feature selection methods include the univariate filter “Partial-Least-Squares based Feature Selection” (PLSS), the combinatorial filter “Correlation-based Feature Selection” (CFS) [160], the embedded feature selection method “Random Forest based Feature Selection” (RFS), and a filter-wrapper combination using a PLS-based pre-filter and an SVM-wrapper with recursive-feature elimination (RFE). Similarly, the four chosen machine learning methods cover a wide range of diverse approaches: An in-house rule-based classifier, BioHEL, a

support vector machine (SVM) [346], a Random Forest classifier (RF) [191] and the “Prediction Analysis of Microarrays” method (PAM) [239].

In the last step of the protocol, a post-analysis is applied on the most informative genes to identify putative oncogenes and tumour suppressors, investigating the frequency of occurrence of gene identifiers in different types of prediction rules. Finally, an example literature mining is presented, showing that almost all of the top-ranked genes have known functional associations with the studied cancer diseases.

## 4.0.2 Datasets

All methods are evaluated on three public microarray cancer datasets representing three different types of cancer: Diffuse large B-cell lymphoma (DLBCL) [347], prostate cancer [348] and a breast cancer dataset obtained from the collaborating Queens Medical Centre in Nottingham [19, 349–351]. Below, details are given for each dataset and pre-processing method used in this comparative evaluation.

Table 4.1: Datasets used for comparative evaluation

Dataset	Platform	No. of genes	No. of samples class 1; class 2	references
DLBCL	Affymetrix	7,129	58 (D) ; 19 (F)	[347]
Prostate	Affymetrix	12,600	52 (T) ; 50 (N)	[348]
Breast	Illumina	47,293	84 (L) ; 44 (N)	[19, 349–351]

### Diffuse large B-cell lymphoma (DLBCL)

The DLBCL dataset [347] contains expression values for 7,129 genes and 77 microarray samples, 58 of which were obtained from patients suffering from diffuse large B-cell lymphoma (D), while the remaining samples are derived from a related B-cell lymphoma type, termed follicular lymphoma (F). The experiments were carried out on an Affymetrix HU6800 oligonucleotide platform [352].

To pre-process the raw data, the *Variance stabilizing normalisation* method [80] was applied to filter out intensity-dependent variance. This was done using the *vsr* library and the *expresso* package in the R statistical learning environment [353]). Moreover, a thresholding was applied using the suggestions in the supplementary material of the original publication associated with the dataset [347], and a “fold change”-filter used to remove features with low variance (all gene vectors with less than a 3-fold change between the maximum and minimum expression value were discarded), resulting in 2647 remaining genes (see section 4.2.3 for comparative classification results from the literature on this dataset). Figure 4.2 shows a 3D scatter plot of the first 3 principal components in the data.

### Prostate cancer

The prostate cancer dataset [348] consists of expression measurements for 12,600 genetic probes across 50 normal tissues and 52 prostate cancer tissues. All experiments have been carried out on Affymetrix Hum95Av2 arrays [352]. Due to the large number of samples, the fast GeneChip RMA (GCRMA) normalisation algorithm was applied [354], a method that combines stochastic and sequence-based physical models to estimate the mRNA abundances. Moreover, thresholding was employed using the suggestions of the original publication associated with the dataset [348] and a fold change filter to remove all probes with less than a 2-fold change between the maximum and minimum expression value, providing 2135 remaining

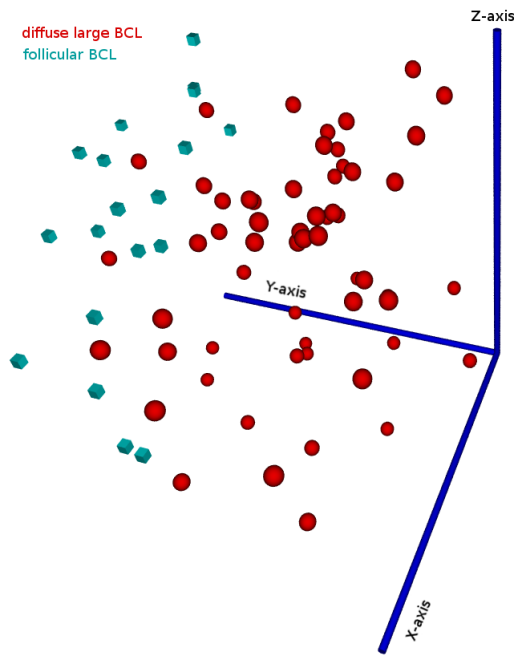


Figure 4.2: 3D visualisation of the first three principal components of the B-cell lymphoma dataset (created using the VRMLGen software [24], see chapter 8).

genes (see section 4.2.3 for comparative classification results from the literature on this dataset). Figure 4.3 shows a 3D scatter plot of the first 3 principal components in the data.

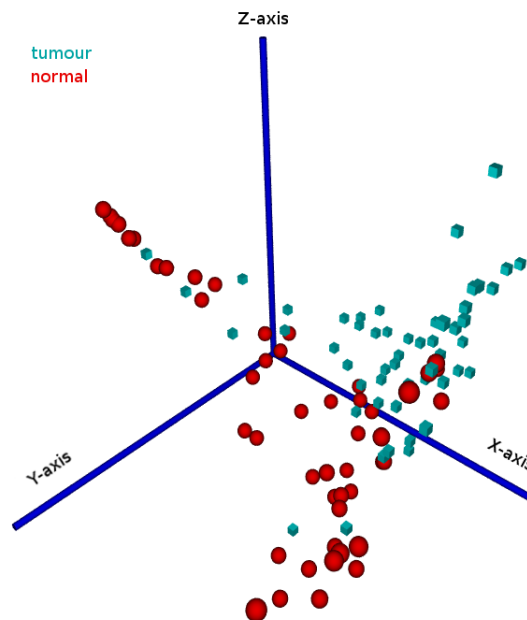


Figure 4.3: 3D visualisation of the first three principal components of the prostate cancer dataset (created using the VRMLGen software [24], see chapter 8).

**Breast cancer**

The breast cancer dataset from the collaborating Queen’s Medical Centre [19, 349–351] provides gene

expression values for 128 primary breast tumours across 47,293 genetic probes. Two groups of tumour samples can be distinguished in the data, the luminal group (L, 84 samples), which is characterised by oestrogen receptor expression, and the non-luminal group (N, 44 samples, no oestrogen receptor expression). The expression profiling procedure has previously been described in detail [349–351], and has also been applied in a recent ensemble gene selection analysis of this dataset [19]. Since the probe level data was not obtained from a conventional DNA chip, but from a Sentrix Human-6 BeadChip platform (v1.0, Illumina, San Diego, CA), the data was normalised and summarised using the dedicated Bioconductor “beadarray” package (see section 4.2.3 for comparative classification results). Figure 4.4 shows a 3D scatter plot of the first 3 principal components in the data.

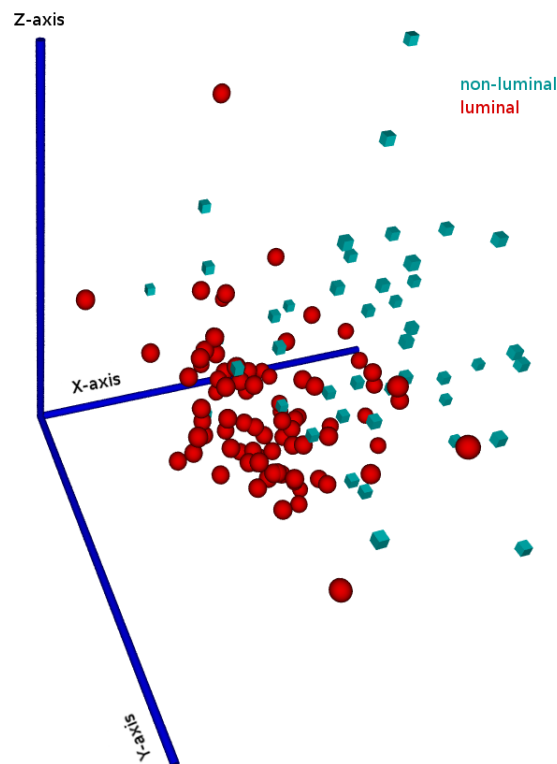


Figure 4.4: 3D visualisation of the first three principal components of the breast cancer dataset (created using the VRMLGen software [24], see chapter 8).

## 4.1 Comparative Evaluation of Feature Selection Methods

### 4.1.1 Feature selection methods

As discussed in the literature review (see chapter 3), microarray data typically contain a large number of uninformative genes with regard to the biological question to be addressed, causing statistical problems subsumed under the term “curse of dimensionality” (see [355]). Therefore, after normalisation and pre-filtering of the datasets in the first part of the analysis pipeline, microarray feature selection methods are applied prior to the classification methods (unless classification and attribute selection are already tied together in one algorithm), even if the user is only interested in the prediction results and not in the interpretation of the gene selection outcome.

To account for the diversity of existing feature selection methods, four different types of approaches are considered separately in this study. These include a classical univariate filter (PLSS [356]), a combinatorial filter (CFS [160]), a tree-based feature ranking approach (RFS [190]) and a filter/wrapper combination (PLS-filter in combination with an SVM-RFE wrapper). For all feature selection methods the maximum feature subset size was set to 30 to prevent overfitting, reduce the probability of including uninformative features and facilitate the comparison and interpretation of the results from different algorithms (however, the methods are allowed to flexibly select less than 30 features). This decision was also motivated by recent studies estimating the approximate number of features to be selected in different types of microarray studies to obtain only genes with significant informative value on the outcome attribute (based on different models to compute p-values for the significance of genes, see [357–359]). The chosen selection methods are described in detail in the following paragraphs.

**Partial-Least-Squares based Feature Selection (PLSS):** As a representative of a classical univariate filter, a method using the Partial Least Squares (PLS) [360] algorithm is employed. Specifically, the features are ordered by the absolute values of the weight vector defining the first latent component in a PLS model that was built upon the training data. As previously shown [361], the ordering of features obtained from this approach is equivalent to the F-statistic used in analysis of variance (ANOVA). Thus, instead of the PLS-calculation the F-statistic itself could have been used, but PLSS provides a much faster way of performing the calculation (the fast SIMPLS algorithm [362] is used for this purpose).

**Correlation based Feature Selection (CFS):** The CFS algorithm [160] searches for subsets of features that have high correlation to the response variable but low correlation amongst each other (see also the section on combinatorial filters in chapter 3). This concept is formalized by the following feature subset score:

$$CFS_S = \frac{k \cdot \overline{c_{rf}}}{\sqrt{k + k(k-1)\overline{c_{ff}}}} \quad (4.1)$$

where  $S$  is the selected subset with  $k$  features,  $\overline{c_{rf}}$  is the average feature-class correlation and  $\overline{c_{ff}}$  the average feature-feature correlation. While the denominator reduces the score for correlated features to eliminate redundant variables, the numerator promotes features with high correlation to the class variable to retain them as powerful discriminators. As proposed in the original CFS publication, a greedy best-first search strategy was employed to explore the feature subset space [160].

**Random Forest based Feature Selection (RFS):** In contrast to the CFS and the PLSS algorithm, the attribute selection obtained from the Random Forest classifier [190] uses a method directly embedded into the prediction algorithm. Specifically, a Random Forest model is built by training many binary, unpruned decision trees on bootstrap sub-samples of the training data. The importance of a feature is evaluated using the *Gini index* node impurity measure [192], by calculating the mean decrease in this measure from parent nodes to their direct descendent nodes over all tree nodes. A feature subset is obtained from the corresponding attribute ranking by selecting the top  $n$  features.

**Filter/wrapper combination** In order to assess the benefits of the wrapper methodology for feature selection, while retaining relatively short runtimes, a combination of a univariate filter with a wrapper, an approach that is frequently used in practice, was employed. Specifically, a PLS pre-filtering is used (see

PLSS method above) to pre-select 300 genes, among which highly predictive subsets are searched using an SVM-wrapper with a recursive feature elimination (RFE) search procedure [189] (see also the feature selection section in chapter 3). Employing an SVM-RFE wrapper selection without any pre-filtering would result in infeasible runtimes for typical microarray data and in an unfair comparison with the other selection methods, which require at most a few minutes computation time on standard desktop machines, hence, a “wrapper-only” approach was not considered here.

### Feature selection results

When using feature selection to pre-process microarray data prior to supervised classification, the average accuracy can vary greatly with the choice of the selection method, since the performance does not only depend on the inclusion of informative features, but can also be affected negatively by the selection of redundant and irrelevant features. To compare the selection methods considered in this study, the Friedman test was applied to the average classification accuracies (once for 10-fold CV and once for LOOCV) across all datasets and all four prediction methods (BioHEL, SVM, RF and PAM; with the exception of the filter/wrapper approach, in which the selection is tied to a single algorithm).

In summary, for 10-fold CV, no significant differences in performance were observed between the selection methods, but for LOOCV the PLSS approach was significantly superior to all other methods at 95% confidence level according to a subsequent Holm-test (see the ranking in table 4.2). The observation that a univariate ranking method could not be outperformed by combinatorial, embedded and filter/wrapper selection methods justifies the still widespread popularity of univariate selection methods. Relatively high performances of simple selection strategies had already been noted in a similar study by Wessels *et al.* [363] when comparing other selection methods on microarray data. In particular, the F-statistic has been used frequently in highly successful machine learning systems. For example, in the “NIPS 2003 Feature Selection Challenge” [364], a method using the F-statistic and an SVM-classifier [365] was ranked among the top 5 entries, performing better than many multivariate selection strategies. Thus, if the independence assumption represents a good approximation for some of the most informative features, or if multivariate methods fail to correctly capture the dependence structure between different variables, a classical fast univariate selection approach may still be the method of choice for complex, high-dimensional microarray data.

The wrapper/filter combination (SVM-RFE) provided the best average accuracy on the DLBCL dataset, but on both other datasets the average performance was lower than for most other method combinations. However, in order to obtain a fair comparison between SVM-RFE and filter approaches, the runtimes were kept approximately similar, hence, the wrapper approach might achieve better results, by running it for a longer time or using a more stringent pre-filtering (notably, the DLBCL dataset, on which SVM-RFE performed best, has a smaller dimensionality than the other datasets). These observations match to previous findings in the literature according to which filter methods tend to provide a better trade-off between predictive *effectiveness* and runtime *efficiency* in comparison to wrappers on most real-world datasets [178] (see also chapter 3, section 3.3.2).

Table 4.2: Comparison of feature selection methods

method	Average ranks		
	CFS	PLSS	RFS
10-fold	2.1	<b>1.9</b>	2.0
LOO	2.4	<b>1.6</b>	2.0

Average rank scores resulting from a Friedman test to compare feature selection methods in terms of classification accuracy across different datasets and prediction methods. The best average ranks for each row are shown in bold typeface.

## 4.2 Comparative Evaluation of Classification Methods

### 4.2.1 Methods

The main purpose for the comparative evaluation of classification methods conducted in this study was to assess whether classification methods using easily interpretable “if-then-else” decision rules can reach similar accuracies as other state-of-the-art microarray sample classification methods, e.g. kernel-based approaches like SVMs. Therefore, the following section will first describe the rule-based in-house classification method BioHEL [366–369] and then the benchmark approaches (SVM, PAM, RF) employed for comparison, and the evaluation methods and implementation parameters will be discussed.

**Rule-based evolutionary machine learning (BioHEL)** BioHEL [366–369] is an evolutionary machine learning system employing the Iterative Rule Learning (IRL) paradigm [370, 371].

The IRL procedure begins with an empty rule set and the complete set of observations as input. Classification rules are added iteratively to the set of rules until their combination covers all samples. The final outputs are structured rule sets, also known as *decision lists* [372], a knowledge representation inherited from BioHEL’s predecessor software GAssist [373] (a small example rule set is shown in figure 4.5). Each time a new decision rule has been learnt and added to a corresponding rule set, the observations it covers are removed from the examples set.

To explore the search space of possible rules efficiently, BioHEL uses a standard generational Genetic Algorithm (GA) which is applied in each IRL iteration to find the best rule for samples which have not yet been covered by rules found in previous iterations. Since GAs are non-deterministic, multiple repetitions of the rule learning process with identical training sets can be used to increase the probability of finding the optimal rule. Additionally, repetitions of the complete learning process (i.e. generating a complete rule set and not just a single rule) can also be applied, in order to combine several rule sets to a majority-vote consensus prediction and benefit from the variance-reducing effects of ensemble learning [8] (see section “Ensemble learning methods” in chapter 3).

In order to find the best rule in each IRL iteration, the fitness function used in the GA accounts for both the accuracy and the generality, i.e. the number of covered observations, of a rule. In BioHEL, this fitness function accounts for the Minimum Description Length (MDL) principle [374] and has the following basic structure:

$$Fitness = TL \cdot W + EL \quad (4.2)$$

**Rule 1:**  
Att 40282\_s\_at ∈ [1.37,8.58]    ^    Att 38087\_s\_at ∈ [3.67,8.49]    ^  
Att 37366\_at ∈ [5.18,9.00]    ^    Att 38407\_r\_at ∈ [1.05,9.00]    ^  
Att 38322\_at ∈ [3.69,6.57]    ^    Att 38406\_f\_at ∈ [2.80,8.73]    ^  
Att 38291\_at ∈ [3.01,4.71]  
→ **class 1**

**Rule 2:**  
Att 32598\_at ∈ [3.15,9.00]    ^    Att 41468\_at ∈ [3.31,9.00]    ^  
Att 38087\_s\_at ∈ [3.67,8.49]    ^    Att 37366\_at ∈ [8.05,8.95]    ^  
Att 38406\_f\_at ∈ [2.79,9.00]    ^    Att 37394\_at ∈ [3.64,8.44]  
→ **class 1**

**Rule 3:**  
Att 38406\_f\_at ∈ [2.79,5.93]  
→ **class 1**

**Rule 4:**  
Att 41468\_at ∈ [3.31,9.00]    ^    Att 33767\_at ∈ [2.90,8.23]    ^  
Att 38322\_at ∈ [5.97,8.74]    ^    Att 39120\_at ∈ [2.39,6.81]  
→ **class 1**

**Rule 5:**  
Att 38291\_at is [3.10,8.78]  
→ **class 2**

Figure 4.5: Example for a BioHEL classification rule set for the prostate cancer dataset (“Att” is short for “Attribute”, “^” represents the conjunctive AND-operator, “[x,y]” is an interval of expression values in which the value of the attribute must lie to fulfil one premise of the rule, and “→” is a class assignment operator, followed by the output class of the rule).

where  $TL$  represents the theory length (reflecting complexity, see precise definition below),  $EL$  stands for exceptions length (reflecting accuracy and coverage, see precise definition below) and  $Fitness$  is a score to be minimised.  $W$  is a weighting factor to adjust the relation between  $TL$  and  $EL$ , which is set automatically using a previously introduced heuristic [373]. More specifically,  $TL$  is given by the following formula:

$$TL(R) = \sum_{i=1}^{NA} \frac{NumZeros(R_i)/Card_i}{NA} \in [0, 1] \quad (4.3)$$

where  $NA$  is the number of attributes of the domain,  $R$  is a rule,  $R_i$  is the predicate of rule  $R$  associated to attribute  $i$ ,  $NumZeros$  measures the rule specificity by counting the number of bits set to zero in the binary vector representation for the rule predicates (see [375] for details on this representation), and  $Card_i$  is the cardinality of attribute  $i$ .

The second scoring term,  $EL$ , is designed to maximise both the accuracy ( $acc$ ) of rules and their sample coverage ( $cov$ ). Rules which cover a certain minimum percentage of observations receive a high reward, but after surpassing this threshold, the additional reward for covering more samples is smaller:

$$EL(R) = 2 - acc(R) - cov(R) \quad (4.4)$$

$$acc(R) = \frac{corr(R)}{matched(R)} \quad (4.5)$$



$$cov = \begin{cases} minCovRatio \cdot \frac{rawCov}{covBreak} & \text{if } rawCov < covBreak \\ minCovRatio + (1 - minCovRatio) \cdot \frac{rawCov - covBreak}{1 - rawCov} & \text{if } rawCov \geq covBreak \end{cases} \quad (4.6)$$

$$rawCov = \frac{matched(R)}{|T|} \quad (4.7)$$

where  $corr(R)$  is the number of examples correctly classified by rule  $R$ ,  $matched(R)$  is the number of examples matched by  $R$ ,  $minCovRatio$  is a weighting factor applied depending on whether the minimum coverage is achieved,  $covBreak$  is the minimum coverage threshold and  $|T|$  is the total number of training examples.

In summary, BioHEL combines the benefits of evolutionary algorithms, IRL and ensemble learning to obtain rule sets for sample classification that are both informative and accurate.

**Benchmark machine learning methods for microarray sample classification** In order to compare BioHEL against commonly used methods for microarray sample classification, the cross-validation procedure was applied to three alternative classifiers in addition to BioHEL: A *support vector machine* (SVM) [346], a *random forest* classifier (RF) [191] and the *nearest shrunken centroid* classifier (“Prediction Analysis of Microarrays”, PAM) [239].

The used support vector machine is a linear kernel C-SVM from the e1071-package [376] of the R statistical learning environment [353], a wrapper for the well-known LibSVM library [377]. Other polynomial kernels and the radial basis function kernel were tested without providing superior results (data not shown). This observation matches well to earlier findings in the literature according to which linear kernel SVMs often perform similar or better on microarray data than SVMs using higher degree polynomial kernels [189, 378, 379]. For the RF and PAM methods, the corresponding R packages *randomForest* and *pamr* were used. Moreover, BioHEL was also compared with alternatives from the literature using previously published results. For this purpose, only average cross-validation accuracies were considered, since evaluation methods involving only a single random training/test-set partition are now widely regarded as unreliable [380]. For the same reason, methods from the literature using internal cross-validation instead of external cross-validation were excluded from the comparison, wherever this was clearly stated by the authors.

### 4.2.2 Evaluation methods and implementation parameters

The main evaluation method used in this study is a cross-validation scheme known as *two-level external cross-validation* [25]. In an *external* cross-validation, feature selection is applied independently to each training set generated within the cross-validation cycles, avoiding the selection bias of classical internal cross-validation, where feature selection is only applied once to the whole dataset prior to any further analysis [380]. *Two-level* external cross-validation uses an additional nested cross-validation procedure to optimise the parameters for each prediction algorithm. We apply this second level of cross-validation to fit the parameters for the benchmark predictors SVM, RF, and PAM.

Since BioHEL employs an evolutionary algorithm depending on the initialisation of a stochastic random number generator, the entire cross-validation process is repeated 10 times for different random number

seeds. The results are averaged over all repetitions and cross-validation cycles, e.g. for 10-fold cross-validation the final accuracy is an average over 100 runs of BioHEL. Other algorithm specific parameters used for the evolutionary learning procedure are shown in table 4.3.

Table 4.3: **Parameters used for BioHEL**

population size	100
iterations	100
cross-over probability	0.6
mutation probability	0.6
selection type	tournament selection
coverage ratio	0.9
coverage break	0.2

(for more details on BioHEL's workflow and default parameters see [381] and [369])

Moreover, apart from single rule-set classifiers, ensemble models are trained by assigning samples to the majority-voting class of rule-sets obtained from multiple runs of BioHEL. The number of base models in ensemble learning can be seen as a trade-off parameter determining the balance between the bias and variance of the model (see ensemble learning section in chapter 3). Importantly, using a high number of base models does not necessarily reduce model interpretability, e.g. more robust statistics on the occurrence of selected genes in different rule sets can be obtained with a larger number of base models (see section "Identification of potential oncogenes and tumour suppressor genes" below). Since the best results were obtained with a 100-times ensemble, this was used as the default configuration.

Importantly, the obtained prediction models are only applicable to samples from the same platform, cell type, environmental conditions and experimental procedure, since this study was designed for the comparative evaluation of different analysis techniques and not for cross-platform data integration. However, as BioHEL supports both continuous and discretized input data, it is compatible with most of the cross-study normalization methods that have been proposed in the literature.

### 4.2.3 Comparison of prediction results

An overview of the comparative prediction results is given in table 4.4 for 10-fold CV and table 4.5 for LOOCV. Below the results for both datasets are discussed.

#### **Classification results: Diffuse large B-cell lymphoma (DLBCL)**

On the DLBCL dataset, the best prediction results with BioHEL were reached in combination with the PLSS filter, providing average accuracies of 92% (10-fold CV) and 93% (LOOCV). These results are both comparable to those for parameter-optimised conventional methods considered in this study (SVM, RF and PAM, see tables 4.4 and 4.5) and results reported in the literature for this dataset (see table 4.6). In the comparative analysis, only some SVM models outperformed BioHEL by a slight margin (SVM-RFE and SVM in combination with PLSS-feature selection, reaching an average accuracy of 96%, respectively 94%, with LOOCV). From the results reported in the literature for the DLBCL dataset (see table 4.6) only the approach by Liu *et al.* [382] reached a slightly higher accuracy (94%) than the best BioHEL models.

Table 4.4: 10-fold CV classification results

Dataset	Feature Selection	Classification	AVG (%)	STDDEV
PROSTATE	CFS	BioHEL	92	8
	PLS	BioHEL	<b>92</b>	12
	RF	BioHEL	89	11
	CFS	SVM	90	10
	CFS	RF	92	11
	CFS	PAM	91	10
	PLS	SVM	90	11
	PLS	RF	92	9
	PLS	PAM	<b>94</b>	8
	RF	SVM	88	8
	RF	RF	93	9
	RF	PAM	90	11
	SMV-RFE	SVM	88	8
DLBCL	CFS	BioHEL	83	19
	PLSS	BioHEL	<b>92</b>	11
	RFS	BioHEL	87	17
	CFS	SVM	87	12
	CFS	RF	87	16
	CFS	PAM	78	17
	PLSS	SVM	91	13
	PLSS	RF	87	8
	PLSS	PAM	86	11
	RFS	SVM	91	13
	RFS	RF	89	13
	RFS	PAM	86	14
	SMV-RFE	SVM	<b>95</b>	9
BREAST	CFS	BioHEL	<b>86</b>	8
	PLSS	BioHEL	82	11
	RFS	BioHEL	83	10
	CFS	SVM	86	9
	CFS	RF	86	7
	CFS	PAM	<b>89</b>	7
	PLSS	SVM	84	7
	PLSS	RF	<b>89</b>	5
	PLSS	PAM	88	7
	RFS	SVM	80	17
	RFS	RF	<b>89</b>	5
	RFS	PAM	88	7
	SMV-RFE	SVM	86	9

10-fold cross-validation results obtained with BioHEL, SVM, RF and PAM on the three microarray datasets using four feature selection methods (CFS, PLSS, RFS, SMV-RFE); AVG = average accuracy, STDDEV = standard deviation. The highest accuracies achieved with BioHEL and the best alternative method are both shown in bold typeface for each dataset.

Table 4.5: LOOCV classification results

Dataset	Feature Selection	Classification	AVG (%)	STDDEV
PROSTATE	CFS	BioHEL	89	31
	PLS	BioHEL	<b>92</b>	28
	RF	BioHEL	91	29
	CFS	SVM	89	31
	CFS	RF	<b>95</b>	22
	CFS	PAM	90	30
	PLS	SVM	93	25
	PLS	RF	93	25
	PLS	PAM	93	25
	RF	SVM	89	31
	RF	RF	91	29
	RF	PAM	91	29
	SMV-RFE	SVM	85.3	36
DLBCL	CFS	BioHEL	82	39
	PLSS	BioHEL	<b>93</b>	26
	RFS	BioHEL	84	37
	CFS	SVM	88	32
	CFS	RF	87	34
	CFS	PAM	84	37
	PLSS	SVM	94	25
	PLSS	RF	90	31
	PLSS	PAM	86	35
	RFS	SVM	90	31
	RFS	RF	92	27
	RFS	PAM	83	38
	SMV-RFE	SVM	<b>96</b>	19
BREAST	CFS	BioHEL	<b>84</b>	36
	PLSS	BioHEL	<b>84</b>	36
	RFS	BioHEL	<b>84</b>	37
	CFS	SVM	84	37
	CFS	RF	84	36
	CFS	PAM	<b>90</b>	30
	PLSS	SVM	81	39
	PLSS	RF	88	33
	PLSS	PAM	86	35
	RFS	SVM	86	35
	RFS	RF	87	34
	RFS	PAM	88	32
	SMV-RFE	SVM	78.9	41

Leave-one out cross-validation results obtained with BioHEL, SVM, RF and PAM on the three microarray datasets using four feature selection methods (CFS, PLSS, RFS, SMV-RFE); AVG = average accuracy, STDDEV = standard deviation. The highest accuracies achieved with BioHEL and the best alternative are both shown in bold typeface for each dataset.

A common problem in the classification of high-dimensional data with small sample sizes is the high variance in cross-validation error estimates, especially in LOOCV [380, 383]. This observation is also made on the three datasets considered in this study and applies both to BioHEL and the alternative prediction methods. However, in comparison to random classification assignments, all BioHEL models showed significant discriminative power on the DLBCL dataset.

Table 4.6: Comparison of prediction results from the literature for the DLBCL dataset

Author (year)	Method	(Avg.) accuracy	(Avg.) number of genes
Wessels <i>et al.</i> [363]	RFLD(10), Monte-Carlo CV	95.7	80
Liu <i>et al.</i> [382]	MOEA+WV	93.5	6
Shipp <i>et al.</i> [347]	SNR+WV, LOOCV	92.2	30
Goh <i>et al.</i> [384]	PCC-SNR + ECF, LOOCV	91	10
Lecocke <i>et al.</i> [385]	GA+SVM, LOOCV	90.2	**
	GAGA+DLDA, LOOCV	89.8	**
	GAGA+3-NN, LOOCV	86.3	**
Hu <i>et al.</i> [386]	WWKNN, LOOCV	87.01	12
	ECF, LOOCV	85.71	12
our study	PLSS+BioHEL, LOOCV	89.68	*30
	PLSS+BioHEL, 10-fold CV	92.48	*30

\*maximum no. of genes per base classifier in ensemble learning model

\*\*evaluation results averaged over feature subsets using different numbers of genes

#### Classification results: Prostate cancer

On the prostate cancer data, the highest average accuracies of 92% (10-fold CV) and 92% (LOOCV) with BioHEL were again obtained using the PLSS feature selection. From the alternative prediction methods considered in the comparative analysis, only the PLS/PAM combination reached a slightly higher accuracy (10-fold avg. acc.: 94%. LOOCV avg. acc.: 93%). Similarly, in the results reported in the literature for this dataset using external cross-validation methods, only Shen *et al.* (2005) [387] and Paul *et al.* (2005) [388] (see table 4.7) obtain a slightly higher average accuracy. However, Shen *et al.* employ a singular value decomposition (SVD) instead of feature selection, which includes more genes from the original data than the maximum of 30 considered here, and which can be more difficult to interpret (unless the derived features can be linked to biological processes). Paul *et al.* use original features in their models, but the average number of included genes also exceeds 30 features (48.5). Thus, considering both accuracy and model complexity, BioHEL performs well on this dataset in comparison to the benchmark classifiers and alternative approaches in the literature.

#### Classification results: Breast cancer

For the breast cancer dataset, obtained from the Nottingham Queen's Medical Centre, the best average accuracies obtained with BioHEL were 86% (10-fold CV) and 84% (LOOCV). These results were similar to those of other benchmark classifiers, with some methods being slightly superior and some slightly inferior (the most successful approach was CFS/PAM with 89% acc. for 10-fold CV and 90% acc. for LOOCV). Importantly, independent of the feature selection and cross-validation method, BioHEL always provided average accuracies of at least 82% on the breast cancer data.

Table 4.7: Comparison of prediction results from the literature for the Prostate dataset

Author (year)	Method	(Avg.) accuracy	(Avg.) number of genes
T.K. Paul <i>et al.</i> [388]	RPMBGA, LOOCV	96.6	48.5
Wessels <i>et al.</i> [363]	RFLD(0), Monte-Carlo CV	93.4	14
Shen <i>et al.</i> [387]	PLR, Monte-Carlo-CV (30 iterations)	94.6	***
	PLR, Monte-Carlo-CV (30 iterations)	94.3	***
W Chu <i>et al.</i> [389]	Gaussian processes, LOOCV	91.2	13
Lecocke <i>et al.</i> [385]	SVM, LOOCV	90.1	**
	DLDA, LOOCV	89.2	**
	GAGA+3NN, LOOCV	88.1	**
	our study	PLSS+BioHEL, LOOCV	92.8
	PLSS+BioHEL, 10-fold CV	93.2	*30

\*maximum no. of genes per base classifier in ensemble learning model

\*\*evaluation results averaged over feature subsets using different numbers of genes

\*\*\*singular value decomposition used instead of classical feature selection

Since this dataset was obtained from a collaborating institute, no external cross-validation results for alternative methods are available in the literature, however, the dataset has been published online and can freely be used for comparative evaluation and analysis purposes [19, 349–351].

On the whole, in all experiments the BioHEL ensemble classification models provided high and robust classification accuracies, comparable to those for a selection of some of the most popular microarray classification methods. The good agreement between the performance estimates obtained from 10-fold CV and LOOCV provides further support for these observations made for individual combinations of datasets and cross-validation methods. To compare the predictors across all datasets and different feature selection methods, a Friedman test [390, 391] over the average classification accuracies (once for 10-fold CV and once for LOOCV) was applied in addition to the direct comparison of accuracies within each dataset (the wrapper-based approach, which is tied to a single classifier, was disregarded in this analysis). According to this test, at a 95% confidence level no significant differences between the performances of the different classification methods were detected (see the average ranks in table 4.8). These results suggest that in spite of using simple “if-then-else”-rules, BioHEL’s performance for microarray sample classification is approximately comparable to that of parameter-optimised SVM-, RF- and PAM-models.

With regard to BioHEL’s runtime, even the most time-consuming experiment (combining LOOCV with the 100-times ensemble and repeating this 10 times for different random seeds) required less than one day on a 2 GHz dual-core CPU (other experiments lasted between several minutes and a few hours, depending on the number of CV-cycles and base models in ensemble learning). Across all methods, the overall memory requirements and the runtimes for applying the trained models were similar and negligibly small.

Table 4.8: Comparison of prediction methods - Friedman test

method	Average ranks			
	SVM	RF	PAM	BioHEL
10-fold	2.9	<b>1.8</b>	2.6	2.7
LOO	2.6	<b>1.9</b>	2.3	3.2

Average rank scores resulting from a Friedman test to compare prediction methods across different datasets and feature selection methods. The best average ranks are shown in bold typeface.

### Identification of potential oncogenes and tumour suppressor genes

Analysing genes solely based on their differential expression patterns in samples from different disease conditions can only provide a rough indication of their potential functional roles in the disease. Therefore, using the rule sets obtained from the BioHEL classification models, occurrence patterns of frequently selected genes in different rule types were investigated to obtain further insights.

For this purpose, top-ranked genes for each dataset, which had been chosen most frequently across all selection methods, were studied for their occurrence in the rule predicates of the BioHEL ensemble models. The example rule set in figure 4.5 contains some of these top-ranked gene attributes (the attribute names are specified as Affymetrix gene identifiers). This example also shows that each occurrence of a gene in a rule has an associated value range, in which the expression value of the gene must lie to fulfil one premise of the rule. Moreover, each rule has an associated class label corresponding to the conclusion of the rule.

Thus, for pairs of selected genes and possible class assignments, sets of upper and lower expression value bounds can be extracted from the rules and analysed statistically. In some cases, more general rules can be derived from these extracted value ranges, i.e. “less than”- or “greater than”-rules according to which the expression value of a gene must always be greater or smaller than a certain threshold-value, when the sample is assigned to a specific class. In a previous machine learning approach [392], this idea had already been used to identify potential *promoter*- and *blocker*-genes, whose gene products are likely to promote the disease (potential oncogenes, which are mostly over-expressed in the tumour) or block the disease (potential tumour suppressor genes, which are mostly under-expressed in the tumour).

Here, an alternative approach is used to extract information from the rule-based models, providing an easily interpretable bar plot visualisation. For each pair of a selected attribute and a class assignment, the median of the upper and lower bound of the expression value range is extracted from all occurrences of this gene/class-pair in conditionals of the rule-sets in the 100-times ensemble model. Thus, for both the “tumour”- or the “healthy”-class a robust estimate is obtained for the value ranges in which the expression value of a gene should lie according to the ensemble model. Every gene-class pair is represented by a single vertical bar in the bar plot, extending from the median of the lower bounds to the median of the upper bounds.

In contrast to a binary categorisation of genes into *promoters* and *blockers* of a disease [392], this visualisation of median value ranges for the sample classes using coloured bars does not only contain information on whether a gene tends to be up- or down-regulated in a certain sample class, but the distance between the median lower and upper bound also provides information on the narrowness of the value range for a certain gene/class-pair and the size of the overlap for the value ranges of different classes. Thus, this analysis also provides an intuitive measure of confidence (see example plots in figures 4.6 and 4.7).

More importantly, the plots in figure 4.6 and 4.7 reveal that in the corresponding datasets the differences between the expression values across the sample classes are not large enough in relation to the within-class variance for any of the top-ranked individual genes to obtain reliable classification results based on a single gene as predictor (the expression value ranges for the assignment of a sample to the different classes overlap for each of these genes). This observation explains why groups of genes are required to build models with substantial predictive power for these datasets.

In summary, analysing the occurrence of gene attributes in the ensemble rule sets enables the identification of putative oncogenes and tumour suppressor genes, whose expression values tend to be associated with a certain disease state. Based on the difference between the upper and lower bounds of the value ranges

in both classes and the overlap of these value ranges, genes can also be prioritized as potential diagnostic markers.

### Example literature mining for frequently selected genes

To illustrate the usefulness of the analysis pipeline for the biological interpretation of the data, an example literature mining was performed for the top-ranked genes on the prostate cancer dataset. Table 4.9 shows the top 20 genes that were chosen by at least two different selection methods among the genes selected most frequently across the LOOCV cycles (Tables 4.10 and 4.11 show the corresponding results for the DLBCL and the breast cancer dataset). Since this approach combines results from all cross-validation cycles and diverse selection methods, the identified consensus gene list is expected to represent a more robust selection than conventional rankings obtained when applying only a single selection strategy once on the whole data.

Table 4.9: List of high scoring genes for the Prostate cancer dataset

Gene identifier	No. of occurrences	Annotation
37639_at	3	<i>hepsin (transmembrane protease, serine 1)</i>
32598_at	3	<i>nel-like 2</i>
41706_at	3	<i>alpha-methylacyl-coa racemase (AMACR)</i>
38634_at	3	<i>retinol binding protein 1, cellular (CRBP1)</i>
37366_at	3	<i>pdz and lim domain 5 (PDLIM5)</i>
40282_s_at	2	<i>complement factor d (adipsin)</i>
38087_s_at	2	<i>s100 calcium binding protein a4 (S100A4)</i>
41468_at	2	<i>T cell receptor gamma (TCR-gamma)</i>
38827_at	2	<i>anterior gradient 2 (AGR2)</i>
38406_f_at	2	<i>prostaglandin d2 synthase 21kda (PTGDS)</i>
34840_at	2	<i>we38g03.x1 homo sapiens cdna, 3' end</i>

List of genes that were chosen by at least two different selection methods among the 20 features with highest Z-scores on the Prostate dataset (column 1: the Affymetrix gene identifier, column 2: number of feature selection methods for which the gene appeared among the 20 top-ranked genes, column 3: gene annotation)

Table 4.10: List of high scoring genes for the DLBCL dataset

Gene identifier	No. of occurrences	Annotation
X02152_at	3	<i>lactate dehydrogenase a (LDHA)</i>
V00594_at	2	<i>metallothionein 2a (MT2A)</i>
HG1980-HT2023_at	2	<i>tubulin, beta 2c (TUBB2C)</i>
U63743_at	2	<i>kinesin family member 2c (KIF2C)</i>
X05360_at	2	<i>cell division cycle 2, g1 to s and g2 to m (CDC2)</i>
M63379_at	2	<i>clusterin</i>
M13792_at	2	<i>adenosine deaminase (ADA)</i>
L19686_rna1_at	2	<i>macrophage migration inhibitory factor (MIF)</i>
D14662_at	2	<i>peroxiredoxin 6 (PRDX6)</i>
S73591_at	2	<i>thioredoxin interacting protein (TXNIP)</i>

List of genes that were chosen by at least two different selection methods among the 30 features with highest Z-scores on the DLBCL dataset (column 1: the Affymetrix gene identifier, column 2: number of feature selection methods for which the gene appeared among the 30 top-ranked genes, column 3: gene annotation)





Figure 4.6: Medians of lower and upper bounds of rule-based assignment intervals in both sample classes for the Prostate cancer dataset (red area = expression values between median lower and upper bound in tumour class, green = expression values between median lower and upper bound in normal class). Known and putative tumour suppressor genes often have a lower and narrower value range in the “tumour”-class and a higher and wider value range in the “normal”-class, whereas the opposite pattern is observed for oncogenes.

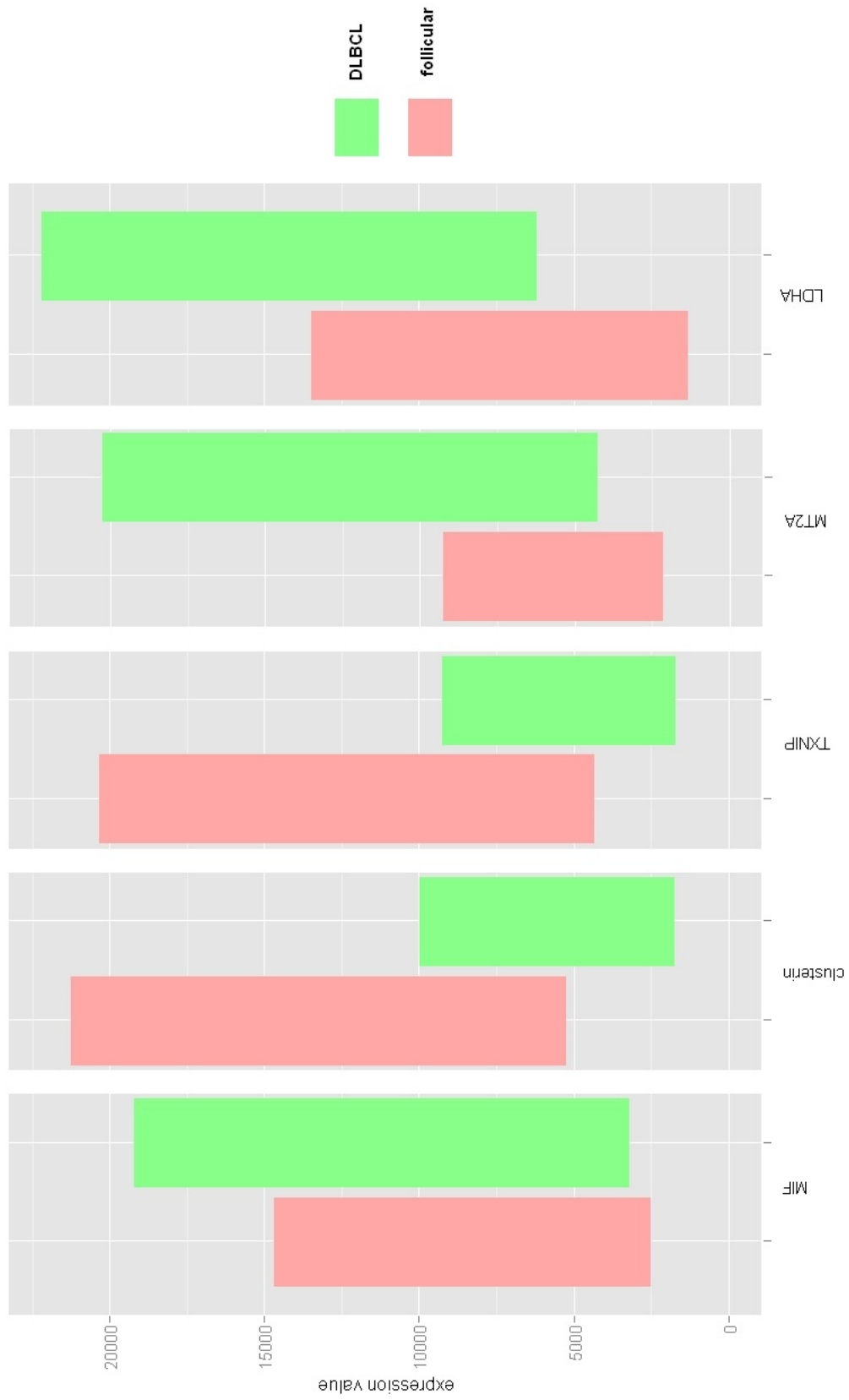


Figure 4.7: Medians of lower and upper bounds of rule-based assignment intervals in both sample classes for the DLBCL cancer dataset (red area = expression values between median lower and upper bound in tumour class, green = expression values between median lower and upper bound in normal class).

Table 4.11: List of high scoring genes for the Breast cancer dataset

Gene identifier	No. of occurrences	Annotation
GI.4503602-S	3	<i>Homo sapiens estrogen receptor 1 (ESR1)</i>
GI.14249703-S	3	<i>Homo sapiens RAS-like, estrogen-regulated, growth-inhibitor (RERG)</i>
GI.16507967-S	3	<i>Homo sapiens potassium channel, subfamily K, member 15 (KCNK15)</i>
GI.22779933-S	2	<i>Homo sapiens WD repeat membrane protein PWDMP (PWDMP)</i>
GI.42657473-S	2	<i>Uncharacterized protein (C6orf115)</i>
GI.7706686-S	2	<i>Homo sapiens Enah/Vasp-like (EVL)</i>
GI.40788002-S	2	<i>Homo sapiens proteasome (prosome, macropain) activator subunit 4 (PSME4)</i>
GI.33620752-S	2	<i>Homo sapiens hypothetical protein FLJ10876 (FLJ10876)</i>
GI.13236596-S	2	<i>Homo sapiens hypothetical protein MGC10765 (MGC10765)</i>
GI.29029609-A	2	<i>Homo sapiens pyrimidinergic receptor P2Y, G-protein coupled, 6 (P2RY6), variant 1</i>
GI.37551139-S	2	<i>Homo sapiens hypothetical protein PRO2013 (PRO2013)</i>
GI.40255152-S	2	<i>Homo sapiens potassium channel tetramerisation domain containing 6 (KCTD6)</i>
GI.30410031-S	2	<i>Homo sapiens prostate-specific membrane antigen-like protein (PSMAL/GCP III)</i>
GI.4503928-S	2	<i>Homo sapiens GATA binding protein 3 (GATA3), mRNA</i>
GI.42659459-S	2	<i>Homo sapiens hypothetical gene supported by AK128810 (LOC399717)</i>
GI.29738585-S	2	<i>Homo sapiens hypothetical protein LOC143381 (LOC143381)</i>
GI.38455428-S	2	<i>Homo sapiens breast cancer membrane protein 11 (BCMP11), mRNA</i>
GI.22035691-A	2	<i>Homo sapiens GDNF family receptor alpha 1 (GFRA1), transcript variant 2</i>

List of genes that were chosen by at least two different selection methods among the 30 features with highest Z-scores on the Breast cancer dataset (column 1: the Genbank identifier, column 2: number of feature selection methods for which the gene appeared among the 30 top-ranked genes, column 3: gene annotation)

In the prostate cancer dataset, five genes were found in the intersection set of the top 20 ranking lists for all feature selection methods: *Hepsin*, *nel-like 2*, *alpha-methylacyl-coa racemase*, *retinol binding protein 1* and *pdz and lim domain 5*. Annotations for these and all other genes on the list were obtained from the Gene Cards web-service [26], the DAVID functional annotation database [393] and from the supplementary material of the microarray dataset. Using this information, the biomedical literature in the PubMed database was mined to identify articles reporting functional associations of these genes with the disease under consideration.

The results reveal that almost all of the genes on the list have either known functional associations with cancer, have already been used as diagnostic markers or considered as candidates for new markers. Specifically, genes or associated genetic products on the list which are already used or have been proposed as tumour markers in the literature include the cell surface serine protease *Hepsin*, the *Neural epidermal growth factor-like 2 (nel-like 2)* gene, the enzyme *Alpha-methylacyl-CoA racemase (AMACR)*, the *T-cell receptor gamma (TCR- $\gamma$ )*, and the secretory protein *Anterior gradient homolog 2 (AGR2)*.

Other genes or corresponding proteins which could be linked to cancer-related processes like apoptosis (programmed cell-death) or cell cycle progression are the carrier protein *Cellular Retinol binding protein 1 (CRBP1)*, the enzyme *Prostaglandin d2 synthase 21kda (PTGDS)* and the *S100 calcium binding protein A4 (S100A4/Mts1)*. Similarly, some top-ranked genes are reported to have known or putative functional roles in cell growth and cell proliferation, e.g. the *PDZ and LIM domain 5 (PDLIM5)* gene and *adipsin*.

The genetic probe with identifier *34840\_at* corresponds to the only nucleotide sequence without additional annotation information occurring in the list of high-scoring attributes. However, the same genetic probe also occurred among the top-ranked genes obtained by another research group using other microarray analysis methods [394], suggesting that the corresponding gene could be a promising candidate for further investigations.

In summary, in the experiments conducted here, all except one of the genes appearing in multiple lists of top-ranked genes from different selection algorithms have either putative or known functional associations with cancer or related cellular processes. Although this does not imply that all high-scoring genes and their corresponding products are suitable markers for the diagnosis or monitoring of cancer diseases, it suggests the selection and ranking methods are useful in identifying and prioritizing putative markers.

While classical gene prioritization methods use a single feature selection method and a single confidence measure [395], the approach employed here uses information from multiple selection methods, multiple prediction models given by ensemble rule sets, and multiple sample subsets. Moreover, potential tumour suppressor genes and oncogenes can be identified by analysing the occurrence of gene attributes in different conditionals of the prediction rules. The corresponding post-processing procedure exploits the information content of a multitude of rule sets from an ensemble learning model and provides a bar plot visualisation to facilitate the interpretation and a confidence measure to prioritize genes.

Finally, since the lists of high-scoring candidate genes which were detected by multiple feature selection methods are confined to relatively small sets of attributes, it would be feasible to conduct a quantitative polymerase chain reaction (qPCR) to refine the gene prioritization using more accurate gene expression measurements than those provided by conventional microarray platforms.

#### 4.2.4 Supervised analysis - summary and conclusion

Overall, the empirical results on three public microarray datasets using three feature selection methods and two external cross-validation schemes show that a rule-based learning method, BioHEL, can reach classification accuracies comparable to current state-of-the-art prediction methods for gene array data. These results are corroborated by comparisons across multiple types of feature selection methods, as well as by comparisons to other methods in the literature.

As an added value, in contrast to other state-of-the-art benchmark methods, BioHEL's prediction models use easily interpretable conjunctive *if-then-else*-rules. Genes which are frequently selected as informative features in rule sets across different cross-validation cycles and different ensemble base classifiers provide robust and informative predictors with regard to the outcome attribute. In this context, using a high number of base models combined to an ensemble can even be beneficial for data interpretation due to the variance-reducing effects of ensemble learning [8] which result in more robust statistics on the importance of single features in the predicates of the decision rules. This matches well with the results of the example literature analysis for the prostate cancer data, showing that all except one of the top-ranked genes have a known or putative functional association to the studied cancer disease.

More importantly, biological insights can be gained from ensemble rule sets that are not obtainable by most other popular machine learning algorithms such as SVM and PAM. Potential oncogenes and tumour suppressor genes can be identified within the set of selected genes by summarizing and visualising information from the rule sets using easily interpretable bar plots (e.g. the *nel-like 2* gene, which is predicted as a potential tumour suppressor gene by the method proposed here, is already used as a marker in a patented diagnostic method for prostate cancer [396]). Thus, simple statistical analyses which can easily be applied to ensemble rule learning models, can help to accelerate the identification of candidate biomarkers for clinical diagnosis.

Although the main goal behind the analysis protocol was to show that an evolutionary machine learning method using simple decision rules can compete against other benchmark microarray classifiers across a

diverse group of feature selection approaches and multiple datasets, as a by-product of the experiments, the performance of different types of attribute selection methods were also compared. The PLSS approach provided significantly better results in LOOCV, while no method significantly outperformed the other approaches in the 10-fold CV experiments. These results justify the widespread popularity of fast filter-based feature selection methods in terms of providing a good balance between runtime efficiency and predictive effectiveness, while wrapper-based selection methods (SVM-RFE) tend to achieve higher accuracies only with significantly higher runtimes.

Possible future extensions for the machine learning system BioHEL include integrating prior clinical or biological knowledge into the analysis and directly combining the system with automated literature mining tools to better exploit the information content of the generated models. On the whole, BioHEL's performance in comparison to other successful predictors and the benefits in terms of interpretability show that rule-based evolutionary machine learning algorithms can be profitably applied for microarray sample classification.

### 4.3 Comparative Evaluation of Clustering Methods

In unsupervised analysis of microarray data, the comparison of diverse clustering approaches and the combination of multiple methods into a consensus can provide similar benefits with regard to algorithmic performance and biological data interpretation as corresponding techniques for supervised analysis (see above). For this purpose, this section will discuss the results of comparing *partition-based clustering methods*, including k-Means, PAM, CLARA and SOM, and *hierarchical clustering methods*, including Average linkage agglomerative clustering, DIANA, hybrid hierarchical clustering and SOTA (as a hybrid between SOM and hierarchical clustering), as well as the combination of these algorithms into a consensus clustering using a Simulated Annealing based aggregation method. Moreover, different standardisation and dimensionality reduction methods will be evaluated, including a gene set analysis based data transformation, which improves the robustness of the analysis, while retaining interpretable attributes representing cellular pathways and processes.

The methods are compared by employing both internal and external validity indices and suitable selections for the number of clusters are estimated using multiple of these indices and multiple clustering methods. Additionally, to simplify the interpretation of the data, low-dimensional visualisations using different state-of-the-art dimensionality reduction methods are presented.

Again, all methods are part of the integrative analysis framework developed during the doctoral project, and can be accessed online on the ArrayMining.net web-server [18].

#### 4.3.1 Unsupervised analysis - Methods

For the comparative evaluation of clustering methods in this study all possible combinations of 2 standardisation methods, 2 dimensionality reduction methods, 8 clustering methods and 5 validity indices were considered. The standardisation methods include the *classical standardisation* to mean 0 and standard deviation 1 (CL), and a *robust standardisation* method computing the median absolute deviation from the data's median (MD) [397]. For computing the dimensionality reduction, a classical *variance filter* (VF) is used, selecting the 2000 genes with the highest variance across all samples, and a *gene set analysis based*

*dimensionality reduction* (GSA), mapping the microarray probes onto 37 gene sets representing cancer-related cellular pathways and processes, obtained from the Van Andel Institute in Michigan, and using the “Parametric Gene Set Analysis” approach (PAGE) to summarise the gene expression values for each gene set [310].

The clustering methods include the following partition-based and hierarchical approaches:

**Partition-based clustering methods:**

- k-Means
- Partitioning around Medoids (PAM)
- Clustering Large Applications (CLARA)
- Self-Organising Maps (SOM)

**Hierarchical clustering methods:**

- Average linkage agglomerative clustering (AVL)
- Self-Organising Tree Algorithm (SOTA)
- Divisive Analysis clustering (DIANA)
- Hybrid hierarchical clustering (HYBRID)

Additionally, a self-devised consensus clustering approach (CONSENSUS) was implemented, combining the above methods using a cluster agreement matrix and a Simulated Annealing optimisation method (a detailed description of this method can be found in chapter 5, section “Ensemble and Consensus Analysis of Microarray Data”).

Moreover, in order to evaluate the benefit of different clustering methods and find a number of clusters, providing compact and well-separated clusters, the following internal cluster validity indices were computed for each clustering result, across cluster numbers varying between 2 and 8:

- Average silhouette width (SILHOUETTE), value range [-1, 1], to maximise
- Calinski-Harabasz index (CH), value range [0,  $\infty$ ], to maximise
- Dunn index (DUNN), value range [0,  $\infty$ ], to maximise
- C-index (C-INDEX), value range [0, 1], to minimise
- kNN-Connectivity (CONNECTIVITY), value range [0,  $\infty$ ], to minimise

The rationale behind this choice of indices was to obtain both a diverse selection of validity measures and to only use measures which are unbiased and have already been shown to perform well on synthetic data [230]. A detailed explanation of the above validity indices can be found in the literature survey in chapter 3 (section “Cluster validity / Selection of the number of clusters”).

In addition to these *internal* validity indices, the agreement with an *external* reference partition was measured using the adjusted rand index (ARI, for details see the section on external validity indices in chapter 3). Importantly, the reference partition corresponds to a partly subjective categorisation of microarray cancer samples according to clinical experts and not to a completely reliable “gold-standard”. Moreover, as outlined in chapter 3, there might be multiple meaningful structures and patterns in a microarray dataset,

hence, the ARI only provides a further confirmation for a clustering result, if the identified cluster structure happens to match well to the given reference partition (for a good match the ARI does not necessarily need to be close to the maximum value 1.0, but should be higher than ARIs computed from random model clustering solutions).

Finally, for the direct visual inspection of the data, four dimensionality reduction methods for obtaining low-dimensional visual representations were compared: *Principal Component Analysis* (PCA), *Independent Component Analysis* (ICA), *Locally Linear Embedding* (LLE) and *Isomap*. These approaches were chosen by taking into account the results of a previous comparative analysis of dimensionality reduction methods [398] and enable a user to directly identify natural groupings in 2D- and 3D-representations of the data. Moreover, the visualisations highlight outlier samples and patterns in the data density and variance across sample groups, which do not always become apparent in algorithmically obtained cluster groupings. Like all other algorithms above, these unsupervised methods are also part of the clustering module on the ArrayMining.net web-server for integrative microarray analysis [18].

All methods were applied to the breast cancer dataset obtained from the collaborating Nottingham Queen's Medical Centre [19, 349–351] (see the dataset description in the supervised analysis section), in order to identify informative structures in the data and the number and composition of sample clusters which can best be separated. Previously, different categorisations have been proposed in the biomedical literature for breast cancer tumour subtypes, including two-class groupings (e.g. luminal vs. non-luminal samples, see supervised analysis section) and three-class groupings (e.g. clinical grades 1 to 3, with 1 being the mildest form of breast cancer and 3 the most severe subtype). Thus, the identification of the number of clusters which results in the best cluster compactness and cluster separation across multiple clustering methods and validity indices is of great practical interest for clinical research. For future breast cancer microarray data analyses, similar comparative evaluations can help to choose subtype definitions that match well to natural grouping in the data and are useful for clinical prognosis and diagnosis, enabling accurate sample classifications using computational methods.

### 4.3.2 Unsupervised analysis - Results and discussion

The estimates for the optimal number of clusters on the QMC breast cancer dataset and the optimal validity index scores for all combinations of clustering methods and validity measures are shown in tables 4.12 (using the classical standardisation, CL) and 4.13 (using the robust median absolute deviation standardisation, MD). These tables include both the results for a simple variance-filter dimensionality reduction (VF) and a dimensionality reduction involving the extraction of cancer-related gene sets (GSA, see Methods section above). The optimal scores for a certain validity index are always shown in bold typeface in the tables.

The results reveal that across the great majority of validity indices and clustering methods, the optimal cluster number estimate is 2. This applies both to the CL and the MD standardisation and to both dimensionality reduction methods. Only the C-index tends to estimate the cluster number for optimal separation to be significantly higher (3, 5, 6 or 8), but this method also displays a higher variance in its estimates and has previously been shown not to perform as well as the other validity measures on simulated datasets [230]. Overall, 7 out of 10 method combinations on both the CL and the MD standardised data agree in their estimate that the optimal number of clusters is 2.

In general, for the CL and the MD standardisation, the qualitative results (i.e. the optimal cluster number estimates) were relatively similar, but the CL approach provided slightly better validity scores, hence, only

the CL standardisation was considered for further analysis.

When comparing the clustering results after variance filtering (VF) and after gene set analysis filtering (GSA) in detail, although both dimensionality reduction approaches provide the same estimate of the optimal number of clusters, the variance in the estimates across the different clustering methods is higher for VF, whereas the combination of functionally similar genes to a “meta-genes” representing a functional process in GSA tends to provide more robust estimates. This can also be seen when plotting the estimates for different validity indices against the number of clusters. Figure 4.8a) shows these estimates for the Calinski-Harabasz index when using the VF method and figure 4.8b) for the same index when using the GSA method (the score is scaled to range [0, 1]), and the trend from a maximum score for 2 clusters and decreasing scores for higher numbers of clusters is much clearer for the GSA approach than for VF, where the agreement between different methods is smaller. A similar trend can also be observed for validity indices that are to be minimised rather than maximised: In figures 4.9a) and b) the knn-connectivity validity scores were scaled to range [0, 1] and inverted by subtracting the scaled scores from 1, in order to make the results comparable to other validity measures like the Calinski-Harabasz (CH) index shown in figure 4.8. Although the variance in the estimates is generally higher than for the CH index, again the GSA method tends to provide more robust estimates, with a clear maximum at 2 and decreasing scores for higher cluster numbers.

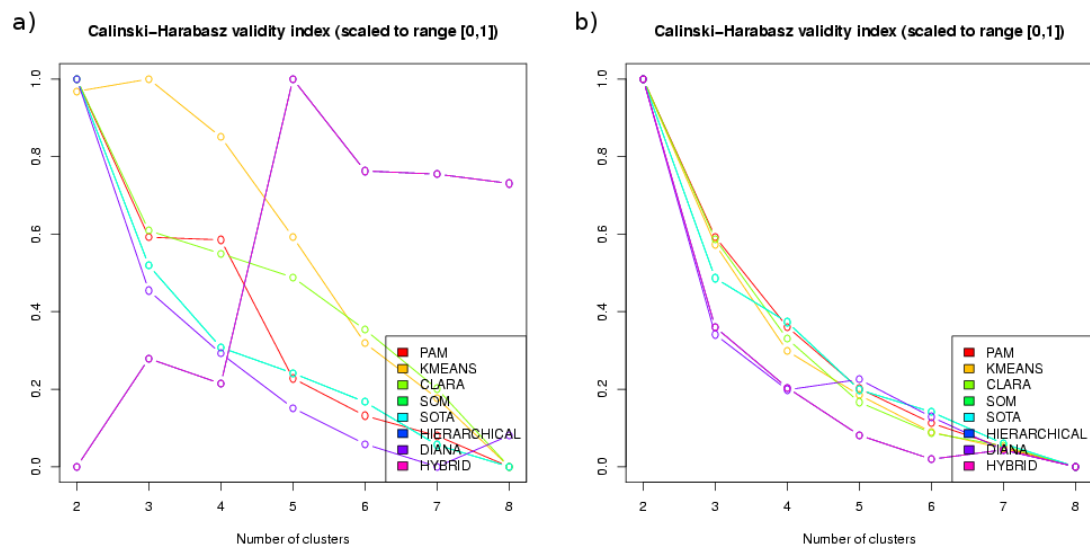


Figure 4.8: Visualisation of clustering results - Calinski-Harabasz index (scaled to range [0, 1]) vs. number of clusters: a) after variance filtering (VF); b) after gene set analysis filtering (GSA); In both cases the optimum number of clusters for most clustering methods is estimated to be 2.

In addition to the validity scores for complete clustering results, the different methods can also be compared based on confidence scores for the cluster assignments of single samples. This can be achieved with the *silhouette width* validity index, which provides a confidence measure for every single observation (see definition in chapter 3, in the section “Cluster validity / Selection of the number of clusters”). These silhouette widths have also been used to compute a global validity measure from their average across all samples, as reported in tables 4.12 and 4.13. However, additionally investigating the silhouette widths for single samples reveals specific groups of high-confidence sample assignments, which are reliably clustered together, and low-confidence sample assignments which might result from low-quality experimental measurements



Table 4.12: Comparison of clustering methods (standardisation: CL, dimensionality reduction: VF or GSA, QMC breast cancer dataset)

Validity index	Method	Opt. cluster number (VF)	Scores (VF)	Opt. cluster number (GSA)	Scores (GSA)
Calinski-Harabasz	PAM	2	10.19	2	58.05
	KMEANS	3	9.48	<b>2</b>	<b>58.94</b>
	CLARA	<b>2</b>	<b>10.62</b>	<b>2</b>	<b>58.94</b>
	SOM	2	10	2	50.34
	SOTA	2	10	2	50.34
	AVL	5	4.65	2	51.48
	DIANA	2	9.52	2	52.27
	HYBRID	5	4.65	2	51.48
Dunn	PAM	8	0.47	2	0.29
	KMEANS	7	0.48	2	0.28
	CLARA	7	0.47	2	0.28
	SOM	6	0.52	4	0.26
	SOTA	6	0.52	4	0.26
	AVL	<b>2</b>	<b>0.65</b>	2	0.28
	DIANA	2	0.47	<b>3</b>	<b>0.34</b>
	HYBRID	<b>2</b>	<b>0.65</b>	2	0.28
Silhouette	PAM	2	0.15	2	0.37
	KMEANS	3	0.07	2	0.36
	CLARA	2	0.13	2	0.36
	SOM	2	0.13	2	0.31
	SOTA	2	0.13	2	0.31
	AVL	<b>2</b>	<b>0.23</b>	2	0.38
	DIANA	2	0.14	<b>2</b>	<b>0.39</b>
	HYBRID	<b>2</b>	<b>0.23</b>	2	0.38
knn-Connectivity	PAM	2	25.67	2	16.93
	KMEANS	3	97.93	2	25.77
	CLARA	2	40.08	2	25.77
	SOM	2	35.85	2	39.08
	SOTA	2	35.85	2	39.08
	AVL	<b>2</b>	<b>2.93</b>	<b>2</b>	18.51
	DIANA	2	27.36	2	<b>12.15</b>
	HYBRID	<b>2</b>	<b>2.93</b>	<b>2</b>	18.51
C-index	PAM	<b>5</b>	<b>0.43</b>	<b>8</b>	<b>0.25</b>
	KMEANS	3	0.45	6	0.27
	CLARA	6	0.46	8	0.26
	SOM	5	0.47	3	0.27
	SOTA	5	0.47	3	0.27
	AVL	2	0.64	6	0.31
	DIANA	3	0.48	5	0.3
	HYBRID	2	0.64	6	0.31

Table 4.13: Comparison of clustering methods (standardisation: MD, dimensionality reduction: VF or GSA, QMC breast cancer dataset)

Validity index	Method	Opt. cluster number (VF)	Scores (VF)	Opt. cluster number (GSA)	Scores (GSA)
Calinski-Harabasz	PAM	2	8.95	3	31.9
	KMEANS	3	8.97	<b>3</b>	<b>33.63</b>
	CLARA	<b>2</b>	<b>9.55</b>	3	31.25
	SOM	2	7.35	2	28.43
	SOTA	2	7.35	2	28.43
	AVL	5	4.46	2	29.99
	DIANA	2	7.92	2	30.65
	HYBRID	5	4.46	2	29.99
Dunn	PAM	3	0.48	7	0.25
	KMEANS	8	0.45	2	<b>0.28</b>
	CLARA	6	0.48	8	<b>0.28</b>
	SOM	3	0.53	6	0.24
	SOTA	3	0.53	6	0.24
	AVL	<b>2</b>	<b>0.65</b>	2	0.26
	DIANA	2	0.46	<b>8</b>	<b>0.28</b>
	HYBRID	<b>2</b>	<b>0.65</b>	2	0.26
Silhouette	PAM	2	0.14	2	0.28
	KMEANS	3	0.11	2	<b>0.34</b>
	CLARA	2	0.14	2	0.2
	SOM	2	0.13	2	0.25
	SOTA	2	0.13	2	0.25
	AVL	<b>2</b>	<b>0.23</b>	2	0.33
	DIANA	2	0.14	<b>2</b>	<b>0.34</b>
	HYBRID	<b>2</b>	<b>0.23</b>	2	0.33
knn-Connectivity	PAM	2	29.17	3	47.19
	KMEANS	3	56.07	2	15.48
	CLARA	2	32.45	2	35.84
	SOM	2	34.1	2	35.99
	SOTA	2	34.1	2	35.99
	AVL	<b>2</b>	<b>2.93</b>	<b>2</b>	24.38
	DIANA	2	28.07	2	<b>16.92</b>
	HYBRID	<b>2</b>	<b>2.93</b>	<b>2</b>	24.38
C-index	PAM	8	0.48	<b>6</b>	<b>0.23</b>
	KMEANS	<b>3</b>	<b>0.46</b>	4	0.3
	CLARA	3	0.48	7	0.26
	SOM	2	0.49	3	0.28
	SOTA	2	0.49	3	0.28
	AVL	3	0.64	6	0.32
	DIANA	3	0.49	7	0.32
	HYBRID	3	0.64	6	0.32

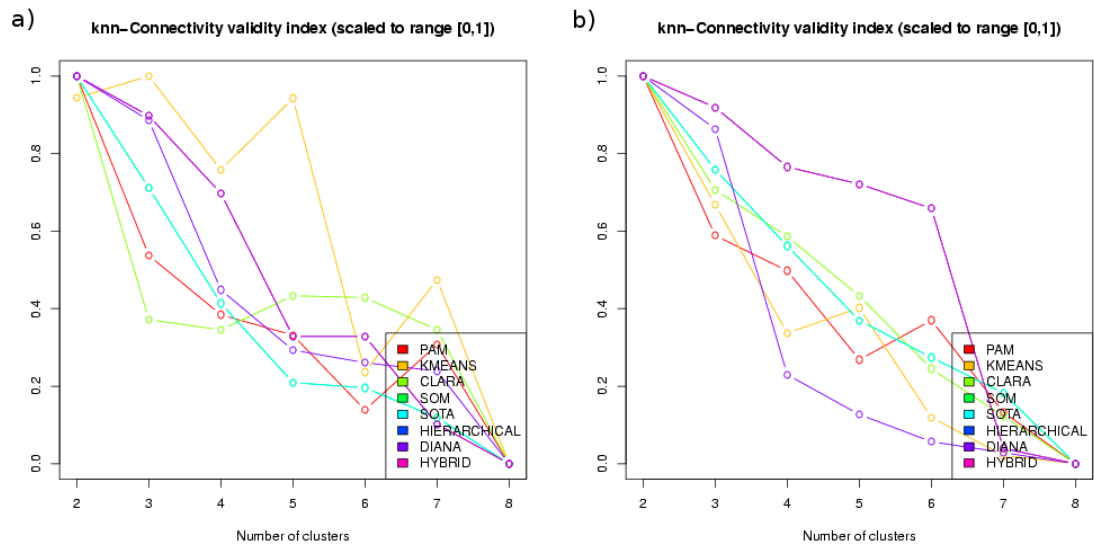


Figure 4.9: Visualisation of clustering results - kNN-Connectivity index (inverted and scaled to range [0, 1]) vs. number of clusters: a) after variance filtering (VF); b) after gene set analysis filtering (GSA); In both cases the optimum number of clusters for most clustering methods is estimated to be 2.

and/or membership in a sample group that displays high variance in the gene expression values. Figure 4.10 shows a *silhouette plot* visualisation of these sample validity scores as horizontal bars with lengths corresponding to the silhouette widths (using the consensus clustering results obtained from the combination of all 8 clustering methods, once for the VF and once for the GSA dimensionality reduction). The cluster number for this consensus clustering was set to 2, as suggested by the results from the validity index analysis of the single clustering methods (see above). Interestingly, the sizes of the two clusters are approximately similar for VF and GSA, however, the silhouette widths (i.e. confidence scores) for the sample assignment in the GSA method are much larger than for the VF approach (the average silhouette width 0.37 for GSA, is more than twice as high as for VF with 0.15). This result again suggests that the robustness can be increased when using the GSA-based dimensionality reduction.

Apart from the validity measures, another possibility to analyse the quality of tentative clustering solutions is to plot the data projected onto the first two principal components and compare how well the clusters are separated in this visualisation. Corresponding *PCA plots* for the consensus clustering results after VF and GSA dimensionality reduction are shown in figure 4.11. Clusters are represented by ellipses in these plots and the overlap between the two ellipses is clearly smaller for the GSA method than for the VF method, indicating that the GSA method does not only provide more robust clustering results but also a better separation between the clusters. Moreover, the first two principal components (PCs) for the data reduced with the GSA method explain 62% of the point variability, whereas for the VF method only 25% are covered by the first two PCs.

Finally, as mentioned in the Methods section, clustering results can also be compared using external validity measures like the adjusted rand index (ARI), if a known reference partition (i.e. a biological categorisation of the samples) is already available. Importantly, a partly subjective categorisation into reference partitions cannot be considered as a reliable “gold standard”, and multiple biologically meaningful cluster structures might occur in the data. However, if the agreement of the reference partition to a microarray-data based sample clustering is significantly higher than the agreement with random cluster assignments, the ARI

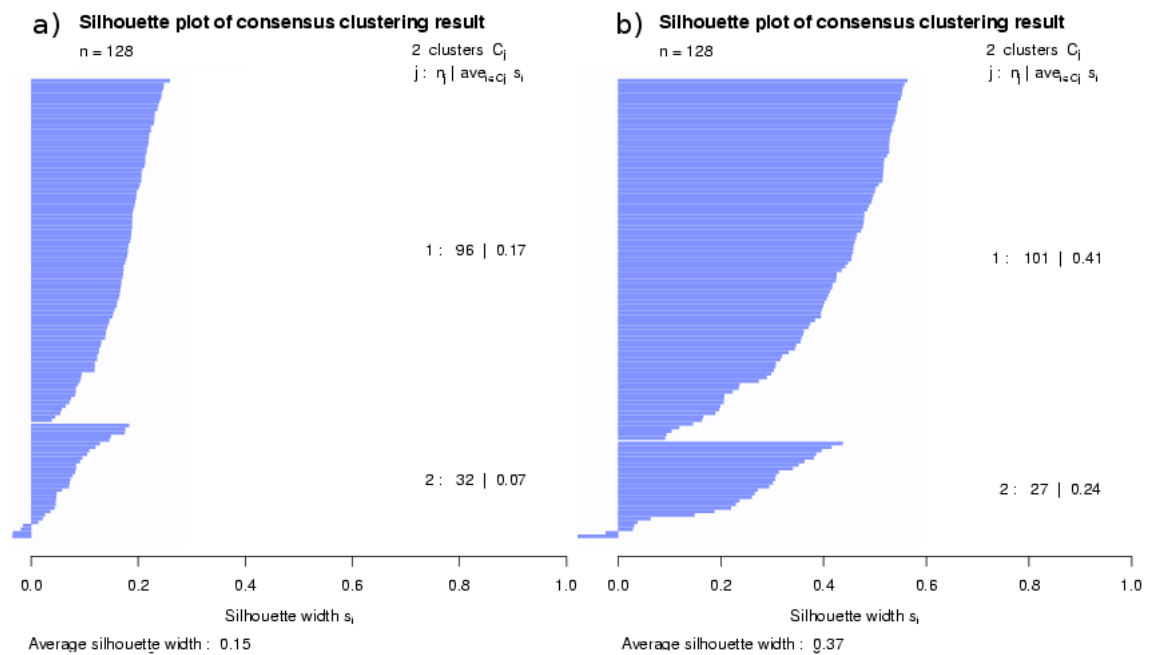


Figure 4.10: Visualisation of clustering results - Silhouette plot (length of horizontal bars represents confidence for each cluster assignment): a) after variance filtering (VF); b) after gene set analysis filtering (GSA); The average confidence (silhouette width) is more than twice as high after GSA dimensionality reduction in comparison to a variance filter reduction

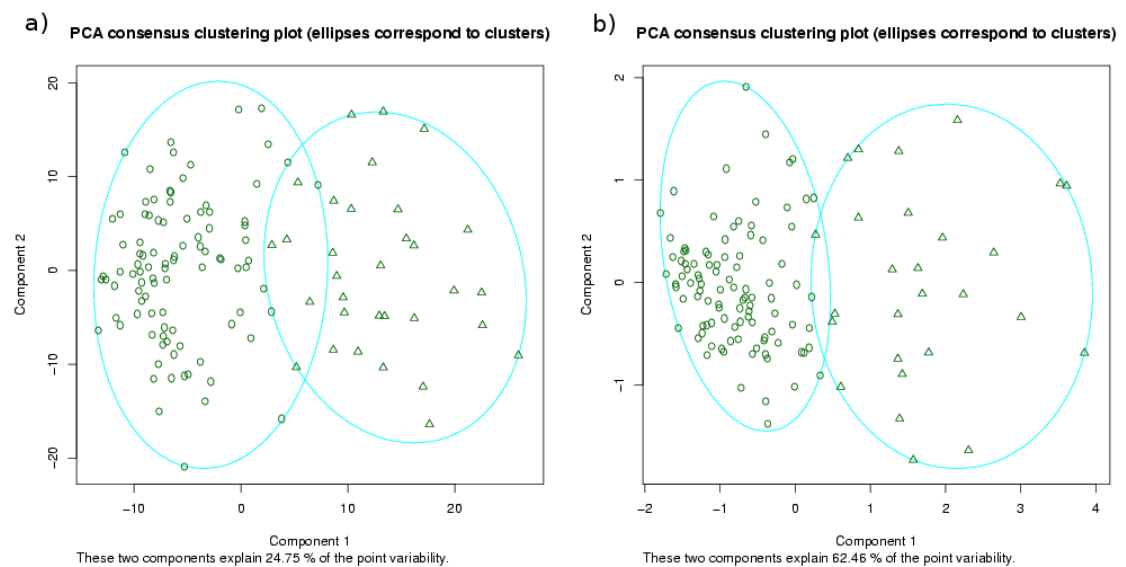


Figure 4.11: Visualisation of clustering results - PCA plot of consensus clustering results (principal component 2 vs. principal component 1): a) after variance filtering (VF); b) after gene set analysis filtering (GSA).

can provide greater confidence in sample cluster assignments, if an agreement between the categorisations derived from different methods and data sources is found.

For the breast cancer data, a reference assignment of the samples into three clinical tumour grade categories had been provided in combination with the data (33 samples belong to grade 1, 52 samples have grade 2, and 43 samples grade 3; the higher this clinical grade, the more severe the tumour type). Accordingly, the clustering results which provided the best validity scores (for a cluster number of 2, see above) could not be used for a comparison against this 3-class reference partition, but the validity measures also showed that a clustering into 3 groups corresponds to the second best number of clusters. Thus, the corresponding 3-group clustering results were combined into a consensus clustering and both this consensus and the single-method clusterings were compared against the reference. As expected, the obtained ARI scores were generally low ( $< 0.25$ ) in relation to the maximum achievable score (1.0), since the reference clustering does not represent a reliable ground truth and necessarily match to natural groupings in the data. However, the majority of clustering methods, including the consensus clustering, provided significantly larger ARI scores than 10000 random clusterings (created using a Mersenne-Twister stochastic random number generator [399]) compared against the reference. The results are shown in a graph in figure 4.12 for the VF dimensionality reduction, and in figure 4.13 for the GSA reduction.

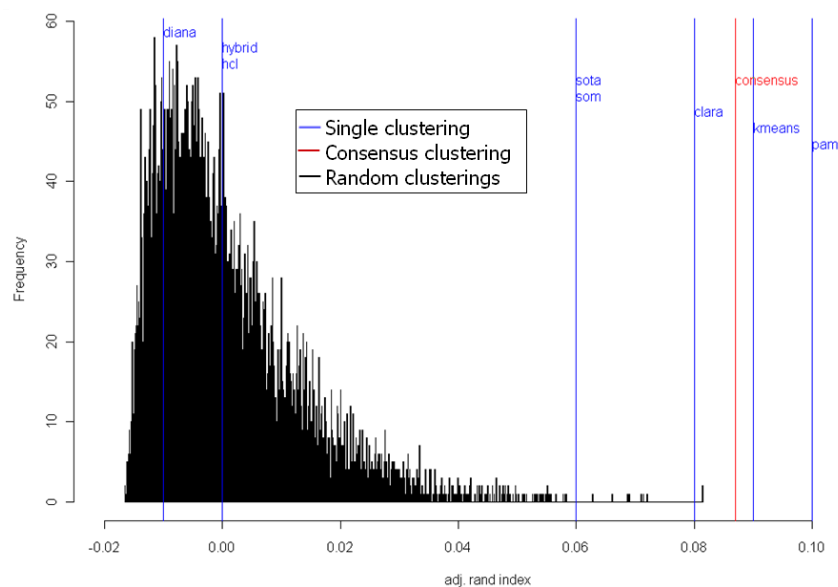


Figure 4.12: Histogram of adjusted rand indices (ARIs) between the tumour grade 3-class reference partition and 10000 random clustering results (VF dimensionality reduction, the ARIs for microarray-based clustering methods are shown as blue vertical lines (in contrast to the random clusterings, the height does not have a frequency interpretation)

For the VF reduction, all clustering methods except the hierarchical approaches provided significantly higher ARI scores than the 10000 random clusterings. The consensus clustering belonged to the methods with the highest scores, and only two single-method clusterings had a slightly higher ARI than the consensus (K-Means and PAM). Although similar trends were observed for the GSA reduction, with the partition-based methods again reaching higher ARIs than the hierarchical methods, in comparison to the VF reduction all clustering methods reached larger ARI scores, and all methods performed significantly better than the great majority of random clusterings (see figure 4.13, partition-based methods achieved ARIs exceeding the best scores for the random model by multiple standard deviations). Again, the consensus clustering across all methods provided significantly better results than the random clusterings, but the

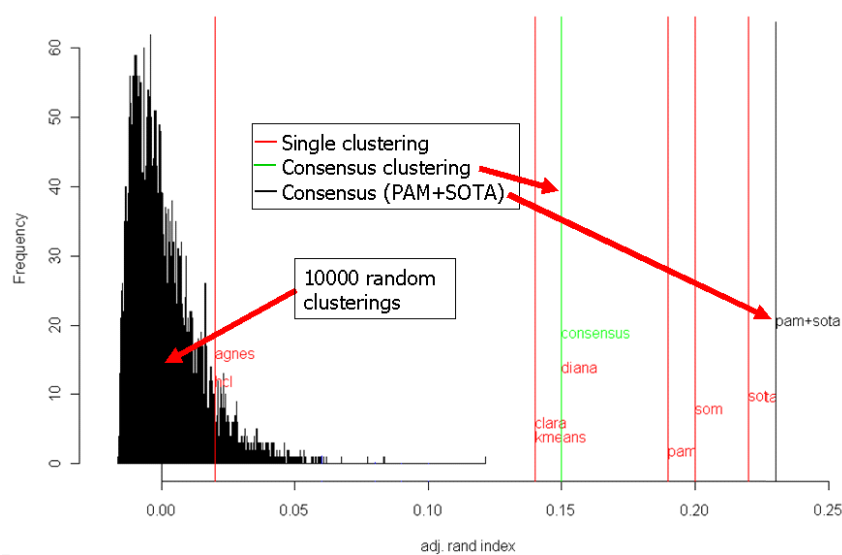


Figure 4.13: Histogram of adjusted rand indices (ARIs) between the tumour grade 3-class reference partition and 10000 random clustering results (**GSA** dimensionality reduction, the ARIs for microarray-based clustering methods are shown as blue vertical lines (in contrast to the random clusterings, the height does not have a frequency interpretation)

overall best results were achieved by some of the single-algorithm methods. Since the performance of the hierarchical clustering methods in terms of the ARI was again low, and the consensus clustering did not only include hierarchical methods but also methods with a low diversity in their clustering results, a second consensus clustering was computed, including only two diverse clustering methods (PAM+SOTA, see figure 4.13). For this algorithm combination, the consensus clustering provided the best overall clustering result, outperforming the single-method clusterings. However, since the user does not know *a priori* which types of clustering methods will perform well on a new microarray dataset, the consensus clustering involving all clustering methods provides a more generally applicable, robust solution, providing good clustering solutions on average, even if not all of the single-method clusterings are outperformed. Moreover, the consensus clustering methodology has room for further improvements, since very recently, new extensions for consensus clustering have been proposed (e.g. link-based clustering [222], see literature review in chapter 3), that could be combined with the current implementation in the framework for integrative biological data analysis.

Apart from the clustering results, the dimensionality reduction of the data can also be regarded as an unsupervised analysis technique for data interpretation, producing low-dimensional representations for visual inspection, e.g. to identify outlier samples, high-density groupings of samples and other structures in the data. Thus, four dimensionality reduction methods were used to create 3D data representations, including *Principal Component Analysis* (PCA, figure 4.14), *Independent Component Analysis* (ICA, figure 4.15), *Isomap* (figure 4.16) and *Locally Linear Embedding* (LLE, figure 4.17).

As in the previous evaluations, all results were compared between the variance-filtered data (VF) and the gene set analysis based pre-processing (GSA). Generally, the data visualisations for a specific dimensionality reduction method tend to be similar across the two pre-processing methods, and in none of the 3D representations a perfect separation between the three sample classes is obtained (the classes are represented by different colours and shapes in the plots: green box = tumour grade 1, red sphere = tumour grade 2, blue pyramid = tumour grade 3). However, in some plots a certain degree of *assortative mixing* [344]

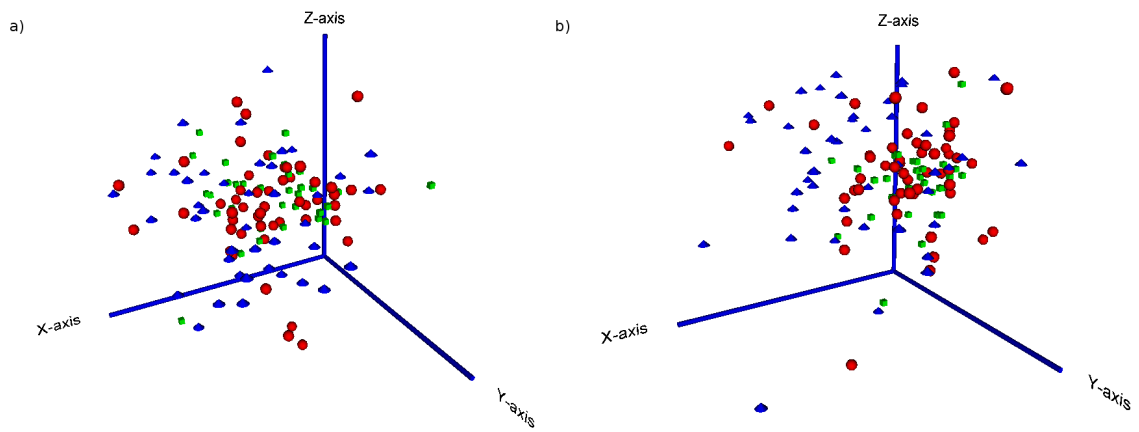


Figure 4.14: Principal Component Analysis (PCA), 3D representation of the breast cancer dataset a) after variance filtering (VF), b) after gene set analysis (GSA); green = tumour grade 1, red = tumour grade 2, blue = tumour grade 3.

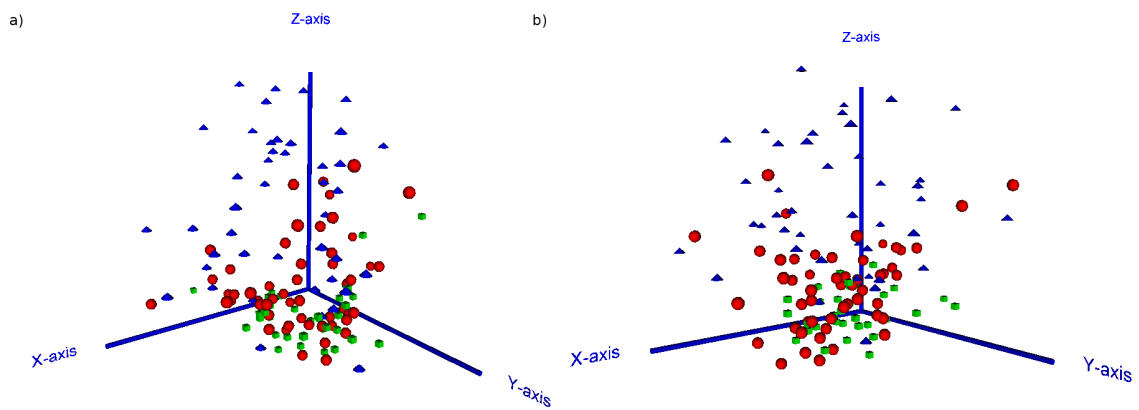


Figure 4.15: Independent Component Analysis (ICA), 3D representation of the breast cancer dataset: a) after variance filtering (VF), b) after gene set analysis (GSA); green = tumour grade 1, red = tumour grade 2, blue = tumour grade 3

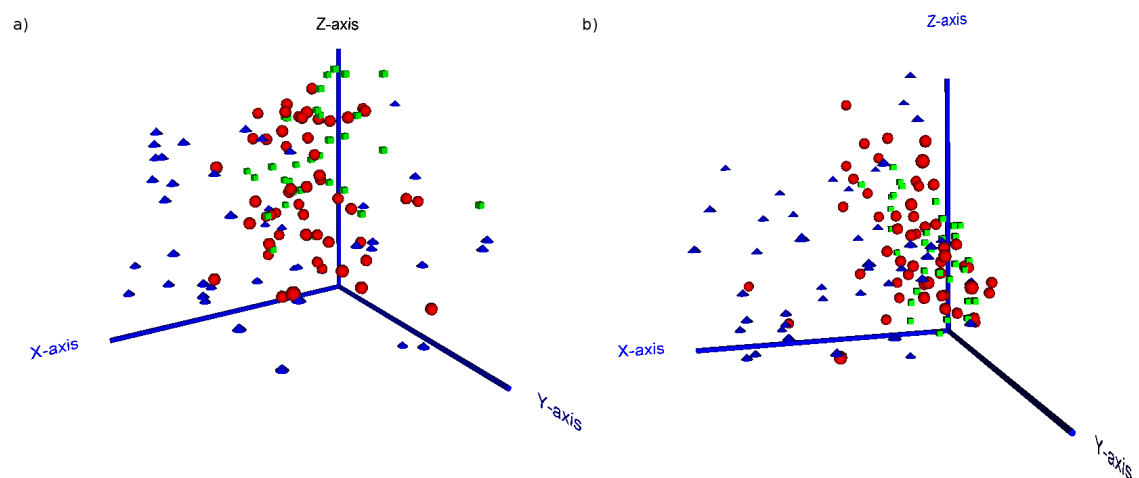


Figure 4.16: Isomap 3D representation of the breast cancer dataset a) after variance filtering (VF), b) after gene set analysis (GSA); green = tumour grade 1, red = tumour grade 2, blue = tumour grade 3

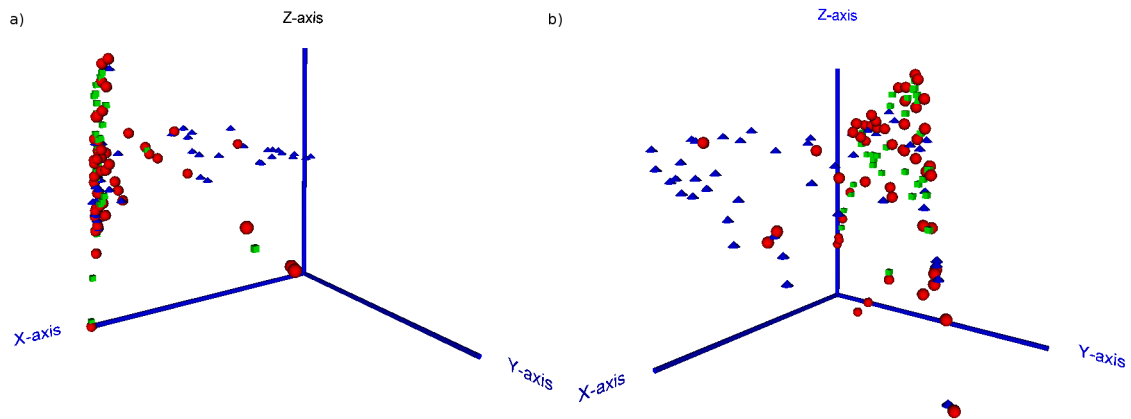


Figure 4.17: Locally Linear Embedding (LLE) 3D representation of the breast cancer dataset a) after variance filtering (VF), b) after gene set analysis (GSA); green = tumour grade 1, red = tumour grade 2, blue = tumour grade 3

can be observed, i.e. samples from the same class tend to be located closer together in space than samples from different classes. The tumour grade 1 samples tend to be most homogeneous, and are clustered in a relatively dense region in space, whereas the grade 2 and grade 3 samples display a higher variance in their positions. As expected, the best separation is obtained between the grade 1 and grade 3 samples, representing the most diverse tumour sub-types, while the grade 2 samples largely overlap with both of the other classes.

The ICA visualisations tend to provide the best separation, with a very dense grade 1 group having only a small overlap with the grade 3 group. PCA and Isomap provide relatively similar representations, where the grade 1 and grade 2 groups have a large overlap, and the grade 3 group displays a larger variance than the other groups. The LLE visualisations differ markedly from the other representations, displaying a long vertically stretched sample group consisting mainly of group 1 and group 2 samples, whereas most of the group 3 samples are separated from this group in horizontal direction. Marked differences between the VF and the GSA pre-processing in terms of class separation and class homogeneity cannot be observed in most visualisations, apart from the Isomap representation, where the grade 1 and grade 2 sample groups are clearly more homogeneous after the GSA pre-processing than after the variance filtering.

Although these visualisations suggest that the breast cancer classes are not easily separable in three dimensions, a certain extent of class separation can be observed between grade 1 and grade 3 samples, in particular in the ICA plots (this can also be seen in figure 8.2 in chapter 7, where only the grade 1 and grade 3 samples from the same dataset are visualised). The results match to earlier findings by Zervakis *et al.*, according to which breast cancer tumour classes tend to be difficult to separate in comparison to other cancer types, like leukaemia and colon cancer data [400] (figure 1 in the corresponding publication contains a similar plot to the ones shown here). However, as shown in the section on supervised analysis (see above), by using labelled training data to train a prediction model, samples from different tumour classes can be more successfully separated than with unsupervised analysis techniques.

### 4.3.3 Unsupervised analysis - summary and conclusion

In summary, the results of the unsupervised analysis techniques highlight again that ensemble/consensus techniques provide a significant added value for microarray analysis, leading to higher robustness and often



to better scores in terms of multiple performance measures. In particular, the observation that the GSA method tends to provide more robust results than the VF method, and similarly that the consensus clustering tends to provide higher ARI scores than the average of all methods, match well to the intuition that combining the information from multiple sources (i.e. multiple genes in case of GSA method, and multiple clustering algorithms in case of the consensus clustering) can help to overcome the shortcomings of the single input data sources and methods.

Overall, clustering results on microarray data are more difficult to evaluate than supervised analysis, because multiple biologically meaningful patterns might exist in the data and external reference groupings of the samples do not provide a reliable ground truth. However, the agreement between algorithmic clustering results and an expert-based tumour grade categorisation of the samples was significantly better than the agreement of this categorisation with 10,000 random clusterings, providing an increased confidence in the biological meaningfulness of the algorithmic clustering solutions. Moreover, an analysis of the data using low-dimensional visualisations reveals which tumour sub-types can be distinguished best. In combination with the estimation of the optimal number of classes from the clustering analysis (suggesting that two breast cancer classes can best be separated according to the given data), this can help the experimenter to re-define the sample classes in order to obtain more meaningful tumour sub-type definitions, and a better performance in subsequent supervised analyses.

## Chapter 5

# An Integrative Framework for Gene/Protein Expression Data Analysis (ArrayMining)

### Chapter abstract

Given the multitude of publicly available functional genomics datasets and the great variety of higher-level analysis techniques (see discussion in chapter 3 and comparative evaluation in chapter 4), an integrative framework to exploit the synergies of different data sources and algorithms was developed as part of this doctoral project, and its main component, a tool set for microarray data analysis, was made available in the web-application *ArrayMining* [18]. Figure 5 highlights ArrayMining's modules and the links between them and other analysis tools in the framework.

In this chapter, ArrayMining's machine learning and network analysis techniques dedicated to the integrative analysis of high-dimensional gene and protein expression data will be discussed in detail. The implemented algorithms consist of known state-of-the-art statistical learning methods, chosen using the information from the literature survey (chapter 3) and the comparative evaluation (chapter 4), as well as *new algorithms*, *ensemble* and *consensus techniques*, and *new pipelines* developed during the PhD project.

The first part of the chapter will provide an overview of the analysis modules for different problem types, discuss the automatic data processing and gene/protein name normalisation techniques, and the details of the ensemble and consensus methods implemented within these modules (partly using material from a previous publication [18]). In the second part, the possibilities for creating new cross-domain analysis pipelines using sequential combinations of the modules will be laid out. Example illustrations for these pipelines can also be found in the appendix.

Novel data analysis methods in the framework, which extend beyond classical machine learning methods (i.e. the components which are greyed out in figure 5), will be discussed in the following chapters. Specifically, chapter 6 will present more general techniques for the analysis of gene and protein lists, which do not require numerical input data. Chapter 7 will be dedi-

cated to novel approaches for increasing the interpretability of computational models derived from the data.

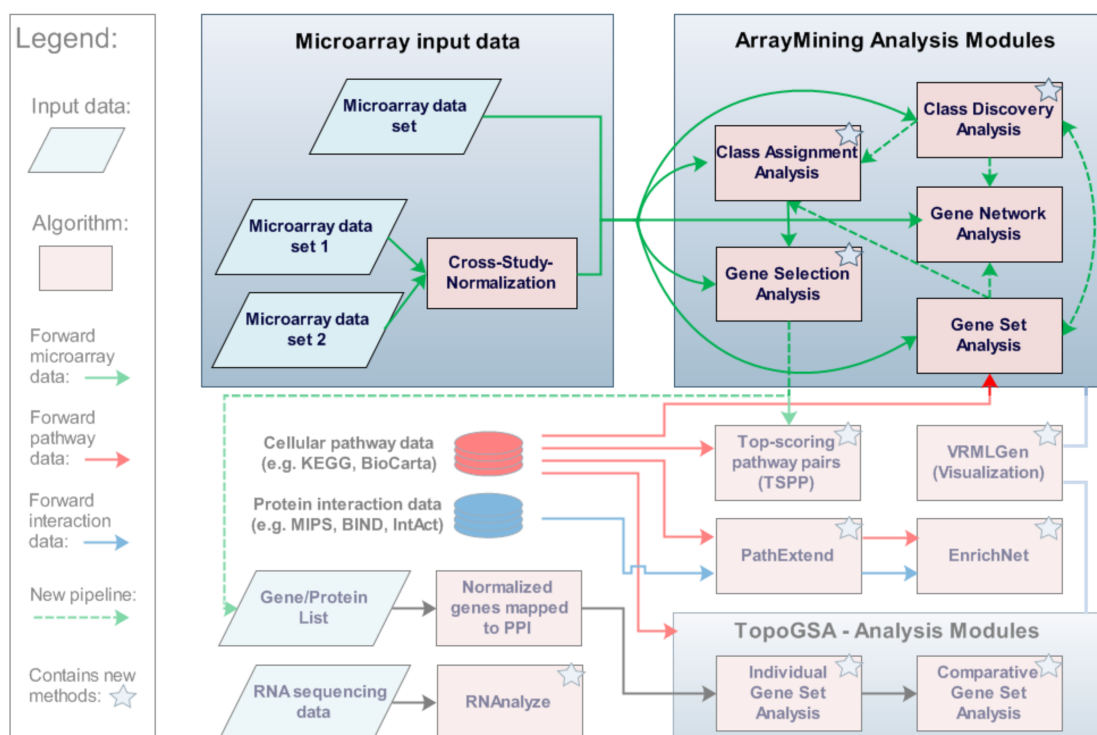


Figure 5.1: The ArrayMining system for microarray data analysis is one of the main components of the integrative data analysis framework presented in this thesis (highlighted by the colour contrast).

#### 5.0.4 ArrayMining - Overview, motivation and implementation

**Overview:** ArrayMining is a software tool set and web-application for microarray data analysis, focussing on ensemble algorithms and cross-domain analysis techniques. It consists of six main analysis modules: *Cross-Study Normalisation*, *Gene selection*, *Class Discovery*, *Class Assignment*, *Gene Set Analysis* and *Network Analysis*. Each of these modules contains multiple algorithms for the same problem type accessible through a unified web-interface. The user can upload new data in different raw or pre-processed formats (tab-delimited flat file format or as zip-compressed Affymetrix CEL-files which will be automatically extracted, normalised and summarised using the Robust Microarray Analysis (RMA) method [401]). Alternatively, various example datasets have been made available directly on the webpage, and access to the GEO database [27], the largest public microarray data base, is provided on the class discovery module. After submitting an analysis task, an output webpage containing the downloadable results as plots, tables, 3D-VRML visualisations etc. is generated. Depending on the chosen module and algorithm, the data can be forwarded to further analysis modules (see data forwarding pipelines in figure 5, new pipelines are highlighted by dotted lines) and is automatically interlinked with annotation data from external web-tools and data bases, e.g. ENSEMBL [28], DAVID [29], Gene Ontology [30] and KEGG [31].

**Motivation:** Although statistical programming frameworks like R [353] and Matlab [402] already allow users to develop and apply complex scripts for expression data analysis, they are difficult to use for non-

experts and there is a high risk of deviating from standard guidelines. To obviate the need for specialized programming skills and manual software installations, several other web-based tools for gene expression analysis have been presented in recent years. Currently available integrative online analysis services include *GEPAS* [403], *Expression Profiler* [404], *ASTERIAS* [405], *EzArray* [406], *CARMAweb* [407], *MAGMA* [408], *ArrayPipe* [409], *RACE* [410], *WebArray* [411] and *MIDAW* [412]. These web-based systems provide methods for a multitude of data pre-processing and analysis purposes ranging from image analysis, missing value imputation, single-study normalisation, gene filtering and gene name conversion to higher-level analysis methods for clustering, gene selection and gene annotation, prediction, data visualisation and gene set enrichment analysis, among others.

Additionally, numerous web-applications have been developed and optimised for specialized analysis tasks, e.g. biclustering of genes and samples [413], co-clustering of genes with similar functional annotations [414], inference of gene regulatory relationships [415] and cross-species clustering [416].

Although various tools provide a choice and comparison between different algorithms for one analysis task, previous integrative analysis software did not enable the user to easily combine multiple methods using ensemble learning and consensus clustering techniques, or exploiting multiple validation approaches. However, studies from the literature have shown that microarray analysis can profit from ensemble feature selection, ensemble prediction and consensus clustering methods in terms of both robustness and accuracy [9–12], suggesting that there is significant potential still to be exploited with these approaches.

Similarly, it would be desirable not only to combine different algorithms but also different data sets related to the same biological problem. Although currently available cross-study normalisation methods use simplified assumptions and are limited in applicability and accuracy, various successful applications [15, 417] have shown that the benefits of an increased sample size can outweigh the loss of information due to the normalisation process.

For these reasons, ArrayMining was developed as a new web-application that provides access to multiple algorithms for each of the most common tasks in statistical microarray analysis, namely gene selection, sample clustering, sample classification, network and gene set analysis, available from a single, easy-to-use interface. In contrast to other web-tools, providing the results of individual methods as outputs, here, alternative techniques can be compared and their different strengths combined using *ensemble* and *consensus* approaches. Likewise, instead of using only data from a single study, different *cross-study normalisation* methods are made available to integrate similar data from different experimental platforms and compare the results using density and quantile-quantile plots.

Apart from these combinations of data sets and methods within an analysis module, different modules have been interlinked in new ways, enabling for example the integration of gene set analysis with clustering or network analysis, or cross-study analysis with gene selection or prediction. These *new analysis pipelines* have a similar practical value as new algorithms and method combinations, because they provide different insights than standard analysis approaches and can lead the experimenter to a new interpretation the data (e.g. by inferring cellular pathway associations from the similarity of deregulation patterns in corresponding gene sets).

Other new features include access to an in-house developed rule-based evolutionary classification algorithm, automatic parameter selection mechanisms in all modules, the availability of specific cancer-related gene sets for enrichment analysis in addition to gene sets from KEGG and GO, and a 3D VRML visualisation of clustering results using a self-devised software package for creating interactive 3D data representations [24] (see the discussion of the VRMLGen tool in chapter 7).

Since the above methods and features are not available in other web-tools or software packages, and similarly, other tool sets include methods distinct from the system discussed here, ArrayMining is designed as a complement rather than an alternative to existing services, and has been interlinked with other external web-applications.

**Implementation:** ArrayMining uses software written in the programming languages R [353] and C++ and a PHP-interface combining all implementations together. All analysis modules run on an Apache web server. The system uses in-house algorithms and implementations as well as standard packages from the Bioconductor project [418] (see the discussion of the single modules below for more details).

## 5.1 Automatic Data Processing and Gene/Protein Name Normalisation

One of the main goals behind the development of ArrayMining.net was to facilitate the analysis of complex, high-dimensional biological data for users with no prior background knowledge in computer science and statistics. Currently, many wet-lab experimenters, including academic and clinical researchers, cannot always benefit from the capabilities of recent analysis approaches, because setting up a corresponding data processing pipeline (with suitable parameter selections and analysis and validation procedures following accepted guidelines) would be a time-consuming and difficult task, often requiring knowledge about the inner workings of several algorithms. Instead, experimenters often apply standard microarray software, which is directly supplied with the array scanning machine and typically only contains a single dedicated algorithm for some of the most common analysis tasks. This prevents the user from comparing different methods, achieving performance gains from ensemble and consensus techniques and gaining new biological insights from integrative cross-domain analysis methods.

For this reason, the ArrayMining framework was designed to facilitate complex analyses by automating analysis tasks wherever possible, and providing an easy-to-use interface to control and run a wide choice of feature selection, clustering, prediction, gene set analysis and cross-study normalisation methods. Setting up an individual analysis pipeline only requires a few mouse-clicks, and the results can be explored interactively in low-dimensional data visualisations and sortable tables, in some cases with dynamic and expandable content.

In this section, the automatic data processing and gene/protein name normalisation features of the framework will be discussed. They allow experienced users to save time in common data analysis tasks, and prevent researchers without prior experience in the field of microarray analysis from deviating from established guidelines and practices.

### 5.1.1 Automatic data processing

Automatic normalisation and pre-filtering of high-dimensional data, and automated parameter selection within the generation of clustering, prediction and co-expression network models requiring fast processing using heuristics that can adapt to different types of input data.

Particularly, when users upload gene or protein expression data on a public web application, a great variety of file formats and experimental platforms have to be supported, and format errors and missing data entries have to be recognised. In the ArrayMining framework, both pre-normalised microarray data in tab- or

space-delimited flat file format, and raw data in the standard CEL-format (compressed in a zip-archive) is accepted and recognised as input. More importantly, the framework identifies common errors and provides the user with easily understandable warning messages, if an uploaded input file does not match with the format specifications (e.g. “Your input file contains missing entries or non-numeric characters in line 5” or “Your input file does not contain any class labels” for a supervised analysis).

Dedicated algorithms are required to find optimal parameter settings for various algorithms. To relieve the user from having to manually adjust internal settings, heuristics for automatic parameter selection were used wherever possible. Specifically, the following techniques were used on the different analysis modules:

- **Feature Selection and Pathway analysis module:** Most feature selection methods do not require many parameter settings, apart from the possibility to limit the number of selected attributes. The ArrayMining feature selection module contains fully automated selection approaches, like the CFS method [160, 161], which can also determine an optimal number of selected features for a predictive analysis by scoring the relevance and redundancy of feature subsets (see feature selection section in chapter 3). However, in order to let the user control the behaviour of the selection methods, a maximum number of features can always be specified additionally, e.g. to reduce the runtime, or to only identify the most significant genes/proteins. In the feature ranking list obtained as the main output of an analysis, the user can sort the identified features using different scores (e.g. significance scores adjusted for multiple testing and relative expression fold changes) and identify attributes whose values exceed an arbitrary threshold for one of these scores. The same techniques are used on the pathway analysis module (see detailed description below), which also solves an attribute selection problem, but on a set of extracted pathway expression fingerprints.
- **Class Assignment module:** In order to choose model parameters for maximum predictive performance automatically, e.g. the number of nearest neighbours in the k-NN algorithm or the capacity constant  $C$  in the C-SVM algorithm, a nested (in most cases leave-one-out) cross-validation using grid search is applied. Importantly, not the parameter values with smallest cross-validation error are chosen, but those parameters for the least complex model within one standard deviation of the optimal settings (e.g. for k-NN this would be the model with largest  $k$  within one standard deviation of the parameter  $k$  with smallest cross-validated error). This common regularisation technique reduces the model complexity helps to avoid overfitting to noisy and spurious structures in the data.
- **Class Discovery module:** Selecting optimal parameters for clustering algorithms on the class discovery module is more difficult than the parameter selection for supervised modules, because there is no “gold standard” for an optimal clustering solution. To solve the problem of finding a suitable selection for the number of clusters, the module computes a combination of five diverse cluster validity indices (Calinski-Harabasz, Silhouette width, Dunn index, C-index and knn-Connectivity; see also chapters 3 and 4), and three consensus rankings of clustering solutions from 2 to 8 clusters using these validity indices (majority vote, median and sum of ranks; see detailed discussion in section 5.2). The user can also inspect the rankings using the individual validity indices (as plots and in tabular form) and enter class labels as input to obtain an estimate of the association between the best clustering result and the given outcome labels using the adjusted rand index.

Moreover, the class discovery module also provides the option to apply an automatic feature-preserving dimensionality reduction on the data using different techniques. To remove uninformative low-variance genes, while at the same time retaining low-variance regulator genes, the automatic filter size detection method SUMCOV [145] can be employed. Alternatively, a sparse PCA filtering [146]

can be computed automatically. If the user wishes to use a classic variance filter with a self-defined threshold value, this is also supported as a further option (although not recommended, due to the risk of removing low-variance regulators).

- **Network Analysis module:** The task of identifying co-regulated or co-expressed genes and visualising these relations in a network representation also involves the selection of different parameters. On the ArrayMining Network Analysis module this task is facilitated by only requiring the user to specify an edge adjacency threshold, corresponding to a trade-off between the significance of the included edges and the coverage of feature relations. Other internal parameters are chosen automatically by the algorithm (see section 5.2 for details on the network analysis).

### 5.1.2 Gene/Protein name normalisation:

As discussed in the literature survey (see section 3.1.4 in chapter 3), the mapping of gene and protein identifiers in different formats onto a single standardized format is a common task in cross-domain integrative analyses. Due to the multitude of naming conventions, and deviations from these conventions in human-annotated datasets, often only a subset of the input identifiers can be converted into a standard format.

Importantly, since gene and protein annotation databases are continuously updated, fast automatic access to public databases is required to obtain the best conversion results. For this purpose, the ArrayMining framework is connected to different mirror web-servers of the ENSEMBL database [28], and always attempts to retrieve the most current annotation data for an analysis. Moreover, the framework is interlinked with the DAVID functional annotation web-service and database [29] and enables the user to forward a list of genes/proteins, obtained from an analysis, to DAVID for further functional annotation analysis or gene name conversions. If these external data repositories are temporarily inaccessible, the software will attempt to convert the identifiers using a locally stored mapping database. These mappings are less comprehensive and not up-to-date, but provide reasonable results for the interpretation of common queries (e.g. for most Affymetrix gene expression platforms, a majority of the genetic probes can be converted to a standard identifiers). Importantly, genes which could not be annotated will still be taken into account in a standard feature selection analysis, but if they are selected, they appear with the annotation “unknown” in the ranking list.

To prevent false positive annotations, literature mining based synonym identification has not been employed here, because even for the most common model organisms, the accuracy of corresponding methods does normally not exceed 80% (see section 3.1.4 in chapter 3).

## 5.2 Ensemble and Consensus Analysis of Microarray Data

Both the results from microarray analysis studies in the literature (see chapter 3) and the comparative analyses conducted for this doctoral project (see chapter 4) have shown that new insights and improvements in terms of robustness and accuracy can be obtained by comparing different methods and combining them into ensemble techniques.

This section will present the methodology behind the new ensemble and consensus techniques for feature selection, clustering and classification, that have been integrated into ArrayMining and made accessible via a unified web-interface.

The main new biological findings obtained with this tool set (see chapter 8), and other novelties of the framework, including the new analysis pipelines connecting different analysis modules (see section 5.4), and novel cross-domain analysis methods (see chapter 6) will be described in other dedicated sections.

Figure 5.2 shows ArrayMining's main interface and the six analysis types it covers, including three general machine learning modules (feature selection, clustering, prediction) and three modules for analysis tasks specific to microarray data (cross-study integration, gene set analysis and network analysis). In the following, first the generic machine learning modules will be discussed, focussing on the ensemble and consensus techniques, and then the modules tailored to more specific biology-oriented analysis types.

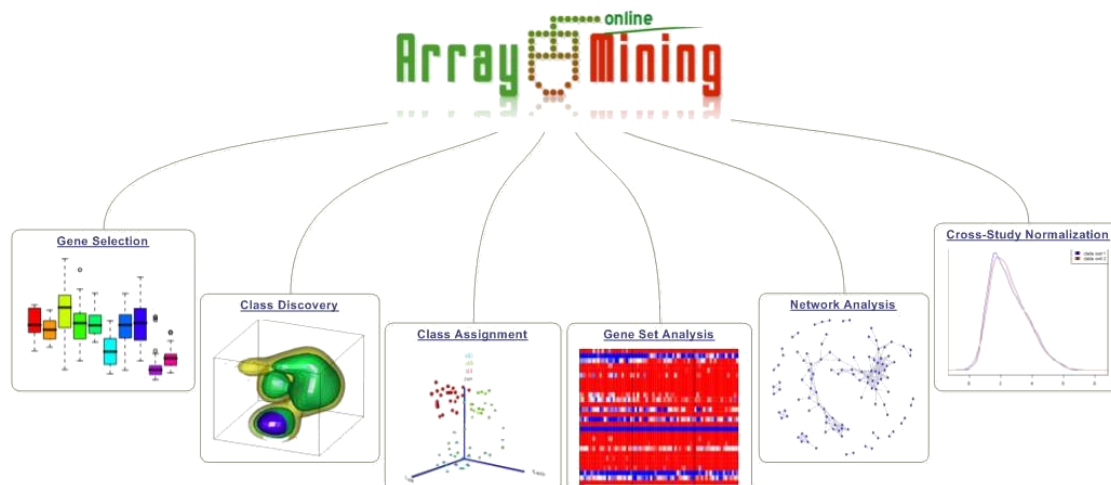


Figure 5.2: The main web-interface of the ArrayMining system for microarray data analysis, covering six different analysis types.

### Feature selection module

The gene/feature selection module enables the comparison and combination of a diverse choice of methods for identifying differentially expressed genes, proteins or metabolites from a high-throughput dataset with sample labels and replicates. These methods include the empirical Bayes t-statistic (*eBayes*) [150, 151], the Significance Analysis in Microarrays method (*SAM*) [419], correlation-based feature selection (*CFS*) [160], Random Forest embedded feature selection (*RF-MDA*) [191] and a Partial-Least-Squares based filter (*PLS-CV*) [356] using the weight vectors defining the first latent components in cross-validated PLS models (see also the literature survey in chapter 3 for details on most of these methods). Moreover, to exploit the synergies of different algorithms, a method to compute aggregated gene ranks from the sum of ranks of the individual methods was implemented (*ENSEMBLE* [18]).

The main result generated by the web-application is a ranked list of genes, in which known gene identifiers become clickable navigation items, referring the user to related entries in functional annotation databases and literature search engines. Additionally, box plots and heat maps (see examples in figure 5.3 and 5.4) visualise the expression values of top-ranked genes across different sample groups. If the supplied data uses common gene identifiers (Entrez Gene ID, NCBI GI accession, RefSeq Genomic ID, etc.), the list of selected genes can be forwarded to external analysis tools, e.g. the DAVID functional annotation clustering service [29].



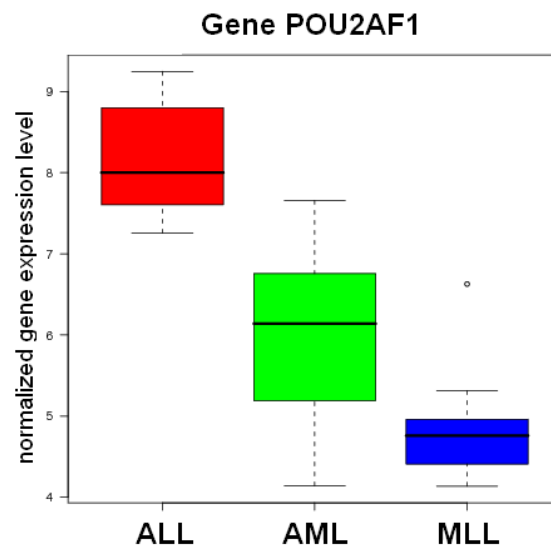


Figure 5.3: Box plots: Example of a box plot illustrating the spread of a gene's expression values across three classes of leukaemia samples: Acute Lymphoblastic leukaemia (ALL), Acute Myeloid leukaemia (AML) and Mixed Lineage leukaemia (MLL); data set by Armstrong *et al.* [420]

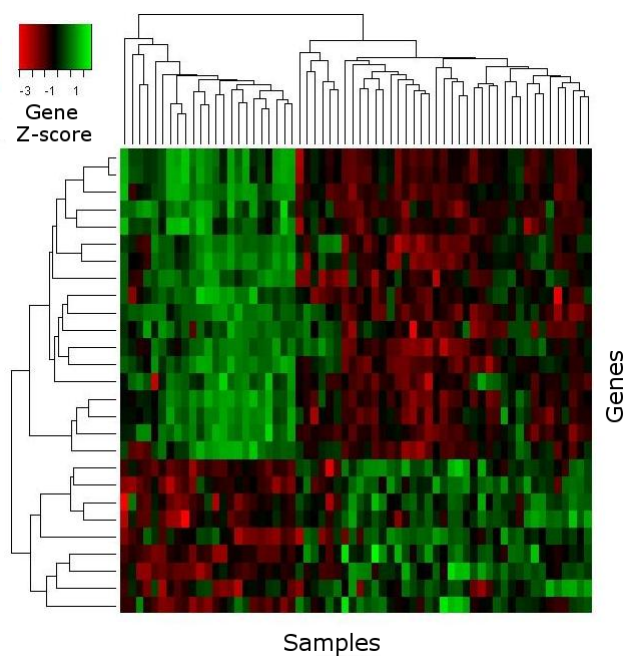


Figure 5.4: Heat map: Example of a heat map visualising the expression values of selected genes (rows) across samples (columns); data set by Armstrong *et al.* [420]

### Class discovery module

The class discovery module is designed to account for the great variety of existing scoring and search space exploration methods for microarray sample clustering, by including both partition-based and hierarchical clustering algorithms, an evaluation using multiple validity indices and a consensus clustering method. Currently, the partition-based clustering methods available are *k-Means*, Partitioning around medoids (*PAM*) [239] and Self-organising maps (*SOM*) [209]; and the hierarchical clustering methods include *Average Linkage Agglomerative Clustering*, *Divisive Analysis Clustering* and a combination between the agglomerative and divisive approach, *Hybrid Hierarchical Clustering* [421] (the Self-organising tree algorithm (*SOTA*) [422] is additionally included as a hybrid between *SOM* and hierarchical clustering). To combine the information content from multiple clusterings into a single representative solution, a self-devised consensus clustering approach was implemented, which maximises a score for the agreement between sample-pair assignments of the tentative solution and all input clusterings using a simulated annealing (SA) approach. This SA approach is a variant of the classical SA algorithm by Kirkpatrick [423] and uses an exponential cooling scheme and a stochastic random number generator creating Cauchy-distributed numbers [424]. The score for each cluster of size  $s_j$  is computed according to Swift *et al.* [12] using the so-called agreement matrix  $A$ , which counts the number of times the cluster assignments for two samples  $i$  and  $j$  agree across all input clusterings (see also section 3.4 in chapter 3). Specifically, the fitness score is obtained as a sum over entries of the upper triangle of  $A$  that correspond to sample-pairs occurring in the tentative cluster  $G_i$ :

$$f(G_i) = \begin{cases} \sum_{j=1}^{s_j-1} \sum_{k=j+1}^{s_j} (A_{G_{ij}G_{ik}} - \beta), & s_j > 1 \\ 0, & \text{otherwise} \end{cases} \quad (5.1)$$

where  $G_{ij}$  is the  $j$ -th element of cluster  $i$  and  $\beta$  is a user-defined trade-off parameter with a value between  $\min(A)$  (i.e. the score promotes the assignment of samples to a single, large cluster) and  $\max(A)$  (i.e. the score promotes the assignment of samples into different, small clusters). The total score for a clustering is the sum of scores for each single cluster, and has to be maximised. Simulated Annealing algorithms have been shown to work well for optimising this scoring function, hence, by testing different variants of Kirkpatrick's Simulated Annealing (SA) algorithm as part of this doctoral project (comparing linear, exponential and geometric cooling schedules, and thermodynamic SA [425] and adaptive SA [426]), the *Fast Simulated Annealing* approach by Szu *et al.* [424] was found to provide the best scores for a given runtime across multiple datasets.

Apart from providing access to single-algorithm and consensus clustering methods, the clustering module also estimates the number of clusters automatically by means of multiple validity indices. In particular, robust estimates of this number can be obtained by combining all pairs of algorithms and validity indices (method "ALL" on the web-interface). In the results page for a submitted clustering task, the user is provided with the output of three different methods to aggregate the validity scores for different numbers of clusters using all input algorithms and all validity indices: *A majority vote ranking* (i.e. the number of clusters is the one that most frequently obtained the best ranking score across all combinations of clustering algorithms and validity methods), *a median score ranking* (i.e. the estimated number of clusters is the median of the vector of best solutions across all method combinations) and *a sum of ranks aggregation* (i.e. the chosen number of clusters is obtained by summing up all ranks the different numbers of clusters received across all method combinations). Since there might be multiple meaningful clustering results, the user will also be

informed about the next best solutions according to these rankings.

Prior to an analysis, the user can optionally apply different types of data *standardisation* and *dimensionality reduction*. Apart from the classical standardisation to mean zero and standard deviation one, a more robust method using the median absolute deviation [397] can alternatively be applied to pre-process the data. The dimensionality reduction methods include the automatic, parameter-free COVSUM approach [145], which can distinguish between uncorrelated, uninformative genes and regulators with high correlation to other genes, a sparse PCA based filtering approach [146], and a classical variance-based filter, removing genes with low variance across all samples (see also section 3.4 in chapter 3). Moreover, as an alternative filtering approach, the user can first upload the input data on the gene set analysis module (see description of this module below) to extract “meta-genes” representing cellular processes, complexes and pathways, and then forward this data to the class discovery module.

As a result for each analysis, the user will obtain a tabular summary of the calculated validity indices and clustering results and various graphical outputs including a silhouette-plot [227], a 2D principal components plot and 3D VRML visualisations (see example in figure 5.2) generated with the self-devised visualisation software package *VRMLgen* [24] for different dimensionality reduction methods including *Principal Component Analysis* (PCA), *Independent Component Analysis* (ICA), *Locally Linear Embedding* (LLE) and *Isomap* (see comparative evaluation in chapter 4).

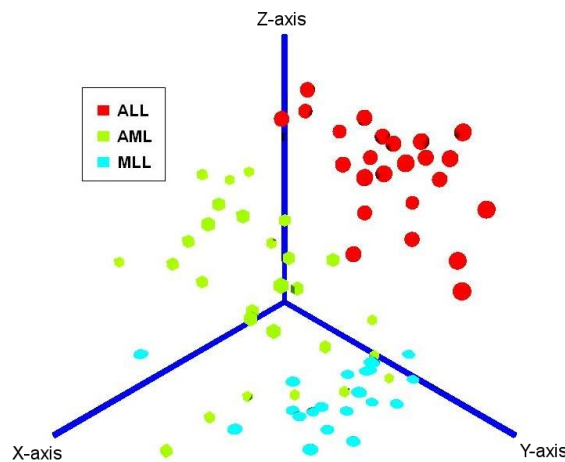


Figure 5.5: Independent Component Analysis: Example of a VRML-visualisation for an Independent Component Analysis; data set by Armstrong *et al.* [420]. The three sample types correspond to the leukaemia subtypes *Acute Lymphoblastic Leukaemia* (ALL), *Acute Myeloid leukaemia* (AML) and ALL with *mixed-lineage leukaemia gene translocation* (MLL).

### Class Assignment module

One of the main driving forces behind supervised microarray analysis is the goal to improve the diagnosis of diseases with genetic components by classifying the disease type for new samples using labelled training data. The third module on the ArrayMining web server is therefore dedicated to microarray sample classification, providing access to popular machine learning methods like *SVM* [377], *RF* [191], *PAM* [239] and *kNN*. Additionally, an in-house developed rule-based machine learning approach, BioHEL [366–369], is made available to the user (see the detailed description of BioHEL in section 4.2 in chapter 4). BioHEL had previously been shown to achieve high prediction accuracies on other complex biological data sets [427],

providing additional benefits in terms of the high interpretability of the decision rules in its prediction models.

All classification methods can be combined with any of the attribute selection methods from the feature selection module. The user has the choice of evaluating a chosen algorithm combination on a dataset either using the widely accepted external two-level cross-validation methodology [25], an automatic parameter optimisation within a nested cross-validation (see the “automatic data processing” section above) or a user-defined training/test set partition of the data. Moreover, since prediction models derived from training data of a single study can typically not be applied to samples from other platforms and laboratories, the combination of the cross-study normalisation module (see corresponding description below) with the class assignment module provides a means to obtain more general models derived from a larger number of samples.

The results for an analysis contain several performance measures for the evaluation and comparison of prediction methods, including the mean accuracy and standard deviation, Cohen’s Kappa score [428], the sensitivity and specificity, as well as a classification p-value according to Huberty *et al.* [429]. Moreover, Z-scores are computed for the genes that were most frequently selected across different cross-validation cycles [430]. To obtain more insights on these genes, similar analysis plots and links to external database entries are available as on the feature selection module.

## 5.3 Specialised Analysis Methods for Microarray Data

### Gene Set Analysis module

Two common problems in microarray analysis are high noise levels for single genes and a high number of redundant and uninformative genes. Using gene set analysis (GSA) to aggregate functionally related genes into gene sets and summarising their expression values to a robust “meta”-gene expression vector is a promising approach to overcome some of these limitations [318] (see chapter 3, section 3.7). Moreover, differentially expressed gene sets can provide insights on the differences between the biological conditions of the samples on the level of cellular processes, pathways and protein complexes represented by these gene sets.

On the ArrayMining server, the GSA module provides access to three annotation databases to extract sets of functionally related genes from a dataset (Gene Ontology [30], KEGG [31], and a collection of 37 cancer-related gene sets from the van Andel Institute in Michigan [310]). Alternatively, users can specify their own gene sets by entering the corresponding gene identifiers into a text box on the web-interface. Since common non-parametric GSA methods are often computationally expensive or provide only rough estimates of a gene’s significance score, the p-values are computed using the parametric PAGE approach [310], requiring a minimum gene set size of approx. 10 genes. To adjust these scores for multiple testing, the Benjamini-Hochberg method [3] is used. Finally, to summarise the information for a gene set into a single *meta-gene* expression vector, a dimensionality reduction method is applied to the original matrix of expression levels (using either the first component from a principal components analysis (PC-GSA) or the first dimension from a multidimensional scaling (MDS-GSA)).

The final results obtained on this module are presented as a list of gene sets, ranked according to their enrichment p-value score, with additional box plot and heat map visualisations, similar to those on the

gene selection module. Moreover, meta-gene expression values obtained from the GSA module can be downloaded or forwarded to other analysis modules, e.g. to use them as predictors for sample classification.

### Network Analysis module

As discussed in chapter 2, many diseases with genetic components are believed to be driven or influenced by a complex network of interacting biological molecules. Thus, the identification of modules of co-expressed genes or proteins can help to identify the cellular processes and their associated molecules, which are modulated in response to changes in the biological conditions of interest.

For this purpose, the ArrayMining framework contains a module for gene co-expression network analysis, which uses a single-parameter method to build a graph representing the co-expression relations between the genes in a dataset. Specifically, the graph nodes (representing genes) are connected by edges, if the corresponding genes are estimated to be significantly co-expressed, given the Pearson correlation of their gene expression vectors. As mentioned in the “Automatic data processing” section above, the user only needs to specify the edge adjacency threshold parameter, representing the trade-off between the reliability of identified co-expression relations and the coverage of all potential co-expression relations, and all other internal parameters are chosen automatically. For this purpose, the one-step automatic network construction approach for weighted gene co-expression networks by Zhang and Horvath [343] is employed.

The user can display a co-expression network in six different automatic graph layouts, using the Fruchterman-Rheingold method [431], Graphopt [432], DrL [433], the Kamada-Kawai layout [434], singular value decomposition (SVD) of the graph’s adjacency matrix, or a simple circular layout. In order to facilitate the interpretation of labelled data, the user can provide binary sample labels in combination with the microarray input data, which will be used to generate a coloured graph as output, where nodes representing genes which have a higher median expression level in the reference condition (class 1) in relation to the target condition (class 2) are coloured in blue, and genes with the opposite relation are coloured in red. Figure 5.3 shows an example graph visualisation created with the network module for the breast cancer dataset used in the comparative analysis in chapter 4 (employing the force-directed layout generation by Fruchterman and Reingold [431]).

Additionally, the network analysis module computes several topological statistics for the generated network, enabling the user to identify the gene with the maximum number of co-expressed neighbours (maximum degree node), the mean number of neighbours per node, and the global clustering coefficient, among others (see section 6.1.2 in chapter 6 for details on these and other topological descriptors).

To use the results for further analysis, a generated network can be downloaded in a variety of common graph/network file formats to export the data for visualisation in other software tools (e.g. Cytoscape [435]). Moreover, gene identifiers belonging to the same connected components are provided as a downloadable text file, enabling the user to analyse these gene sets with other external tools (e.g. functional annotation clustering using the DAVID web-service [29]).

### Cross-study normalisation module

Integrating data from multiple microarray studies can be an effective means to alleviate the common problem of small sample sizes in microarray data analysis (see also section 3.6 in chapter 3). To enable experimenters to benefit from the possibilities of cross-platform integration methods, five of these methods have

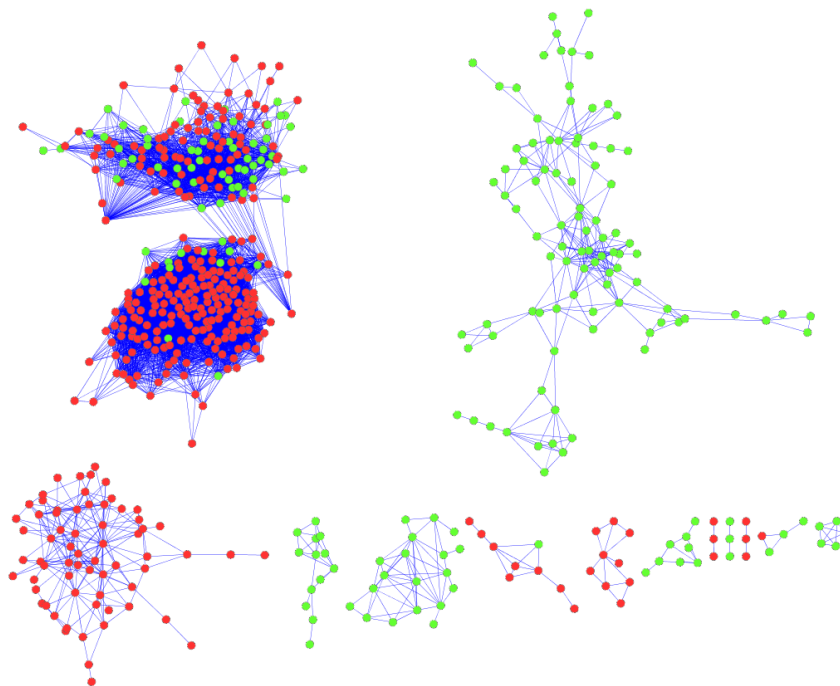


Figure 5.6: Example graph visualisation of gene co-expression network modules identified with the ArrayMining network analysis module on a breast cancer dataset with 128 samples [19] from two tumour subtype classes (luminal and non-luminal) using the force-directed layout generation by Fruchterman and Reingold [431]. Green nodes correspond to genes which are down-regulated in the luminal samples in relation to the non-luminal samples, whereas the opposite relation holds for the genes represented by red nodes. Accordingly, various modules of co-expressed genes exist as potential markers for both classes.

been made available in a dedicated module of the framework. These algorithms combine sample expression values from two different studies into a unified dataset using different strategies to adjust the distributions in the original input data. These include several of the approaches discussed in the literature survey in chapter 3, specifically, a linked gene- and sample-clustering approach (XPN [417]), an empirical Bayes method (EB [16]), a median rank score based method (MRANK [15]), an outlier-removing discretization technique (NorDi [436]) and a quantile discretization procedure (QDISC [15]). While the first three methods provide continuous-valued outputs, the last two use discretization to filter out noise, exploiting the fact that for many higher-level analysis tasks only a general categorisation of gene expression levels in different conditions is required (e.g. “unaltered”, “up”- or “down”-regulated). However, since the discretization of the data might not only remove noise but also biological information from the data, the user should choose the cross-study integration method depending on the analysis techniques to be applied subsequently, and ideally compare the results for both continuous and discretization-based approaches.

Importantly, although the input data sets can originate from different microarray platforms, the genes represented on these platforms need to have a significantly large overlap and the samples should be derived from the same tissue type under the same biological conditions. If these requirements are fulfilled, the module will generate density and quantile-quantile plots to assist the user in evaluating and comparing different algorithms (figure 5.3 shows example density plots before and after an XPN normalisation). If the user is satisfied with the outcome of the normalisation procedure, the combined data can be downloaded or forwarded to other modules for further analysis.

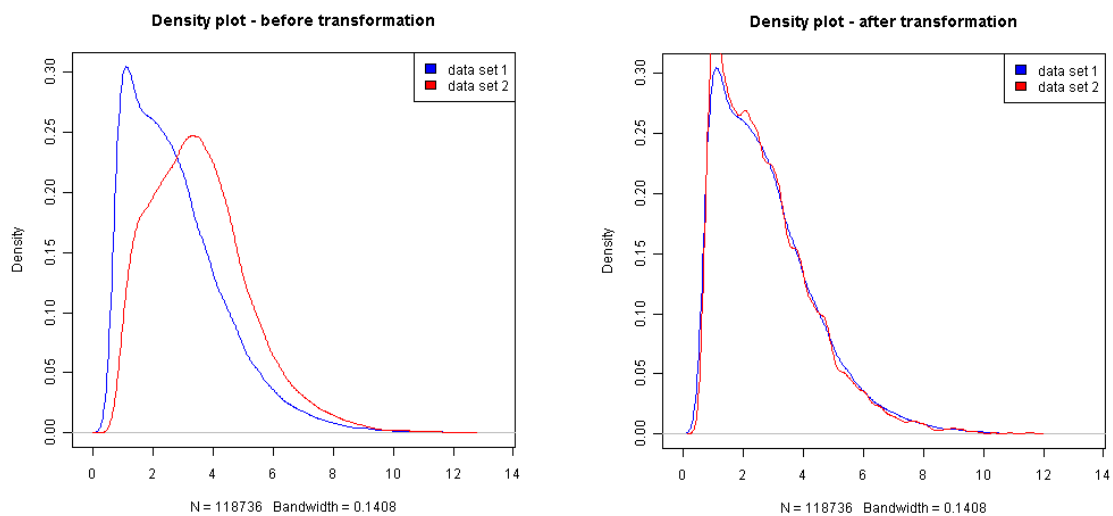


Figure 5.7: Example of kernel density estimation curves for microarray data before (left) and after (right) cross-study normalisation with the XPN method [417]. The normalisation procedure clearly leads to a better match between the two distribution curves.

## 5.4 Modular Combination of Analysis Techniques / Novel Analysis Pipelines

Single analysis types (clustering, gene set analysis, etc.) can only help to extract parts of the biological information within the measurements of a microarray experiment. However, by exploiting interconnections between *different analysis domains* and devising *new analysis pipelines*, novel insights and improvements in terms of robustness and accuracy can be gained using already existing algorithms. Moreover, if a wide selection of possible modular combinations of analysis techniques is made available via a simple interface, an experimenter can also set up a self-devised data analysis pipeline, tailored to a specific biological question, by freely choosing a sequential combination of methods.

ArrayMining enables such combinations of multiple analysis types (highlighted by arrows between the modules in figure 5) by a few mouse clicks, while ensuring that the user cannot use the methods in a manner that violates widely accepted validation guidelines.

For example, if the user wishes to combine a supervised feature selection with a classification algorithm within a cross-validation scheme, the ArrayMining interface will only allow the user to apply the feature selection within the cross-validation procedure (also known as *external cross-validation*, as opposed to a global feature selection using the whole data applied prior to the cross-validation). Thus, information leaks are prevented and more reliable estimates of the attainable classification accuracies on external test data can be obtained.

More novel and interesting combinations of analysis types result from interlinking the gene set analysis module with the classification, the clustering or the network analysis module. For example, the user can extract cancer-related gene sets or gene sets representing cellular pathways of interest from a dataset, summarise their expression values into meta-genes using a dimensionality reduction method (see the section on the Gene Set Analysis module), and then use these gene set fingerprints as predictors to train a machine learning algorithm for sample classification (see figure 11.2 in the appendix for another example pipeline).

This will not only provide more robust classification models, but also enable an interpretation of the data from a different perspective, e.g. revealing associations between pathway deregulations and different disease states in a dataset.

Another example application could be the alleviation of problems with small sample sizes and high levels of noise in input datasets from different studies. An effective means to increase the robustness of the analysis in this case would be to both increase the number of samples and reduce the number of features, by using the cross-study normalisation module to integrate two datasets from different platforms, in combination with the gene set analysis module to aggregate information from multiple functionally related genes into pathway fingerprints.

Interestingly, the combination of modules can also be helpful for a qualitative validation of the results from a single module. An example would be the independent application of the network analysis module on two separate datasets analysing the same cell type and biological conditions, compared to a network analysis of the combined data using the cross-study normalisation module. This comparison will enable the user to see whether similar modules of co-expressed genes in the individual datasets are also maintained in the unified dataset.

More importantly, ArrayMining is interlinked with other software tools and web-applications in the framework. Differentially expressed genes obtained from a feature selection analysis can be forwarded to the TopoGSA web-server for network topological analysis of gene sets [20], or the EnrichNet web-application for network-based gene set enrichment analysis (see chapter 6 for both of these algorithms). Moreover, a pre-filtered dataset can also be used as input for the Top-Scoring Pathway Pairs algorithm (see chapter 7).

Finally, ArrayMining also exploits synergies with external tools, allowing users to forward selected genes to the DAVID web-service [29] for a functional annotation clustering analysis, to load datasets from the GEO database [27] into ArrayMining's clustering module, or to inspect gene functional annotations in more detail using a variety of other web-based annotation services [28, 30].



## Chapter 6

# Integrative Analysis of Gene/Protein Sets

### Chapter abstract

Apart from gene and protein expression microarray analysis, discussed in the previous chapters, a multitude of other experimental methods exist to analyse the involvement of genes and proteins in a biological process of interest. However, many of these techniques do not produce data in a suitable format for machine learning purposes. For example, instead of containing numerical measurements, the data might just provide information on whether a gene is frequently mutated in a certain cancer type, silenced by a hypermethylated promotor, or inhibited by a repressor. Accordingly, the experimenter often just obtains lists with genes or proteins of interest as input for further analysis, instead of a dataset with numerical measurements for different genes/proteins across multiple samples.

Although various computational analysis techniques have been made available to investigate these gene/protein sets using additional annotation data, e.g. functional enrichment analysis methods, the possibilities to use existing interaction network, cellular pathway and literature data to investigate the experimentally derived gene/protein lists in more detail have not been exploited to their full potential.

This chapter therefore presents new analysis techniques dedicated to the general analysis of gene and protein lists from an arbitrary experiment. Although these methods are not designed to profit from additionally available numerical measurements, they are applicable to both data from high-throughput, noisy experiments and small-scale experiments like qPCR, with high sensitivity but low coverage. Moreover, these general analysis approaches can easily be interlinked in an analysis pipeline with more specific analysis methods dedicated to specific experimental platform, e.g. selected genes/proteins on the ArrayMining.net feature selection module (or from other external tools) can be forwarded to any of the analysis modules described below.

In the first part of this chapter, the software TopoGSA [20] for the analysis of network topological properties of gene/protein sets will be presented. Next, EnrichNet [21] is discussed, a web-application extending functional enrichment analysis by including information from interaction networks. Finally, for gene/protein sets representing cellular pathways or processes, a new

method to expand these pathway definitions using protein interaction data, PathExpand [22], will be explained in detail. The chapter uses material from the publications dedicated to each of these software tools.

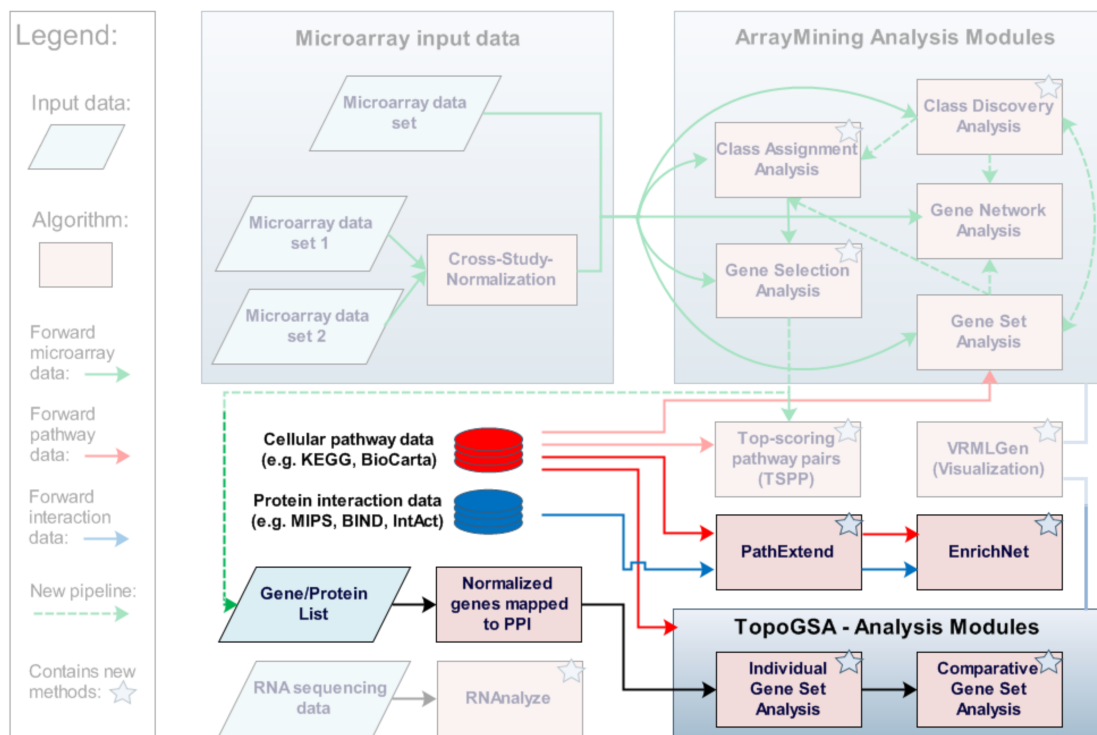


Figure 6.1: The new integrative analysis tools TopoGSA, EnrichNet and PathExpand as components of the integrative data analysis framework (highlighted by the colour contrast).

## 6.1 Network Topological Analysis of Gene/Protein Sets (TopoGSA)

### 6.1.1 Introduction and motivation

Functional genomic experiments provide researchers with a wealth of information delineating gene and protein sets of biological interest. To exploit these data sources, common steps in a functional gene/protein set analysis include the search for enrichment patterns [437], e.g. to identify significant signalling pathways or protein domains, as well as text-mining of the literature [438]. A further alternative approach for the functional interpretation of gene/protein sets is the analysis of molecular interactions in which the genes or their corresponding proteins are involved, in particular protein-protein interactions. In this context, various existing bioinformatics tools already allow users to map genes onto networks of interacting or functionally associated molecules to identify related genes and proteins [439, 440]. However, for the analysis and comparison of gene sets these tools have not taken into account topological properties in interaction networks so far.

This was the motivation for introducing *TopoGSA* (Topology-based Gene Set Analysis), a web-application to compute, visualise and compare network topological properties of gene or protein sets mapped onto in-

teraction networks. TopoGSA was developed as part of a collaboration with the Spanish National Cancer Centre (CNIO, Madrid), focussing on the analysis of cancer-related gene sets, in particular, gene sets known to be mutated in specific tumour types. TopoGSA maps these gene or protein sets onto a protein-protein interaction (PPI) network and computes different topological characteristics, such as the centrality of nodes in the network or their tendency to form clusters, and compares them to those of known cellular pathways and processes. This enables both the identification of genes with outstanding topological properties (e.g. as a post-processing procedure after a microarray feature selection analysis) and a ranking of known gene/protein sets with regard to their topological similarity to a target gene/protein set, which might point to previously undiscovered functional similarities.

### 6.1.2 Workflow and methods

**Analysis of network topological properties:** A network topological analysis on TopoGSA begins with the upload of a list of gene or protein identifiers (e.g. Ensembl IDs, HGNC symbols, among others). Alternatively, a microarray dataset can be used as input and differentially expressed genes will be extracted automatically using the feature selection module from ArrayMining [18] (see chapter 5 and the pipeline illustration in figure 11.2 in the appendix). Moreover, the user can add labels to the uploaded identifiers to compare different sub-sets of genes (e.g. “up-regulated” vs. “down-regulated” genes).

After submitting the list of identifiers, the application maps them onto an interaction network (see description in the *Implementation* section below), and computes topological properties for the entire network, the uploaded gene/protein set and matched-size random protein sets. Specifically, the considered network topological properties are:

- The *degree* of a node (gene or protein) is the average number of edges (interactions) incident to this node. Genes involved in many interactions are likely to have a vital functional role in the network, thus, a high average degree of a gene set can indicate that it contains such genes as members.
- The *local clustering coefficient* provides a measure of the tendency of nodes in a network to cluster together [441]. More formally, the measure quantifies the probability that the neighbours of each vertex are connected, by defining the local clustering coefficient  $C_i$  for a vertex  $v_i$  in an undirected graph  $G = (V,E)$  as:

$$C_i = \frac{2|e_{jk}|}{k_i(k_i - 1)} : v_j, v_k \in v_j : e_{ji} \in E \wedge e_{ij} \in E, e_{jk} \in E \quad (6.1)$$

where  $k_i$  is the degree of vertex  $v_i$  and  $e_{jk}$  is the edge between vertices  $v_j$  and  $v_k$ . Gene sets with a high average local clustering coefficient contain many genes in dense groups of nodes which could represent functional modules in the interaction network.

- The *shortest path length* (SPL) for two nodes  $v_i$  and  $v_j$  in an undirected, unweighted network is defined as the minimum number of edges which have to be traversed to reach  $v_j$  from  $v_i$ . Here, the SPL is used as a centrality measure, computing the average SPL from each node of interest to all other nodes in the network.
- The *node betweenness*  $B(v)$  of a node  $v$  is a centrality measure that can be calculated from the number of shortest paths  $\sigma_{st}$  from nodes  $s$  to  $t$  going through  $v$ :

$$B(v) = \sum_{s \neq v, s \neq t, v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (6.2)$$

Thus, the average node betweenness of a gene sets provides a measure of the network centrality of its corresponding genes with regard to their importance for the information flow across the network.

- The *eigenvector centrality* measures the importance of network nodes by applying a centrality definition, in which the score of each node reciprocally depends on the scores of its neighbours. More precisely, the centrality scores are given by the entries of the dominant eigenvector of the network adjacency matrix (see [442] for a detailed discussion of this property).

More details on network topological descriptors can be found in the book by Junker and Schreiber [345]. In order to visualise these topological properties for each individual gene/protein in an uploaded dataset, TopoGSA displays user-defined 2D and 3D representations in which the plotted data points are interlinked with corresponding entries in an online annotation database.

**Interaction network construction:** To generate a genome-scale interaction network, human protein-protein interactions were retrieved from five public databases. These include MIPS [93], DIP [94], MINT [95], HPRD [96] and IntAct [97]. Following the guidelines in the “Protein interaction data pre-processing” section in the literature review in chapter 3, only experimental methods dedicated to the identification of direct binary protein interactions were considered (see webpage [www.infobiotics.org/topogsa](http://www.infobiotics.org/topogsa), Datasets section). The final protein interaction network contained 9392 proteins (nodes) and 38857 interactions (edges).

**Comparison with known gene sets:** The analysis of network topological properties of only a single gene/protein set (“Individual Gene Set Analysis” module) does not lend itself to direct functional interpretation, although it facilitates the identification of genes with outstanding topological characteristics. However, a second analysis module on TopoGSA (“Comparative Gene Set Analysis”) additionally enables the user to compare the properties of a dataset of interest to a multitude of pre-defined gene/protein sets corresponding to known functional processes from public databases. For the human species, these include signalling pathways (KEGG [36], BioCarta [32]), Gene Ontology functional terms (i.e. *Biological Process*, *Molecular Function* and *Cellular Component* terms [30]) and InterPro protein domains [443]. Corresponding collections of datasets for other model organisms, including plants (*A. thaliana*), worms (*C. elegans*), fly (*D. melanogaster*) and yeast (*S. cerevisiae*), have also been made available.

When applying a comparative analysis using these data sources, summaries of network topological properties are provided for all gene/protein sets, and in the 2D and 3D plots different colours distinguish different datasets. Users can identify pathways and processes similar to the uploaded dataset visually, using these plots, or by inspecting a tabular ranking using a numerical score to quantify the similarity across all topological properties. This similarity score is obtained by computing five ranks for each pathway/process set according to the absolute differences between each of its five median topological properties and the corresponding value for the uploaded dataset. The sum of ranks across all topological properties is then computed and normalised to a range between 0 and 1. Accordingly, the smaller this value, the more similar the corresponding gene/protein set is to the uploaded dataset in terms of its topological properties.

### 6.1.3 TopoGSA - Example analysis

Since TopoGSA was built in co-operation with the Spanish National Cancer Institute (CNIO), an important motivation behind the tool was to use it for the analysis of cancer genes. Both the individual and comparative topological analysis was therefore applied to the complete set of genes currently known to be mutated in cancer [34].

Overall, the results confirmed previous observations according to which proteins encoded by genes which are known to be mutated in cancer have a higher average node degree in interaction networks than other proteins [444]. Moreover, the cancer genes were involved in more than twice as many interactions, on average, than matched-size random subsets of network nodes (with a difference of more than 15 standard deviations for 10 random simulations). Furthermore, the analysis with TopoGSA reveals that the cancer genes occur in closer proximity to each other (in terms of their average pairwise shortest path distances) than random gene sets of matched sizes and occupy more central positions in the interaction network (see Figure 6.2a) for details). In particular, the 3D plot displaying node betweenness, degree and shortest path length highlights the tumour suppressor p53's (TP53) outstanding network topological properties.

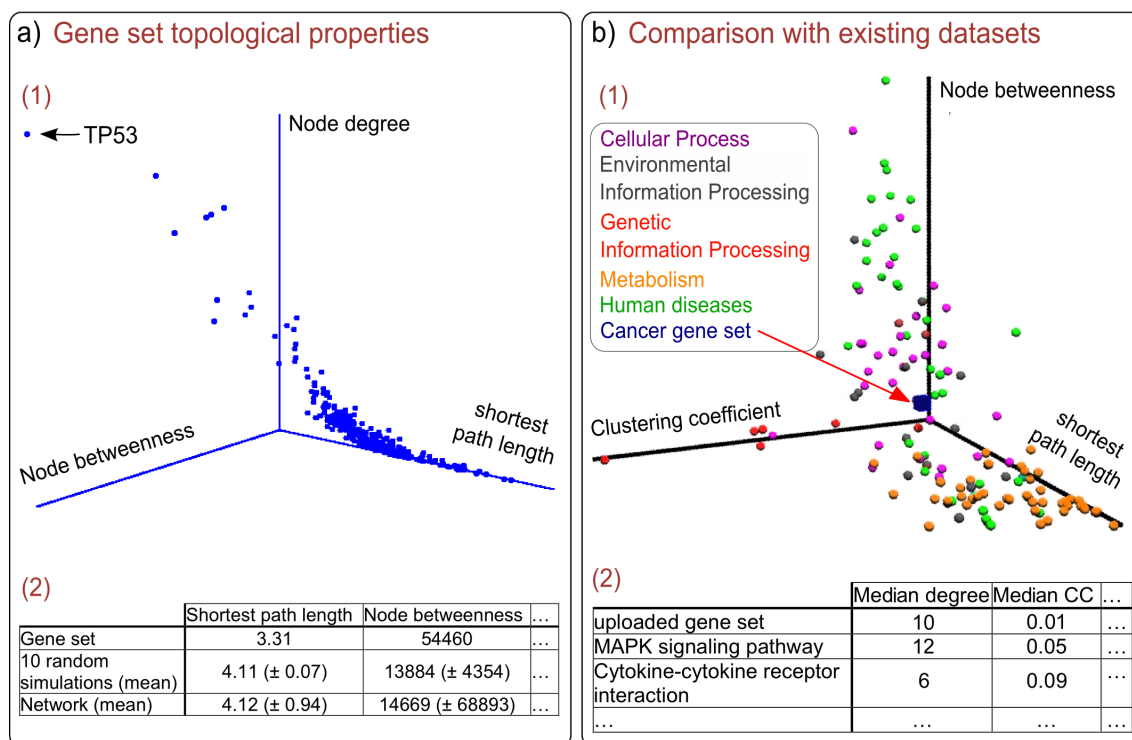


Figure 6.2: Example results generated with TopoGSA using the cancer gene set by Futreal *et al.* [34].

a) Topological properties can be computed and examined as visual (1) and tabular (2) outputs; b) The gene set can be compared with a chosen reference database (here the KEGG database).

When comparing the network topological properties of the cancer proteins with pathways from the KEGG database, representing each pathway by its corresponding set of genes (Figure 6.2b), the cancer proteins have similar network properties as several KEGG cellular processes and environmental information processing pathways (according to the KEGG-BRITE pathway hierarchy, [36], Figure 6.2b, purple and brown), whereas they clearly differ from metabolism related pathways (figure 6.2b, yellow). Interestingly, although the network topological properties of cancer genes are in agreement with their role in promoting cell division

and inhibiting cell death [445], they differ from those of most disease related KEGG pathways (Figure 6.2b, green), which tend to have higher degrees and network centralities.

### 6.1.4 TopoGSA - Implementation

The network analysis and gene mapping was implemented in the programming language R [353] and the web-interface in PHP. To build a human protein interaction network, experimental data from five public databases (MIPS [93], DIP [446], BIND [447], HPRD [448] and IntAct [97]) were combined and filtered for binary interactions by removing entries with PSI-MI codes for detection methods that cannot verify direct binary interactions (these are evidence codes for co-immunoprecipitation or co-localization, for example; the complete list of used method definitions and PSI-MI codes can be found in the “Datasets” section on the TopoGSA webpage, [www.infobiotics.net/topogsa](http://www.infobiotics.net/topogsa)). Additionally, protein interaction networks for the model organisms yeast (*S. cerevisiae*), fly (*D. melanogaster*), worm (*C. elegans*) and plant (*A. thaliana*) have been built using the same methodology as for the human network and using the BioGRID database [92] as an additional data source (see the help sections on the webpage for additional details on these networks). Importantly, users also have the option to upload their own interaction networks for a topological analysis, and instructions can be obtained from a demonstration video and manual available in the “Tutorial”-section on the webpage.

In summary, TopoGSA is a new web-application mapping gene and protein sets obtained from an experiment onto a pre-specified or user-defined molecular interaction network and providing insights on their topological properties and similarities to datasets representing known pathways, processes and complexes. All properties can be inspected and compared visually using 2D and 3D representations or ranked lists of the most similar pre-defined gene sets from cellular process databases.

## 6.2 Integrative Functional Enrichment Analysis of Gene and Protein Sets (EnrichNet)

### 6.2.1 EnrichNet - Introduction and motivation

Using molecular interaction networks in integrative data analysis approaches, gene and protein sets cannot only be analysed in terms of their topological properties (see previous section), but also by using the network distance information to assess their functional associations. In fact, the general task of assessing functional associations between an experimentally derived gene/protein set of interest and a database of known gene/protein sets is a common problem in the analysis of functional genomics data. A classical approach to address this problem is to apply an overlap-based enrichment analysis using the one-sided Fisher’s exact test. However, this method has various limitations:

- it can only score functional associations of overlapping gene sets
- some of the genes are not annotated for any pathways or processes and therefore disregarded
- the network structure of physical interactions between the gene/protein sets is not taken into account

Since publicly available large-scale molecular interaction data provides a source of information both for analysing the network structure between the molecules of interest and for inferring protein function at the

cellular scale, a network-based integrative analysis approach can help to address these problems. This was the motivation to develop *EnrichNet*, a web-application complementing overlap-based enrichment analysis by an association measure using the network structure of interactions between proteins.

### 6.2.2 EnrichNet - Workflow

Similar to TopoGSA, on the EnrichNet web-interface the user only needs to copy and paste a set of gene or protein identifiers into a text box, select a reference database (KEGG, BioCarta, Reactome, or Gene Ontology), and submit the analysis task. The EnrichNet approach will then compute network-based enrichment scores to estimate the functional association between the uploaded gene/protein set and the reference datasets, and if desired, visualise sub-graphs corresponding to the gene/protein sets of interest. This is achieved by the following four steps procedure:

1. The uploaded gene or protein set is mapped onto a molecular interaction network (a protein-protein, protein-DNA or genetic interaction network).
2. The distribution of distances between all protein/gene pairs from the dataset of interest and every reference dataset from the chosen database is computed (possible distance measures: shortest path length, random walk or kernel distance).
3. This distribution is compared against the distribution across all pathway/process protein sets from the chosen database using the Xd-distance [449] (see detailed explanation in the Methods section below).
4. For the chosen pair of gene/protein sets the corresponding network nodes and connecting edges are visualised and can be explored interactively.

In summary, the user obtains two types of outputs: A ranking of the reference datasets in terms of the network-based association with the uploaded gene set (measured by the Xd-distance), and additionally, an interactive visualisation for each entry in this ranking list, displaying the sub-networks of the interaction network corresponding to the uploaded and the reference gene set and highlighting their overlapping genes (nodes) and the interactions (edges) between the non-overlapping nodes. Thus, the user does not only obtain an estimate for the strength of the functional association between two gene sets, but can also investigate the interactions in detail which contribute to this association.

### 6.2.3 EnrichNet - Methods

To obtain a large-scale molecular interaction network as input data, experimentally verified, direct binary protein-protein interactions were assembled from the databases MIPS, DIP, MINT, HPRD and IntAct, and combined into a graph of 9392 nodes (proteins) and 38857 edges (interactions), which has also been used as part of the TopoGSA approach (see previous section). Distances between pairs of gene/protein sets mapped to this network were scored by comparing the distribution of pairwise distances between single nodes using shortest path distances (alternatively, random walk distances or different distance measure induced by kernel matrices could be applied) against a background distribution. This background distribution is obtained from the distances between the gene/protein set of interest and all pathway/process protein sets in a chosen database. In order to compare the distributions, a distance measure that has previously been used in the evaluation of protein contact map predictions, the Xd-distance [449], is employed. This measure is defined as follows:

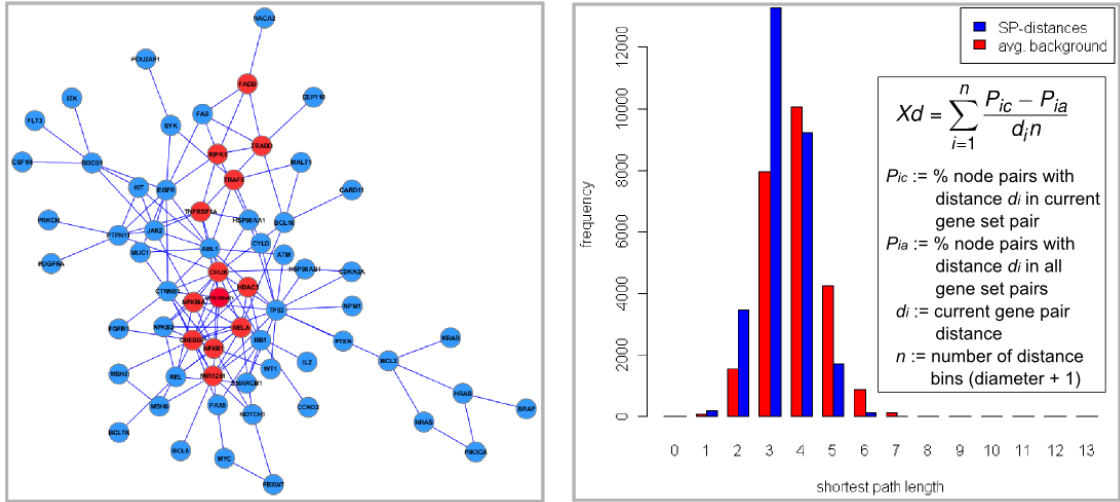


Figure 6.3: Left: Lymphoma mutated genes (blue) and genes/proteins from the BioCarta RE1A pathway (red) mapped onto a human protein-protein interaction network. Right: distribution of shortest path distances for two related gene sets (blue) and the background distribution across all gene set pairs (red).

$$X_d = \sum_{i=1}^n \frac{P_{ic} - P_{ia}}{d_i n} \in [-1, 1] \quad (6.3)$$

where  $P_{ic}$  is the percentage of node pairs with a distance  $d_i$  in the current gene set pair,  $P_{ia}$  is the corresponding percentage across all gene set pairs in the reference database,  $d_i$  is the shortest path distance between a pair of genes/nodes in the network, and  $n$  is the number of distance bins (equivalent to the network diameter + 1). Accordingly, this measure assigns higher weights to small distances and lower weights to large distances, ensuring that large distance outliers do not distort the results. Figure 6.3, right, shows an example distribution of shortest path distances for two closely related gene sets (blue) and the background distribution (red). The  $X_d$ -score, which is also defined in the same figure, rewards gene set pairs with a distance distribution that is shifted towards the left in relation to the background distribution, i.e. short distances (indicating strong functional associations) occur more often than expected, whereas long distances appear less often than expected. Thus, high  $X_d$ -scores point to strong functional associations between the corresponding gene sets and low scores to weak or non-existing functional associations.

Importantly, this definition of the  $X_d$ -score leads to a different interpretation of the ranking scores in comparison to classical p-value significance scores. While p-values lie in a range between 0 and 1, and findings with low p-values have a higher statistical significance than observations with high p-values, the  $X_d$ -score ranges from -1 to 1 and findings with higher scores are more significant than low-scoring results (more specifically, for  $X_d$ -scores larger than 0, the frequency of short network distances between the gene sets is shorter than expected based on the background model). In order to find an approximate cut-off  $X_d$ -score corresponding to a desired p-value significance threshold, the user can compute classical overlap-based significance scores using the Fisher exact test for the overlapping gene sets and fit a linear regression model to relate the overlap-based scores to the corresponding  $X_d$ -scores.

To test this approach on real-world data,  $X_d$ -distances were computed for sets of genes known to be mutated in different tumour types (bladder cancer, breast cancer and lymphoma) compared against pre-defined protein sets from KEGG, BioCarta and Gene Ontology corresponding to cellular pathways and processes.



As an example, figure 6.3 (left) shows the network structure for two non-overlapping gene sets (lymphoma mutated genes (blue), BioCarta RelA pathway (red); Xd-score: 0.2). A large number of edge connections between the genes from the two gene sets suggests a strong functional association between them, which would not have been discovered using an overlap-based enrichment analysis, since the overlap is zero.

### 6.2.4 EnrichNet - Results

Similar to TopoGSA, the EnrichNet methodology and web-application was developed in co-operation with the Spanish National Cancer Institute, and the main biological goal was again to analyse cancer-associated gene sets. Specifically, cancer mutated gene sets were mapped to a large-scale human protein interaction network and their shortest path distances to mapped gene/protein sets from KEGG, BioCarta and Gene Ontology were scored with the Xd-distance measure. These scores were then compared against classical overlap-based enrichment scores using the Fisher exact test, for all cases in which the gene/protein sets had non-zero overlaps.

In agreement with prior expectations, high Pearson correlations between the network-based and the overlap-based significance scores were observed (absolute correlations above 0.75, considering only datasets with non-zero overlap). More importantly, new functional associations were identified for gene/protein sets with small or no overlap. For example, figure 6.4a) displays the network structure obtained when comparing the sarcoma mutated gene set (40 mapped genes) against the BioCarta cell cycle pathway “RacCycD” (26 mapped genes), related to a cell’s transition from the so-called *G1* phase (growth phase) to the *S* phase (DNA synthesis phase).

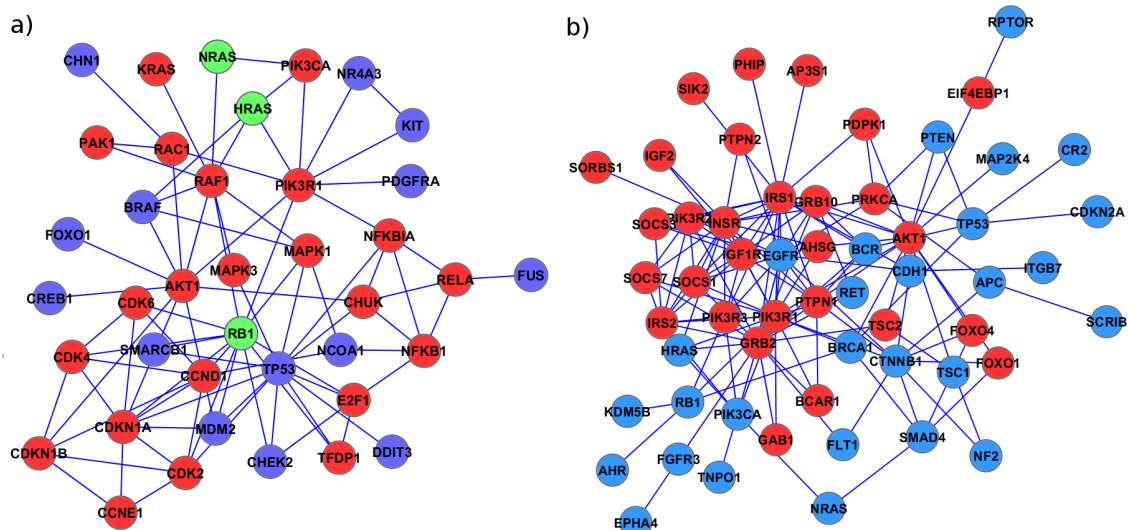


Figure 6.4: Protein-protein interaction sub-network for: a) Sarcoma mutated genes (blue) and genes/proteins from the BioCarta RacCycD pathway (red); b) Bladder cancer mutated genes (blue) and genes/proteins from Gene Ontology term “Insulin receptor signaling pathway” (red, GO:0008286).

These datasets have an overlap of only three genes (HRAS, NRAS and RB1) and would therefore not have been considered as significantly associated by an overlap-based enrichment analysis using the Fisher exact test (q-value of 0.17), whereas the obtained Xd-score is comparatively high (0.18, please note the different interpretation of the Xd-score in comparison to classical p-values described in the Methods section) and suggests a strong association (which is supported by the multitude of interactions between the corresponding

proteins for these gene sets, see figure 6.4a)).

A further example for an identified functional association is shown in figure 6.4b), which reveals strong interconnections between the non-overlapping set of bladder mutated genes (blue) and the dataset representing the Gene Ontology “Insulin receptor signalling pathway” (red, GO:0008286, Xd-score: 0.27). While this association between cancer and insulin receptor signalling is well described in the literature [450,451], there are only few studies reporting an association between lymphoma tumours and the RelA pathway [452] (see figure 6.3, Xd-score: 0.2), whose network mappings are also strongly interconnected.

On the whole, EnrichNet successfully identifies all the functionally associated gene sets detected by an overlap-based enrichment analysis, but additionally finds new associations for non-overlapping gene sets, some of which have not yet been reported in the literature. The associations detected for the sarcoma, lymphoma and bladder cancer gene sets discussed above and similar results for other cancer mutated gene sets suggest that the EnrichNet approach provides an effective means to identify and prioritize previously unknown functional associations between gene and protein sets.

Furthermore, in addition to the ranking of gene set associations using the Xd-score, the user can visualise all network-based associations online and inspect them in further detail using an interactive interface. Specifically, this interface enables the user to zoom into the sub-networks of interest, search for specific genes and highlight them in the network, and view their topological properties and functional annotations by clicking on the corresponding nodes. In contrast to a simple ranking table provided by standard overlap-based enrichment analysis methods, this exploratory data analysis allows the user to investigate the biology behind the ranking scores in more detail and obtain specific insights on the functional and topological properties for the molecules of interest and the network of interactions between them.

### 6.2.5 EnrichNet - Summary and conclusions

EnrichNet is an integrative analysis method for gene and protein sets, addressing limitations of classical overlap-based enrichment analysis by introducing the following new features:

1. Putative gene/protein set associations can be identified and prioritised even if the input data sets have only small or no overlaps.
2. A visual analysis of corresponding sub-networks reveals the structure of molecular interactions and potential *linker proteins* (i.e. proteins involved in the communication between different cellular pathways). Moreover, interactive features like mouse wheel zoom-in and node context menus with network topological information and annotation data for genes/proteins enable an in-depth exploration of the regions of interest in the interaction network.
3. The approach enables the combination of experimental evidence from multiple data sources, different interaction networks (protein-protein, genetic interactions, etc.) and functional genomics experiments.

EnrichNet represents a natural extension of classical overlap-based enrichment analysis, since the classical distinction between overlapping and non-overlapping genes can be understood as a special case of the EnrichNet approach (using a binary distance measure that considers only shortest path distances of zero (overlapping) or greater than zero (non-overlapping)).

Due to the availability of large-scale molecular interaction data for a wide range of species and even specific tissue types, EnrichNet is applicable to a great variety of data analysis tasks in the biosciences, including problems in basic biological research, biotechnology and biomedicine.

## 6.3 Integrative Methods to Extend Gene/Protein Set Based Cellular Pathway Definitions (PathExpand)

### 6.3.1 PathExpand - Introduction, background and motivation

Both in the TopoGSA and the EnrichNet analysis approach, gene/protein sets obtained from an experiment are compared against known gene/protein sets representing cellular pathways, processes and complexes. So far, these “known datasets” have been considered as fixed and correct representations of the corresponding biological processes. However, although these expert-based cellular pathway representations provide a rich source of information, the pathway definitions are partly subjective and inconsistent across the different databases. For example, when comparing the pathway diagrams of the cancer-related *p53 signalling pathway* in the databases KEGG [36], BioCarta [32], Wikipathways [453] and Invitrogen iPath [454], they do not only differ in form and layout but in the molecules involved and the connections between them. In fact, the assignment of a protein to a pathway often relies on the experimental procedure and on a subjective assessment of the protein’s importance for the process. Many associated regulators, effectors or targets of core cellular pathways may therefore have been overlooked or mistakenly not been considered as relevant enough by classical approaches to define pathways.

In addition to the mentioned inconsistencies, classical representations are limited to portraying pathways as independent cascades of proteins transmitting a signal from the cell surface to the nucleus. However, recent functional genomics high-throughput initiatives have identified a large number of interaction partners for signalling proteins, suggesting more complex relationships between cellular pathways than in their traditional representations [455], and challenging the classical view of pathways as independent functional entities. Therefore, an integrative approach to extend cellular pathway definitions, combining information from the original pathway databases and molecular interaction networks and using objective criteria to score pathway definitions, could be an opportunity to create more consistent and informative pathway and process definitions.

For this reason, a new methodology for extending pre-defined protein sets representing cellular pathways and processes, *PathExpand*, was developed as part of this doctoral project. PathExpand amalgamates the information from process and pathway databases with large-scale protein-protein interaction data. Previous approaches for *in silico* generation of cellular processes using molecular interaction data have constructed pathways from scratch [456–459], and related approaches for disease candidate gene prioritisation also rely on interaction network data [460–462]. However, an extension approach which preserves the information content in existing process definitions, but expands these definitions by identifying new strongly associated biomolecules, has previously not been investigated.

This was the motivation behind PathExpand, which maps original pathway definitions from public databases onto a protein-protein interaction network, and extends them to include their most densely interconnected interaction partners (using various graph-theoretic criteria). Both the added proteins and the extended pathway definitions can be used for a wide range of practical analysis tasks, to gain new biological insights

related to the cellular processes represented by these pathways.

In the following sections, the methodology behind PathExpand and statistical and biological results obtained on real-world data will be presented and discussed in detail. Specifically, the analysis of proteins added to pathway definitions by the method reveals that these proteins display distinctive network topological features and molecular function annotations and can be proposed as putative new components of the corresponding cellular processes, and/or as regulators of the communication between different processes. This is illustrated by the prediction of novel Alzheimer's disease candidate genes and the identification of proteins with potential involvement in the crosstalk between several interleukin signalling pathways.

As with the TopoGSA and EnrichNet tool, PathExpand resulted from a collaboration with the Spanish National Cancer Institute (CNIO), and the prime target was therefore to investigate pathways whose deregulation may contribute to the development of cancers [445]. Thus, extended cellular pathways and processes were also used to analyse their enrichment in pancreatic mutated genes from a large-scale resequencing study.

### 6.3.2 PathExpand - Methods

**Implementation:** All data processing and analysis steps in PathExpand were implemented in the programming language R, and the web-interface on [www.infobiotics.net/pathexpand](http://www.infobiotics.net/pathexpand) was developed in PHP. The human protein-protein interaction network used for all network-based computations is the same as the one used for TopoGSA and EnrichNet (see above).

Gene/protein sets corresponding to cellular pathways/processes were extracted from the public databases KEGG [31], BioCarta [32] and Reactome [37] and then mapped onto the protein interaction network. Since the interaction data does not represent the entire proteome, on average about 60% of the pathway proteins could be mapped onto the network.

**Process extension procedure:** Original cellular pathways/processes containing a minimum size of 10 protein members were used as seeds and mapped onto the interaction network. The direct neighbours of these seed nodes were then considered as candidates for the extension procedure and filtered according to multiple graph-theoretic criteria to assess the strength of their association with the pathway nodes. More specifically, in the first filtering step, a candidate node  $v$  has to fulfil condition (6.4) below and at least one of the following conditions (6.7-6.7) to be added to a pathway  $p$  (an illustration of these conditions is shown in figure 6.5).

**node degree:**

$$\text{degree}(v) > 1 \quad (6.4)$$

**direct pathway/process association:**

$$\frac{\text{process\_links}(v, p)}{\text{outside\_links}(v, p)} > T_1 \quad (6.5)$$

**indirect pathway/process association:**

$$\frac{\text{triangle\_links}(v, p)}{\text{possible\_triangles}(v, p)} > T_2 \quad (6.6)$$

**pathway/process node coverage:**

$$\frac{\text{process\_links}(v, p)}{\text{process\_nodes}(p)} > T_3 \quad (6.7)$$

where  $\text{degree}(v)$  is the number of neighbours of node  $v$ ,  $\text{process\_links}(v, p)$  is the number of direct links from  $v$  to a node in process  $p$  and  $\text{outside\_links}(v, p)$  is the number of direct links from  $v$  to a node outside of process  $p$ . In equation 6.6,  $\text{triangle\_links}(v, p)$  is the number of triangles in which  $v$  occurs together with a node in  $p$  and another candidate node, and  $\text{possible\_triangles}(v, p)$  is the number of these triangles which could potentially be formed, if all other candidate nodes would be part of a triangle connecting  $v$  and  $p$ . The thresholds  $T_1$ ,  $T_2$  and  $T_3$  are defined here as  $T_1 = 1.0$ ,  $T_2 = 0.1$  and  $T_3 = 0.3$  (this selection provided a reasonable trade-off between the number of extended pathways and the average size of the extension). For  $T_1 = 1.0$ , equation 6.5 corresponds to a well-known condition in graph theory introduced to define *strong communities* in networks (stating that the number of connections to the pathway/community must exceed the number of connections to the rest of the graph, see [320]). Given that a candidate node can have connections with all the original pathway nodes, the threshold  $T_3$  always has to be smaller than 1 (i.e. the maximum pathway node coverage is 1).

Since the extension procedure should ideally also provide more compact pathway representations in the network, this first candidate protein filter is complemented by applying a second filter to the candidate nodes passing the first. Specifically, a candidate node is only accepted, if the following *compactness score* for a pathway protein set  $P$ , given by the mean of the shortest path lengths between all pairs of proteins belonging to  $P$ , is reduced after adding the candidate:

$$\text{compact\_score}(P) = \frac{\sum_{i, j \in P; i < j} \text{dist}(P_i, P_j)}{|P| * (|P| - 1) / 2} \quad (6.8)$$

Thus, the filtering criteria ensure that proteins added to a pathway are both strongly associated with the original pathway members and provide an extended pathway with a compact network representation. Moreover, in particular the added proteins which increase the compactness by connecting disconnected proteins in the original pathway can be expected to have a very strong association with the original pathway. Finally, to ensure that the extension procedure is deterministic, the order in which proteins are added to a pathway is given by a greedy strategy, i.e. the protein that increases the compactness the most is always added first.

**Topological network analysis:** To quantify local and global topological properties of proteins in the network, the web-application TopoGSA [20] (see above) was used to compute five topological descriptors: the number of connections to other nodes (degree), the tendency of nodes to form clusters (clustering coefficient), their centrality in the network (betweenness and eigenvector centrality) and the distances between them (shortest path length). See the section on TopoGSA above for a detailed explanation of these topological characteristics.

**Cross-validation:** In order to validate the expansion procedure, the extent to which randomly deleted proteins in the original pathways/processes can be recovered by the proposed method is analysed using the following cross-validation strategy:

1. 10% of the proteins from each pathway were removed randomly among those proteins that are connected to at least one other protein in the pathway. If the set of proteins that are connected to other pathway members covers less than 10% of the total number of proteins, we iteratively remove random

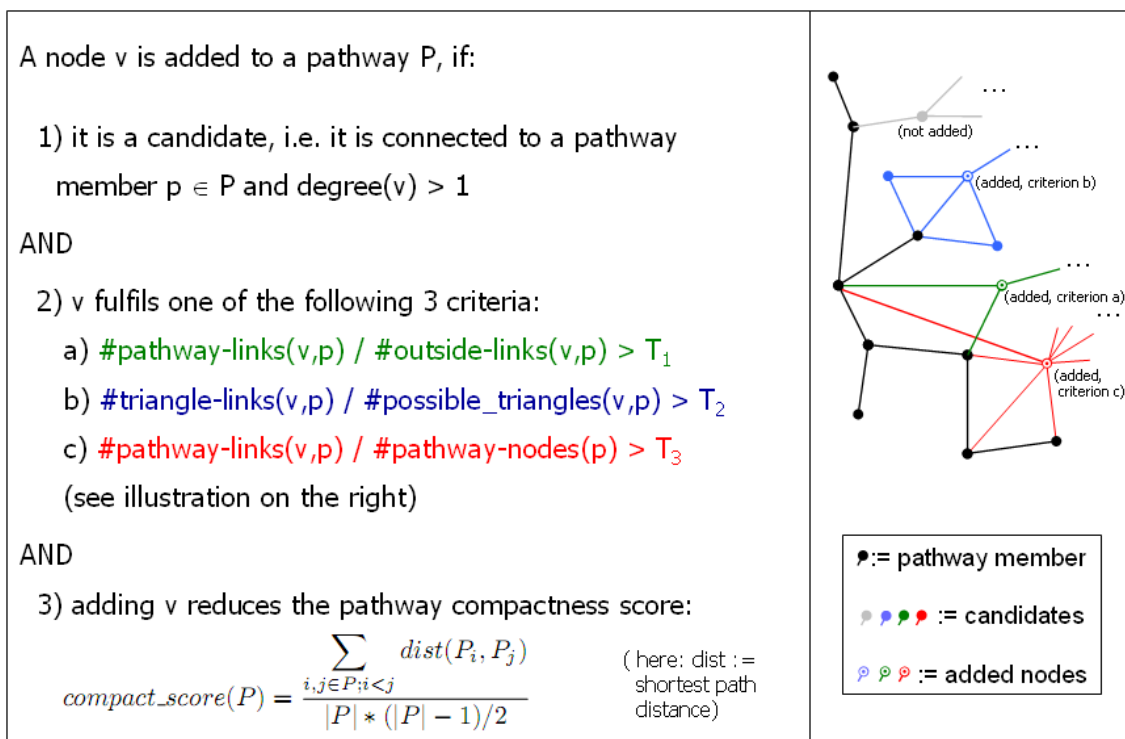


Figure 6.5: Filtering criteria: Visualisation of graph-based filtering criteria used to extend the cellular processes (the process nodes are shown in black, coloured and circled nodes represent cases in which different filtering criteria are fulfilled by a candidate node).

proteins from this set and recompute the set until it is empty.

2. To each reduced pathway the proposed extension procedure was applied as well as 100 alternative random extensions, computed by sampling randomly the same number of proteins from the *candidate* proteins of the reduced pathway (see definition of *candidates* in the previous section).
3. P-value significance scores are estimated as the relative frequency of cases where more proteins were correctly recovered by a random extension than by the proposed extension procedure across all pathways in a database.

**Semantic similarity analysis of Gene Ontology terms:** For further validation, the functional similarity between the added proteins and the original pathway members was assessed against the same background model used in the previous cross-validation analysis. For this purpose, pairwise similarities between protein annotations were quantified with Jiang and Conrath's semantic similarity measure for GO terms [103]. Using this similarity score, the average GO-term similarities between all pairwise combinations of GO biological process (BP) terms from the original proteins in the cellular pathway and the added proteins was computed. The random extension model was again created by randomly sampling the same number of proteins from the candidates for the pathway (see definition of candidates in the pathway extension section) as in the real extension, excluding the proteins from the extended cellular pathway under consideration. Importantly, it is not possible to compare the extensions of real pathways to extensions of random gene/protein sets with similar connectivity in the network, because in most cases these sets would largely overlap with other pathways.

**Enrichment Analyses:**

- The enrichment of molecular functions among the proteins added to the cellular pathways/processes by the extension procedure was tested for all databases independently using the DAVID functional annotation clustering tool [393] (Gene Ontology Molecular Functions and InterPro protein domains), with the proteins from the interaction network. Functional annotation clusters with a more than 2-fold enrichment were selected and manually labelled.
- To estimate the probability of observing certain overlaps between extended or original protein sets representing pathways and other protein sets of interest, e.g. cancer-related proteins, we employed a classical over-representation analysis (ORA) using the one-tailed Fisher exact test. To adjust for multiple testing, we employ the approach by Benjamini and Hochberg [3].

### 6.3.3 PathExpand - Results and Discussion

In the following section, the results obtained by applying the pathway extension approach to cellular pathway and process datasets from the databases KEGG [31], BioCarta [32] and Reactome [37] will be discussed. Across all databases, 1859 different processes were considered (with a minimum size of 10 proteins) and mapped onto a network containing 38857 interactions (see Methods).

**General extension statistics:** The proposed procedure could extend 159 pathways from BioCarta, 90 from KEGG and 52 from Reactome (see table 6.1 and [www.infobiotics.net/pathexpand](http://www.infobiotics.net/pathexpand)). The pathway sizes increased on average from 113% to 126% of the original size.

Table 6.1: Statistics on added proteins across different databases

Property	BioCarta	KEGG	Reactome
no. of examined pathways	322	199	79
no. of extended pathways	195	140	62
avg. pathway size	19	49	75
avg. size after extension	24	61	85
total no. of added proteins	935	1745	622
no. of unique added proteins	280	623	409
Molecular function categories of proteins added by the extension method (2-fold enrichment, see methods)	Phosphatase activity, Regulator activity, Binding, Kinase inhibitor/regulator, Cytokine binding /TNF receptor	Phosphatase activity, Regulator activity, Cytokine binding /TNF receptor	Regulator activity

Statistics on the number of pathways that could be extended, the average extension size, the number of added (unique) proteins and their molecular function categories.

**Network properties of added proteins:** The added proteins in the interaction network had a more than one standard deviation higher node degree, betweenness and average local clustering coefficient (Methods) than 10 matched size random protein sets [20] (see table 6.2). Moreover, the shortest path lengths between the added proteins were smaller by several standard deviations (table 6.2). This tendency of the proteins added by the extension method to occur in more central and dense regions of the network is consistent with similar trends observed for the topological properties of proteins from the original cellular pathways and processes (see table 6.2).

**Functional annotations of the proteins added to the cellular pathways/processes:** A semantic similarity

Table 6.2: Topological properties of BioCarta pathway/process extensions [32]

Property	Proposed extension: Added proteins only (mean)	Random model: Added proteins only (mean / stddev.)	Original cellular processes (mean / stddev.)	All network proteins (mean/stddev.)
Shortest path length	3.68	4.11(0.03)	3.77 (0.51)	4.12(0.94)
Node betweenness	21998	14545(4751)	49888 (153173)	14669(68893)
Degree	10.3	8.11(0.94)	21.53 (32.64)	8.27(16.2)
Clustering coefficient	0.34	0.11(0.01)	0.12 (0.17)	0.11(0.21)
Eigenvector centrality	0.04	0.01(0.04)	0.05 (0.09)	0(0.57)

Comparison of different numerical topological properties for the proteins added by the proposed extension method (column 1) or the random model (column 2), as well as a comparison of these properties for the nodes corresponding to the original cellular processes (column 3) and the entire protein-protein interaction network (column 4).

analysis of the GO terms was used to compare the functional annotations of the original cellular process proteins with the annotations of the proteins added during the extension procedure (see Methods). For almost all cellular pathways, the GO terms of the added proteins are more similar to the GO terms of the original cellular pathway proteins than those of matched-size random protein sets (see figure 6.6). These results confirm that the added proteins belong to similar functional categories as the proteins from the cellular processes they were assigned to. Furthermore, a functional enrichment analysis of the combined set of proteins added to all cellular processes (applied to each database separately) reveals an enrichment in proteins annotated for regulatory activity (see table 6.1). More interestingly, for the databases KEGG and BioCarta, the added proteins are enriched in phosphatases. This result could indicate that phosphatases, which might correspond to negative regulators, have previously been overlooked in the definition of canonical pathways.

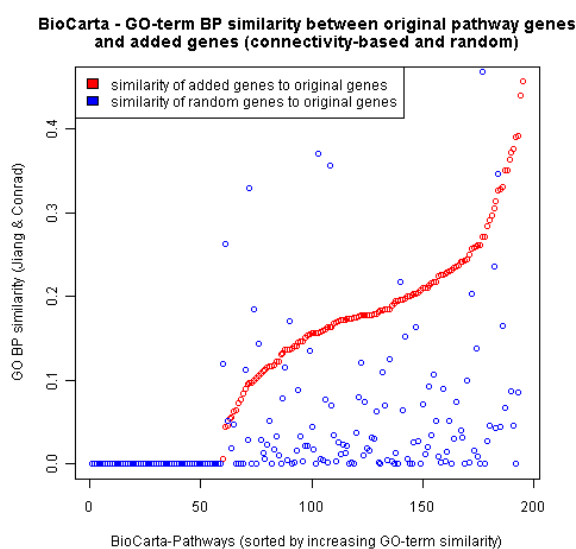


Figure 6.6: Semantic similarity analysis: Similarities in Gene Ontology Biological Process terms between original BioCarta pathway proteins and added proteins (red) and between original pathway proteins and matched-size random protein sets (blue)

### The extension procedure can recover known pathway proteins after deletion

A cross-validation procedure (Methods) showed that the cellular pathway extension recovers a significantly larger number of randomly deleted pathway-nodes in the network than a simplistic extension using a random



selection among the candidate nodes (p-values < 0.01 for all databases). Specifically, the distribution of the number of recoveries across the 100 random model extensions never provided a higher number of recoveries than the proposed extension method.

### Prediction of new pathway/process components

From the following observations we conclude that the proteins added to pathways by the extension procedure can be proposed as new candidate components with a functional role in the corresponding cellular processes:

1. The proteins added by the proposed method are well connected with the original pathway nodes and central in the protein interaction network.
2. The added proteins display gene ontology annotations matching better to the original cellular pathway/process annotations than random proteins, and are enriched in processes known to be related to cellular signalling.
3. The method is able to recover known cellular pathway/process proteins in a cross-validation experiment.

To illustrate the utility of the extension procedure for the prediction of new components, an expert-based pathway diagram modelling the process likely to be deregulated by the most penetrant Alzheimer's susceptibility genes (created using information from the literature [463] and available in the KEGG database [31]) was analysed as example case. The proposed extension method added five different proteins to this cellular map (see figure 6.7, the interactive visualisation of the extension is available online: [www.infobiotics.net/pathexpand](http://www.infobiotics.net/pathexpand)).

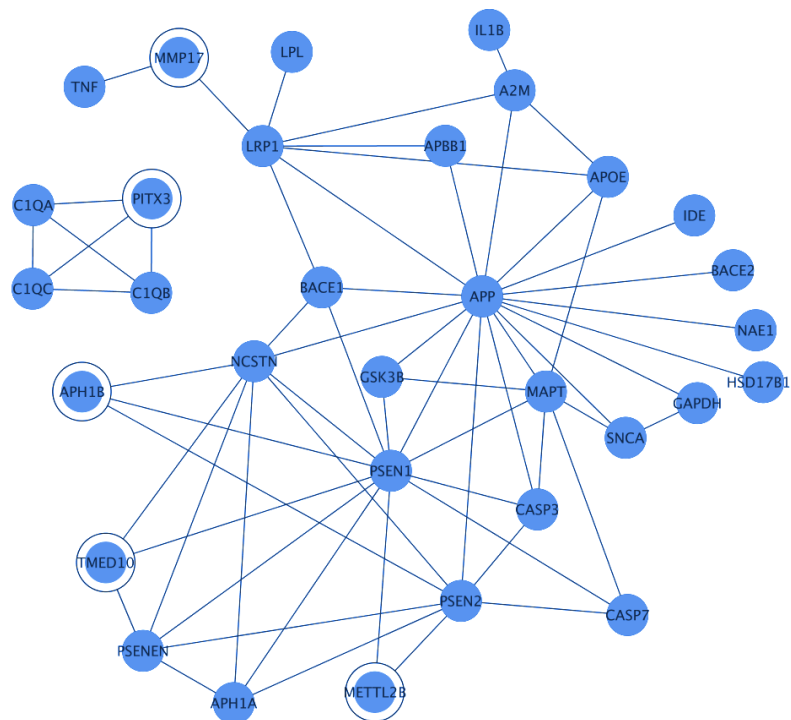


Figure 6.7: Extended KEGG Alzheimer's pathway: Nodes with surrounding circles represent added proteins

Interestingly, three of them have previously been implied in Alzheimer's disease (the proteins TMED10,

APH1B and PITX3). Two other proteins, METTL2B and MMP17, which are also added to the Alzheimer's cellular map by the proposed method, have not been linked to the disease so far. MMP17 is a member of the metalloproteinase protein family involved in the breakdown of the extracellular matrix. According to the *Huge navigator* [464], six other members of this protein family have been associated with the Alzheimer's disease. The other candidate is a methyltransferase-like, METTL2B. Another member of this family, MMETL10 has been associated with Alzheimer's disease in a case-control study [464]. Thus, using the Alzheimer's disease pathway as a first test case of the extension method, MMP17 and METTL2B can be proposed as new candidate disease genes. Recent successes in bioinformatics-based identification of combined biomarkers [465–467] and screening for new drugs against Alzheimer's [468] have already shown, that clinically relevant insights on the disease can be gained by applying new computational biology approaches to existing datasets.

#### **The extension of cellular processes points to extensive inter-pathway communication**

The involvement of some proteins in multiple processes suggests that extensive communication occurs between different cellular processes. Indeed, before applying the extension procedure, about 50% of the cellular process proteins are annotated for more than one cellular process. Interestingly, after the extension procedure, the percentage of unique proteins among all proteins added to the cellular processes ranged from 30% (BioCarta) to 66% (Reactome), revealing that many proteins are added to more than one cellular process. In agreement with the observations for the original process proteins, again about 50% of the added proteins belong to more than one cellular process. Accordingly, many proteins in the protein interaction network are well connected with different cellular processes, and might therefore be expected to have a functional role in the communication between them.

As an example for these types of connections, the class of interleukins (ILs) was considered. ILs are secreted proteins mainly involved in the immune system to regulate the communication between immune cells. They activate different signalling pathways, which can share intracellular signalling cascades (e.g. MAPK, RAS or STAT), but which also display distinct properties (e.g. by binding to different receptors). In this context, some IL-pathway proteins are annotated only for one IL pathway (see figure 6.8, each colour corresponds to an IL pathway), while other proteins occur in multiple pathways (figure 6.8, multiple colour node proteins). Furthermore, all the IL pathways share protein interactions (figure 6.8, blue links). Thus, the analysis of protein interactions between the members of different IL pathways highlights the complexity of this signalling system.

The pathway extension method was applied to the seven interleukin signalling pathways depicted in figure 6.8, and added between 1 to 10 proteins to each pathway. As the figure reveals, some proteins were added to only a single IL pathway. For instance, the CTAG1B (cancer/testis antigen 1B) protein was only added to the IL5-signalling pathway (figure 6.8, green proteins). Interestingly, the added protein is an antigen expressed only in cancer cells and in normal testis cells, and could represent a regulatory member of this pathway in these two particular conditions. Moreover, four other proteins were jointly added to more than one IL pathway. Three of them extend the IL2, IL3 and IL6 pathways, which are all activating the STAT and Ras/MAPK signalling cascades. These proteins are known regulators of these cascades and can also participate in the regulation of the communication between the different interleukin signalling pathways.

#### **Functional enrichment of tumour mutated genes in extended pathways**

Large-scale tumour resequencing projects have revealed a large number of genes mutated in different cancer types [469–471]. To understand the biological significance of these mutated genes, those cellular processes

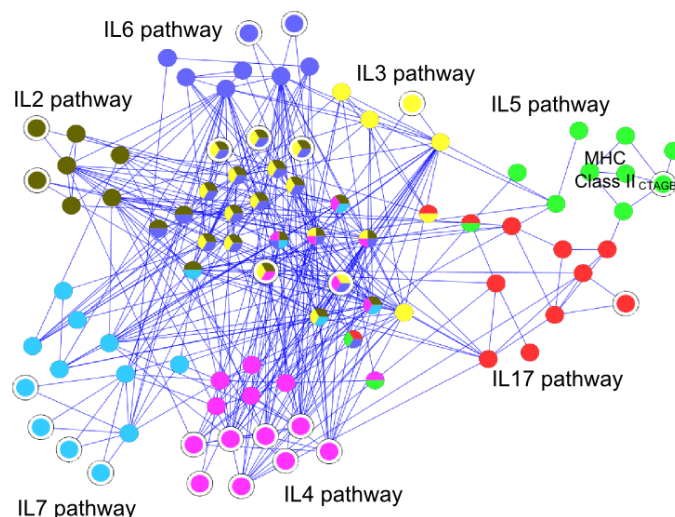


Figure 6.8: Crosstalk between interleukin signalling pathways: Protein interaction sub-network containing the proteins annotated for 7 different Interleukin (IL)-related pathways from the BioCarta database (each colour represents a pathway, nodes for proteins annotated for multiple pathways have more than one colour). Proteins added by the extension method are highlighted by surrounding circles and coloured according to the pathway(s) they were added to (they appear mostly within peripheral clusters or as links between process members). They were not annotated for any of the IL-related pathways before applying the extension procedure, and the original pathway members did not become members in further IL-related pathways. Therefore, to simplify interpretation and provide a compact data representation, the node colours are only used to visualise the pathway memberships after the application of the extension procedure.

containing more mutated genes than expected by chance have been identified previously (see for example [470]).

Here, an enrichment analysis was applied on cancer mutated genes extracted from a pancreatic large-scale resequencing study [470], using extended cellular processes from BioCarta, KEGG and Reactome. This enabled the identification of significant associations between different cancer types and the extended pathways (see also Methods section for details on the methodology). Interestingly, 8 out of 12 core signalling pathways whose association with pancreatic cancer had previously been identified, were retrieved as significantly associated with this disease [470]. An over-representation analysis (ORA) shows that some cellular pathways and processes are more significantly enriched in mutated genes in the extended versions than in the original versions (see table 6.3). These include signalling pathways, like MAPK, p38 MAPK, p53, Wnt, PDGF, FC epsilon receptor I, ErbB or functions like apoptosis and cell cycle G1/S check point (table 6.3). Interestingly, some of the proteins added to these processes by the extension procedure are also pancreatic mutated genes (see last column in table 6.3). These proteins include, for instance, the BCL2-related protein A1, which is added to the Apoptosis Reactome pathway and indeed known to be involved in apoptosis. A less obvious example is the dual specificity phosphatase 19 (DUSP19), a phosphatase added by the extension procedure to different MAPK pathways, the Fc epsilon receptor I signalling pathway and to a pathway known to be activated in response to HIV Nef protein (negative effector of Fas and TNF). This protein is highly expressed in the pancreas [472] and displays a frameshift mutation in pancreatic tumours [470].

Finally, new insights can be gained when analysing the BioCarta cell cycle G1/S check point process (see figure 6.9). This process contains several proteins that were found mutated in large-scale pancreatic resequencing studies (figure 6.9, red nodes), as well as many other proteins known to be involved in cancerogenesis.

The proposed extension procedure adds seven proteins to this process (figure 6.9, circled nodes). All of these

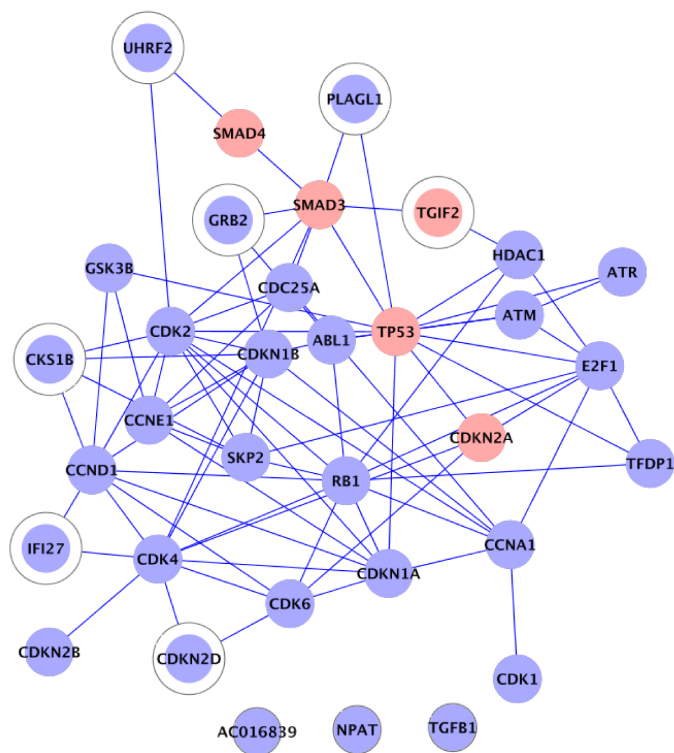


Figure 6.9: Cell cycle G1/S check point sub-network: Protein-protein interaction sub-network corresponding to the proteins annotated for the BioCarta pathway “Cell cycle G1/S check point” and proteins added by the proposed extension procedure (circled). Proteins whose corresponding genes have been found mutated in pancreatic whole-genome resequencing studies [470] are highlighted in red.

proteins are either transcription factors, kinases or other signal transduction regulators, and six of them are known to be involved in cell cycle regulation (all except TGIF2). Interestingly, the cancer resequencing study showed that the TGIF2 gene is mutated in a pancreatic tumour (figure 6.9, circled red node). This transcriptional repressor gene has also been reported to be amplified in some ovarian cancers, and can be recruited by TGF- $\beta$ -activated SMADs [473] (SMADs are intracellular proteins that forward extracellular growth factor signals to the cell nucleus). Accordingly, both the involvement of the corresponding TGIF2 protein in the cell cycle G1/S check point process, and its involvement in cancer through the deregulation of this process can be predicted from these observations.

In conclusion, the network-based extensions of the cell cycle G1/S and other processes provide useful explanatory information for the cancer association of these pathways/processes by adding new regulators that increase the connectivity between cancer mutated genes and other process members in the interaction network. For instance, in the G1/S process, SMAD3 is connected to other process members by adding the proteins TGIF2, GRB2 and PLAGL1; and SMAD4 is connected to the process member CDK2 by adding UHRF2. Thus, the overall coherence of the processes is increased and an expanded view of the influence of different cancer genes in these processes is obtained.

#### 6.3.4 PathExpand - Conclusions

The extension of known cellular pathways and processes with densely interconnected interaction partners in a protein-protein interaction network can provide bioscientific and biomedical researchers with several new insights. Specifically, the example analyses conducted in this study have shown that the extensions lead to the proposal of new putative components and to the identification of mediators of the communication between the processes. These results do not only help to better understand the cellular processes and their interrelations, but the novel extended pathways can also be used as input for bioinformatics tools dedicated to the pathway-based analysis of new experimental functional genomics data (including other tools from the framework, see also the example analysis pipeline in figure 11.2 in the appendix). Thus, by taking into account canonical knowledge as well as large-scale interaction data, the extended pathways can be of direct practical use in a wide range of biological applications, illustrated by their utility in helping to explain the functions of cancer mutated genes.

Table 6.3: Cellular processes enriched in pancreatic mutated genes

Cellular Process database	Cellular process	ORA Q-value before/ after extension	Pathway size before/ after extension	Number of mutated genes in new pathway	Number of mutated genes among added genes	Mutated genes among added genes
Reactome	Hemostasis	0.475 / 5.18e-06	221 / 278	19	4	LRP1B, TFPI2 PON1, SIGLEC11
KEGG	Tight junction	1.48E-4 / 4.5e-05	106 / 126	14	3	RASIP1, RASGRP3, PLEKHG2
KEGG	MAPK signaling pathway	3.35E-4 / 4.87e-05	225 / 279	21	6	DOCK2, MAPKBP1, SLC9A5 RASIP1, DUSP19, PLEKHG2
KEGG	Cell adhesion molecules (CAMs)	2.87E-4 / 1.03E-4	109 / 116	12	2	TNR, SEC14L3
KEGG	Wnt signaling pathway	3.35E-4 / 1.39E-4	123 / 147	14	3	MAPKBP1, PLEKHG2, ANKRD6
KEGG	Neuroactive ligand-receptor interaction	3.35E-4 / 1.72E-4	198 / 217	17	3	EML1, ACE
BioCarta	MAPKinase Signaling Pathway	1.33E-3 / 2.89E-4	81 / 111	8	2	MAPKBP1, DUSP19
Reactome	Apoptosis	3.7E-2 / 4.42E-4	124 / 146	11	2	BCL2A1, RASGRP3
Reactome	Signaling by PDGF	5.72E-3 / 4.43E-4	61 / 121	10	3	VPS13A, LIG3 FMR2
BioCarta	Cell Cycle G1/S Check Point	1.7E-3 / 5.06E-4	27 / 34	5	1	TGIF2
BioCarta	Agrin Postsynaptic Differentiation	1.27E-2 / 8.21E-4	27 / 38	5	2	PGM5, PLEKHG2
BioCarta	p38 MAPK Signaling Pathway	3.25E-3 / 1.13E-3	34 / 42	5	1	PLEKHG2
BioCarta	ALK in cardiac myocytes	2.89E-3 / 1.25E-3	32 / 44	5	1	TBX5
KEGG	Fc epsilon RI signaling pathway	2.69E-2 / 2.71E-3	67 / 114	10	5	DOCK2,MAPKBP1, DUSP19,ATF2,RASGRP3
KEGG	ErbB signaling pathway	2.32E-2 / 3.52E-3	86 / 196	13	7	VPS13A,MAPKBP1,NEK8, LIG3,DUSP19,AFF2,GLTSCR1
KEGG	Regulation of actin cytoskeleton	4.94E-3 / 2.72E-3	184 / 236	15	4	RASIP1. CDC42BPA, PLEKHG2, CYFIP1
BioCarta	HIV-I Nef negative effector of Fas and TNF	7.88E-3 / 4.78E-3	50 / 66	5	1	DUSP19
KEGG	p53 signaling pathway	5.62E-3 / 5.44E-3	59 / 64	7	1	PPP2R4
Reactome	Signaling in Immune system	0.459 / 7.02E-3	228 / 266	12	1	SEC14L3

The complete list of cellular processes that display a statistically significant enrichment in pancreatic cancer mutated genes after applying the proposed extension method (Q-value < 0.01) and improved significance scores in relation to the original pathways (i.e. Q-values decreasing after the extension). The significance scores for the overrepresentation analysis (ORA) and the pathway sizes are shown before and after the extension, and the total number of mutated genes in the extended pathways is provided, as well as the size and the annotations for the set of mutated genes among those that were added to these pathways.

## Chapter 7

# Simplifying Classification Rules to Enhance Model Interpretability

### Chapter abstract

One of the major limitations of many analysis methods for high-dimensional, noisy data is the complexity of the resulting models, which impedes a clear interpretation and the direct extraction of new biological insights. In most biological applications however, the interpretation of computational models has a higher practical value than the exclusive use of a model for predictive tasks.

In particular, since most microarray sample classification models do not reach 100% accuracy on large external test sets, and statistical confidence score estimates are often subject to high variance, confidence in the biological meaningfulness of a model is often an essential requirement for the acceptance of the model by the scientific community. For example, to justify the selection of a specific treatment approach, clinical researchers need to understand from a biological point of view why a model assigns a certain patient sample to a specific diagnostic group. Moreover, the knowledge obtained from interpretable models can improve the understanding of the molecular basis behind a disease and help to design novel therapies and assays for disease monitoring and diagnosis using small sets of biomarkers.

This chapter will therefore present an integrative method to increase the interpretability of machine learning models for high-dimensional biological data analysis. This novel rule-based learning technique, the *Top-scoring pathway pair (TSPP)* algorithm [23], combines information from gene or protein expression data with cellular pathway definitions to learn simple “if-then-else” decision rules for sample classification. The first part of the chapter will introduce the TSPP method and examine the results of its application to microarray cancer datasets using material from the original TSPP publication. The final part will discuss how TSPP relates to other rule learning methods and provide an overall summary. Figure 7 shows how TSPP is integrated into the complete integrative data analysis framework for this doctoral project.

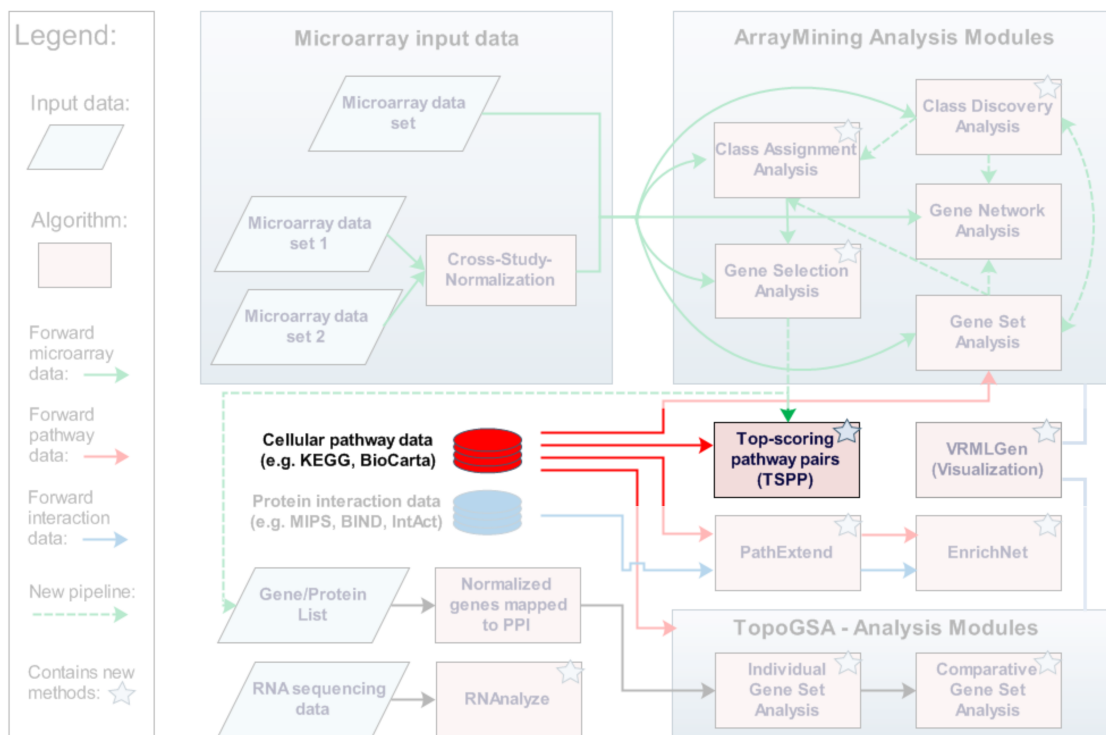


Figure 7.1: The Top-scoring pathway pairs (TSP) method as a component of the integrative data analysis framework (highlighted by the colour contrast).



## 7.1 Integrative Rule Learning for High-Dimensional Biological Data (TSPP)

### 7.1.1 TSPP - Background and motivation

As outlined in the literature survey in chapter 3, the supervised classification of microarray gene expression samples into different biological categories is often hampered by several limitations resulting from a high dimensionality of the data in relation to a small number of available samples and high noise levels.

In recent years, several novel machine learning methods have extended or replaced classical generic approaches, by providing more compact, robust and/or easily interpretable classification models. These approaches reduce the prediction model complexity and increase its robustness by using regularization and shrinkage techniques [238, 474], by generating more human-interpretable machine learning models, which consist of simple decision rules [367, 475], or by using more robust data representations and model formulations, e.g. computing rank score expression values [15] or only considering relative expression values by comparing pairs of genes [284, 285]. One of these methods, the ensemble rule learning method *BioHEL* [366–369], has already been discussed in detail in chapter 4, which focussed on the comparison of standard machine learning methods with ensemble and consensus approaches. The results obtained with BioHEL showed that high classification accuracies obtained with ensemble models do not necessarily come at the expense of model comprehensibility, since a post-analysis of BioHEL’s ensemble models does not only provide robust rankings of informative attributes in the data, but also information to prioritize putative oncogenes and tumour suppressors.

However, the techniques mentioned above and the algorithms presented in chapter 4 are not designed to exploit the information content from diverse biological data sources, including cellular pathway definitions and protein interaction data in addition to microarray expression level measurements. Moreover, these methods analyse microarray data on the level of single genes as attributes, and since the expression measurements for single genes are often very noisy, there is still ample room for improvement in terms of the cross-platform robustness of the corresponding machine learning models.

This motivated the development of the *Top-scoring pathway pairs* (TSPP) classification method, which addresses the problem of low model robustness due to noise and uncertainty in the data by combining ideas from the techniques mentioned above with an approach to analyse the data at the level of pairwise comparisons of cellular pathways, rather than at the single-gene or pairwise-gene level. Briefly, the genes (or proteins) in a microarray study are mapped onto cellular pathways and processes from public databases (e.g. KEGG, Gene Ontology, BioCarta and Reactome) and simple decision rules for sample classification are learned by comparing the expression levels in pairs of pathways against each other. Rules describing single pathway-pairs are then weighted and combined into a unified classification model by applying a boosting algorithm (see literature survey, chapter 3). For an additional post-analysis, the top-ranked pathway-pairs can be mapped onto a protein-protein interaction network for visual inspection or to analyse their functional association with the *EnrichNet* method (see chapter 6).

The proposed approach can be understood as a pathway-based extension of the “top-scoring pairs” (TSP) algorithm [284, 285] (see also the rule-based learning section in the literature review in chapter 3), which identifies discriminative pairs of genes in microarray data, and has therefore been named *Top-scoring pathway pairs* (TSPP). Moreover, the method draws inspiration from functional enrichment analysis approaches, summarising expression values for genes in cellular pathways and processes to “meta-gene” expression val-

ues (e.g. the methods GSEA [476], MaxMean [477] and the global test [478]), and pathway-based sample classification methods [318].

The TSPP procedure generates compact and comprehensible sets of rules, describing changes in the relative ranks of gene expression levels in pairs of cellular pathways across different biological conditions. The rules are robust against monotonic transformations of the data and can be computed efficiently using a simple algorithm.

In the following sections, the TSPP methodology will be explained in detail and example results for two real-world large-scale microarray studies (prostate cancer and B-cell lymphoma sample classification) will be shown. The example results show that the method provides robust rule sets with significant predictive information in relation to random model classification, as well as new insights on differentially regulated pathway pairs. However, the benefit of these predictive models in comparison to other classification methods like support vector machines lies not in the attained accuracy levels but in the ease of interpretation and the insights they provide on the relative regulation of cellular pathways in the biological conditions under consideration.

In summary, the TSPP approach is not designed to compete with existing microarray sample classification and data mining methods in terms of maximising accuracy, but to complement them with the following added benefits:

- New biological insights on the relative up- and down-regulation of cellular pathways in biological conditions of interest can be gained from easily interpretable decision rules.
- The prediction models are applicable to data from other microarray platforms without requiring that all platforms contain the same genetic probes and that cross-study normalisation is applied (the integration takes place at the level of pathways, and the gene expression values are replaced by rank scores).
- By summarising the expression values of multiple genes belonging to the same pathway, the dimensionality of the data is reduced (from about 50.000 genes to a few hundred pathways) and the summarised *pathway expression fingerprints* have a higher robustness than single gene expression vectors (however, at the expense of losing detail; therefore single-gene based methods should be applied additionally).

### 7.1.2 TSPP - Methods

In contrast to previous microarray machine learning methods comparing single gene expression values or summarised expression values for single pathways against fitted threshold values, TSPP provides increased robustness by both combining expression levels of multiple genes into pathway expression fingerprints and making pairwise, relative comparisons between pathways instead of using fitted thresholds. Specifically, the TSPP algorithm identifies, scores and combines decision rules using pathway pairs according to the following five-step procedure (an illustration of the workflow is also shown in figure 7.2):

#### 1. Rank score transformation:

A gene expression matrix  $X$  with dimension  $n \times p$  ( $n$ : number of samples,  $p$ : number of genes) and class labels  $y$  for the samples is read as input and transformed into a *rank matrix*  $R$  by sorting the expression values for each gene across the  $n$  samples and replacing them with their position index

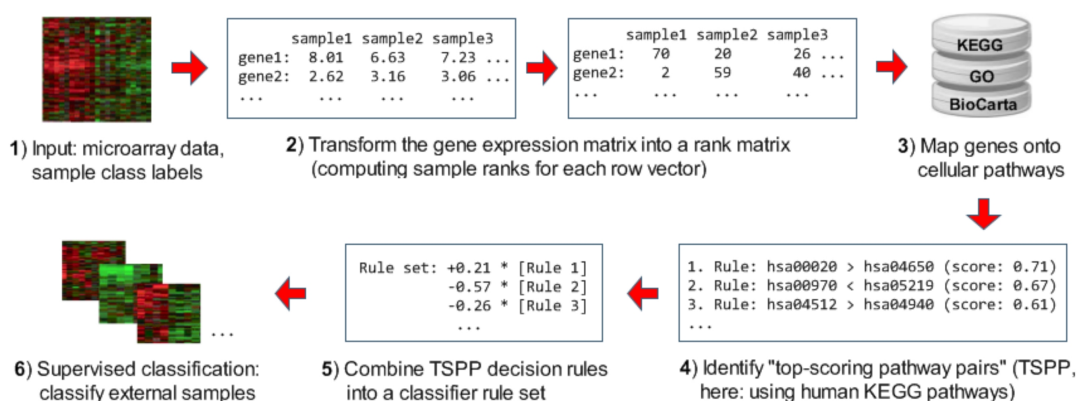


Figure 7.2: An overview of the workflow in the TSPP algorithm (example data is derived from a human prostate cancer microarray dataset [348])

in the sorted vector (ties are handled by replacing equal values by the mean of the corresponding position indices).

## 2. Pathway mapping:

Gene/protein sets representing cellular pathways and processes are extracted from a public database (e.g. KEGG [36], BioCarta [32], Reactome [37] or Gene Ontology [30]). Pathway assignments are obtained for the  $p$  genes or corresponding proteins in the microarray input data by testing whether they occur in any of these known gene/protein sets (for simplicity, in the following the term *genes* will be used instead of *genes/proteins*, but the method is also applicable to protein data). For genes which cannot be assigned to any pathway, the corresponding rows are removed from matrix  $R$ .

## 3. Scoring of pathway pairs:

To score a pair of pathways as being useful for discriminating between two sample class labels 1 and 2, e.g. "tumour (1) vs. normal (2)" or "drug treatment (1) vs. no treatment (2)", the pathway submatrices  $R_1$  and  $R_2$ , corresponding to these two samples classes, are extracted from matrix  $R$  using the mapping data from step 2. The matrices  $R_1$  and  $R_2$  are then reduced to robust pathway expression vectors  $r_1$  and  $r_2$  by replacing each column of expression level ranks by its median value. This median value computation is just one example for a great variety of dimensionality reducing data transformation that can be used at this stage (a principal component or multidimensional scaling reduction could likewise be applied, or the variance across genes could be computed to analyse pathway-variance patterns).

For a two-class prediction problem, the score for a pathway-pair is then obtained by comparing the median ranks in pathway 1 to those in pathway 2 and computing the maximum of two relative frequencies: The percentage of samples which are up-regulated for class 1 and down-regulated for class 2, and vice versa, the percentage of cases which are down-regulated for class 1 and up-regulated for class 2 (i.e. there are two possibilities for the relation of sample ranks in two pathways to differ across the sample classes). Given the sets of column indices for two sample classes  $S_1$  and  $S_2$ , the final score can thus be computed as follows:

$$partial\_score_1 = \sum_{i \in S1} I(r_{1i} \geq r_{2i}) + \sum_{i \in S2} I(r_{1i} < r_{2i}) \quad (7.1)$$

$$partial\_score_2 = \sum_{i \in S1} I(r_{1i} < r_{2i}) + \sum_{i \in S2} I(r_{1i} \geq r_{2i}) \quad (7.2)$$

$$score = \frac{\max(partial\_score_1, partial\_score_2)}{|S1| + |S2|} \quad (7.3)$$

where  $I$  is the indicator function. For a multi-class problem, a similar score can be obtained by computing the mean of the scores obtained for all pairs of sample classes. To obtain significance scores in addition to a relative ranking of pathway-pairs, the user can also apply a non-parametric statistical test, like the Wilcoxon rank sum test (at the expense of higher run-times).

#### 4. Searching for top-scoring pairs:

By default, top-scoring pathway pairs (TSPPs) are identified by performing an exhaustive search across all pairs of cellular pathways. This strategy will be feasible in most practical applications, because the number of pathways is typically much smaller than the number of genes, and the scoring method is kept simple and efficient (see above). Importantly, in spite of the simplicity of the scoring function, the method does not assume that all genes in a pathway are either up- or down-regulated in a sample class in relation to genes in another pathway, but searches for trends across most genes (i.e. identifying pairs of pathways for which many genes occurring in the first pathway change their relation of expression level ranks across the sample classes to genes in the second pathway).

A further variant of the search methodology was considered to investigate whether alterations in the pathway definitions can provide improved results. For this purpose, the user can allow the algorithm to introduce *mutations* into the pathway gene sets, by randomly adding or deleting genes up to a small user-defined maximum number, and replacing the exhaustive search by a previously published evolutionary search algorithm [479]. Only one modification is applied to the evolutionary search method: A genome contains two bit-vectors representing two pathways and mutations are only applied to one of these bit-vectors, selected randomly. Apart from these changes, the scoring function in the evolutionary algorithm is the same as for the exhaustive search variant (see above).

#### 5. Classification model generation:

Each TSPP provides a simple decision rule for classifying microarray samples depending on the relative median expression value ranks of their genes in a pair of pathways. To combine multiple TSPPs into a unified ensemble classification model, the TSPP decision rules are used as “base classifiers” within the Adaboost.M1 algorithm [297], adding one decision rule at a time to the boosting model according to the order of the TSPP-scores computed in the previous step. The main purpose of the boosting scheme is to assign weights to each decision rule in the combined ensemble model, accounting for a rule’s prediction accuracy and capacity to correctly classify samples that were misclassified by decision rules added in previous iterations of the algorithm. Previous experiments with boosting and ensemble techniques applied to microarray data (see section “Ensemble learning methods” in chapter 3) have shown that improvements can be obtained both in terms of robustness and accuracy. Thus, TSPP exploits cross-domain integrative analysis (using pathway and microarray data), ensemble analysis (using the AdaBoost algorithm) and a robust pairwise analysis of features.

## 7.2 Evaluation of Integrative Rule Learning on Microarray Data

In order to evaluate the predictive performance and investigate the insights that can be obtained from an integrative rule learning analysis of microarray data, the TSPP algorithm was applied to two public cancer gene expression datasets. Specifically, the data was retrieved from studies on B-cell lymphoma [347] (7129 genes and 77 samples) and prostate cancer [348] (12600 genes and 102 samples). Both datasets contain samples from two biological classes: In the B-cell lymphoma dataset, 58 samples were obtained from patients suffering from diffuse large B-cell lymphoma (class 1), while the remaining 19 samples are derived from a related follicular B-cell lymphoma (class 2; importantly, this class imbalance complicates the prediction problem). For the prostate cancer dataset, expression measurements were obtained from 50 healthy control tissues (class 1) and 52 tumour tissues (class 2) (for details on the normalisation and pre-processing of the datasets, see section 4.0.2).

To evaluate the predictive accuracy for TSPP models generated for these datasets, an external leave-one-out cross-validation procedure (ELOOCV, i.e. including all modelling steps in the cross-validation) was applied, and repeated using different numbers of top-scoring pairs  $k$  (for  $k = 1, 3, 5, 10$  and  $15$ ). The parameter  $k$  can be regarded as a bias/variance trade-off, enabling the user to control the complexity of the generated classifiers. The cross-validation results, computed both for mappings of genes to KEGG pathways and to Gene Ontology (GO) terms, include the average accuracy, sensitivity and specificity for each LOOCV run and are shown in Tables 7.1 and 7.2.

Table 7.1: Leave-one-out cross-validation results (TSPP on KEGG database)

Dataset	No. of top-scoring pairs	Sensitivity (%)	Specificity (%)	Avg. Accuracy (%)
Prostate cancer	1	83.7	71.7	77.5
	3	87.8	73.6	80.4
	5	85.7	77.4	81.4
	10	77.6	73.6	75.5
	15	79.6	64.2	71.6
Lymphoma	1	64.9	85.0	70.1
	3	68.4	90.0	74.0
	5	78.9	90.0	81.8
	10	77.2	90.0	80.5
	15	75.4	90.0	79.2

Leave-one-out cross-validation results obtained with the TSPP classifier using different numbers of top-scoring pathway pairs on the KEGG database.

In summary, average classification accuracies above 70% were obtained in all cases, and for both datasets the best accuracies (prostate cancer: 81.4%, DLBCL: 81.8%) were achieved when using 5 top-scoring pairs, suggesting that  $k = 5$  represents a reasonable bias/variance trade-off, providing models which are still easily interpretable.

Apart from using the decision rules for class prediction, their simplicity also makes them suitable for direct human interpretation. These rules, i.e. the ten top-scoring pathway pairs for each dataset, are shown in tables 7.4 and 7.5. Interestingly, the top-ranked rule for the prostate cancer dataset contains the KEGG-pathways *Prostate cancer* and *Insulin signaling*, which both are known to be de-regulated in the disease [480, 481]. However, the results also point to relative de-regulations in other pathways with less obvious associations to the cancer disease, e.g. *Pyrimidine metabolism* and *Glycerolipid metabolism*, with a ranking

Table 7.2: Leave-one-out cross-validation results (TSPP on GO database)

Dataset	No. of top-scoring pairs	Sensitivity (%)	Specificity (%)	Avg. Accuracy (%)
Prostate cancer	1	83.7	67.9	75.5
	3	89.8	67.9	78.4
	5	89.8	69.8	79.4
	10	91.8	66.0	78.4
	15	85.7	67.9	76.5
Lymphoma	1	68.4	80.0	71.4
	3	57.9	90.0	66.2
	5	71.9	90.0	76.6
	10	52.6	90.0	62.3
	15	71.9	85.0	75.3

Leave-one-out cross-validation results obtained with the TSPP classifier using different numbers of top-scoring pathway pairs on the GO database.

Table 7.3: Leave-one-out cross-validation results (eBayes &amp; SVM)

Dataset	No. of features (genes)	Sensitivity	Specificity	Avg. Accuracy (%)
Prostate cancer	2	88.0	84.6	86.3
	6	96.0	88.5	92.2
	10	96.0	86.5	91.2
	20	90.0	88.5	89.2
	30	90.0	90.4	90.2
Lymphoma	2	91.4	68.4	85.7
	6	93.1	78.9	89.6
	10	94.8	94.7	94.8
	20	96.6	84.2	93.5
	30	98.3	100.0	98.7

Leave-one-out cross-validation results obtained using eBayes feature selection on single genes and an SVM classifier.

score close to the best-ranked pair.

Similarly, for the B-cell dataset the top-ranked pathway pairs contain processes known to be associated with B-cell neoplasia, e.g. the *Wnt signaling pathway* [482, 483], whereas for other pathways no direct and specific associations with the disease are known. In spite of the class-imbalance in this dataset, the prediction models did not display a preference to assign samples to the majority class; however, similar to other statistical methods for microarray data analysis, problems with robustness can occur when the sample size per condition is very small. Thus, when planning a microarray study, the experimenter might first want to consider techniques for sample size estimation [484], microarray study design [485] and sampling techniques [486] to alleviate these problems.

Importantly, in a top-scoring pathway pair (TSPP), not necessarily both pathways are differentially regulated across the sample classes, but one pathway might have an almost constant expression, while the other pathway is highly de-regulated in one of the sample classes. The motivation for comparing pairs of pathways lies in the possibility to avoid comparing single pathways against fitted thresholds, which would more likely be affected by experimental bias and thus provide prediction models with higher generalisation error. However, if a user's main goal is to identify pathway associations rather than obtaining a prediction model, then TSPPs in which one of the pathways is not differentially regulated across the sample classes can easily be identified and filtered out by computing the variance for the corresponding gene expression vectors and removing TSPPs containing a low-variance pathway expression fingerprint.

When using the evolutionary search methodology and allowing the algorithm to introduce small numbers of random gene deletions and insertions into the pathways (up to five genes), in spite of the higher flexibility of this method, in all experiments the prediction accuracies are either similar or lower than those obtained for the original pathways using an exhaustive search (data not shown). The weaker performance might result from an entrapment in local minima due to the expansion of the search space, but could also suggest that the original pathways and processes are already well defined and therefore hard to optimise using an evolutionary search procedure.

Overall, the results from the cross-validation analysis and the lists of top-scoring pathways show that the method can generate compact predictive models with both high interpretability and high accuracy in comparison with a random model predictor (when applying the *proportional chance criterion* by Huberty [429], p-values < 0.01 are obtained in all cases). To show how these results relate to existing machine learning methods with single genes as predictors, a C-SVM [376] was applied using different kernel functions, including the radial basis function and polynomial kernels with a degree up to 3 (the results for the best kernel, a linear SVM, are reported in Table 7.3). The gene-based SVM-models achieve higher average accuracies than pathway-based models, with the best models reaching more than 90% accuracy on both datasets, which matches with the fact that the model operates at the finer level of detail of single genes instead of pathways. However, these models do not enable an interpretation of the data on the level of cellular pathways and processes, and are not designed to be applied on other array platforms with different genetic probes representing the same pathways. Although the simple decision rules generated by the TSPP algorithm do not reach the highest accuracies obtained by the support vector machine on single genes, their high interpretability and significant predictive information content allow the user to quickly identify cases, in which the relative gene expression in pathway pairs is differentially regulated across different biological conditions.

In order to investigate the biological utility of the top-scoring pathway pairs (TSPPs) in more detail, the genes in these pathways were mapped onto their corresponding proteins in a large-scale protein-protein

Table 7.4: Top-ranked pathway pairs (Prostate cancer data)

Rank	Pathway 1	Pathway 2	Direction	Score
1	hsa05215 Prostate cancer	hsa04910 Insulin signaling pathway	down	0.81
2	hsa00240 Pyrimidine metabolism	hsa00561 Glycerolipid metabolism	up	0.80
3	hsa04540 Gap junction	hsa05210 colorectal cancer	up	0.78
4	hsa04115 p53 signaling pathway	hsa00230 Purine metabolism	down	0.75
5	hsa04510 Focal adhesion	hsa00071 Fatty acid metabolism	down	0.75
6	hsa04514 Cell adhesion molecules (CAMs)	hsa04610 Complement and coagulation cascades	up	0.72
7	hsa03050 Proteasome	hsa01430 Cell Communication	up	0.69
8	hsa04920 Adipocytokine signaling pathway	hsa04730 Long-term depression	up	0.69
9	hsa04810 Regulation of actin cytoskeleton	hsa04530 Tight junction	down	0.65
10	hsa04512 ECM-receptor interaction	hsa04110 Cell cycle	down	0.63

The 10 top-ranked pathways for the prostate cancer dataset according to the TSPP-score (direction “down” means that in the healthy control samples, pathway 1 is down-regulated in relation to pathway 2, whereas in the prostate cancer samples, pathway 1 is up-regulated in relation to pathway 2, and respectively, “up” means the pathways have opposite relations in the two sample classes).

Table 7.5: Top-ranked pathway pairs (B-cell lymphoma data)

Rank	Pathway 1	Pathway 2	Direction	Score
1	hsa00020 Citrate cycle (TCA cycle)	hsa04310 Wnt signaling pathway	down	0.88
2	hsa00052 Galactose metabolism	hsa04664 Fc epsilon RI signaling pathway	down	0.87
3	hsa04670 Leukocyte transendothelial migration	hsa03050 Proteasome	up	0.87
4	hsa04514 Cell adhesion molecules (CAMs)	hsa00030 Pentose phosphate pathway	up	0.86
5	hsa04730 Long-term depression	hsa00240 Pyrimidine metabolism	up	0.85
6	hsa00562 Inositol phosphate metabolism	hsa00051 Fructose and mannose metabolism	up	0.84
7	hsa00220 Urea cycle and metabolism of amino groups	hsa00980 Metabolism of xenobiotics by cytochrome P450	down	0.84
8	hsa04540 Gap junction	hsa00330 Arginine and proline metabolism	up	0.84
9	hsa00252 Alanine and aspartate metabolism	hsa04630 Jak-STAT signaling pathway	down	0.84
10	hsa00970 Aminoacyl-tRNA biosynthesis	hsa04912 GnRH signaling pathway	down	0.81

The 10 top-ranked pathways for the B-Cell lymphoma dataset according to the TSPP-score (direction “down” means that in the DLBCL samples, pathway 1 is down-regulated in relation to pathway 2, whereas in the follicular B-cell lymphoma samples, pathway 1 is up-regulated in relation to pathway 2, and respectively, “up” means the pathways have opposite relations in the two sample classes).



interaction network, consisting of 38857 interactions between 9392 proteins assembled from direct binary interactions in a previous study [20] (see also the network construction methodology for TopoGSA, described in chapter 6). Figure 7.3a) shows the largest connected component of an example mapping for the TSPP with the highest score on the Prostate cancer dataset, *hsa05215 Prostate cancer vs. hsa04910 Insulin signaling pathway* (see also Figure 7.2), revealing a strong network connectivity between these pathways, which also share a significantly large set of overlapping genes/proteins (the q-value significance score is  $5.1E-17$ , when testing the Insulin pathway for overlaps with all other KEGG pathways using the one-sided Fisher exact test and adjusting for multiple testing with the Benjamini-Hochberg method [3]).

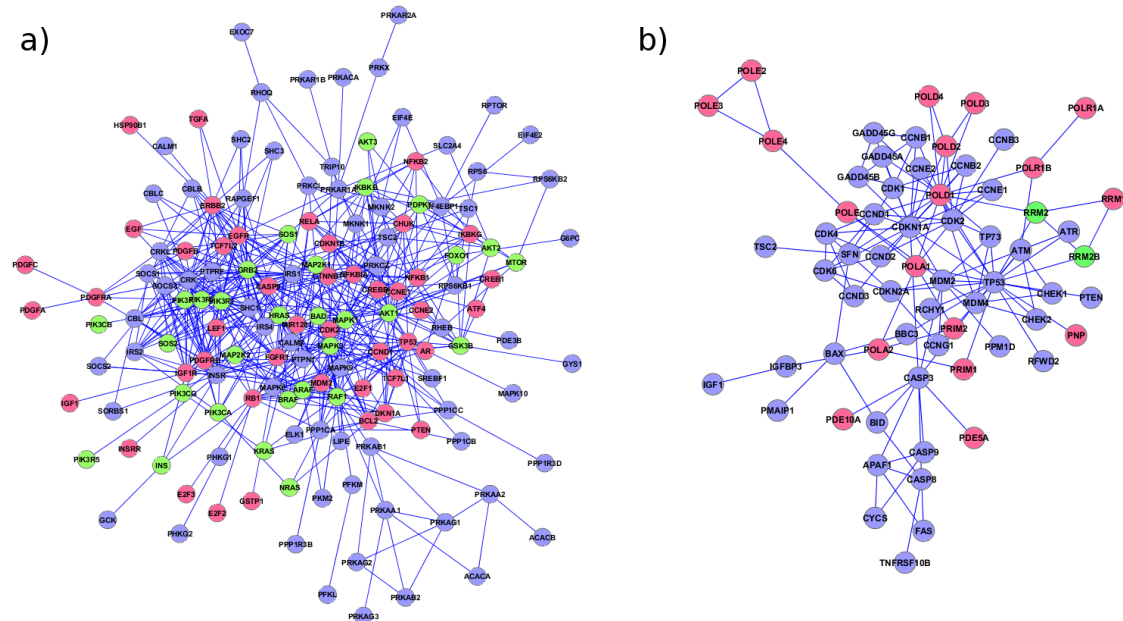


Figure 7.3: Analysing TSPPs in a protein-protein interaction network: a) Largest connected component for the KEGG pathways *Prostate cancer* and *Insulin signaling* (blue: *Prostate cancer*, red: *Insulin signaling*, green: members in both pathways); b) Largest connected component for the KEGG pathways *P53 signaling* and *Purine metabolism* (blue: *P53 signaling*, red: *Purine metabolism*, green: members in both pathways)

However, the TSPP method also points the user to differentially regulated pathway pairs which would not be detected as significantly associated using an overlap-based significance test, e.g. figure 7.3b) shows the largest connected component for the TSPP *hsa04115 p53 signaling pathway vs. hsa00230 Purine metabolism*, with only two overlapping proteins, but a multitude of direct protein-protein interactions between the two pathways. Further experimental evidence for an association between these pathways is provided by a study showing that the inhibition of *de novo* purine synthesis by the drug *AG2034*, which also inhibits prostate cancer cell growth, increases the expression levels of *p53* [487].

Accordingly, although the patterns identified by the top-scoring pathway pairs method do not necessarily result from regulatory relationships between the pathways, the analysis of the TSPPs can help to identify pathway pair relations associated with changes in biological conditions, which would remain unnoticed by other computational/statistical methods.

### 7.2.1 Conclusion

In summary, TSPP is a new method for extracting pathway-based decision rules from combined microarray expression data and gene/protein sets representing cellular pathways and processes. When applying prediction models derived from these decision rules for sample classification on two public microarray cancer datasets, compact and easily interpretable models are obtained with significant predictive information content. The generated decision rules are robust against monotonic transformations of the data, and the algorithm is easy to implement and has a comparatively short run-time due to the reduction of the data dimensionality when summarising gene sets to pathway expression fingerprints. Moreover, the resulting rule-based prediction models enable the interpretation of microarray data from a different perspective, at the level of pairwise relations between pathways. More specifically, the top-scoring pathway pairs can point the user to regulatory relationships or other functional associations between the corresponding pathways, which are associated with changes in the biological conditions, and would not be detected by other functional genomics analysis methods. Thus, in a nutshell, the TSPP algorithm provides both a novel method to generate compact and accurate classification models and a new exploratory tool to analyse microarray data at the level of pairwise pathway relations.

## Chapter 8

# Visualisation and Interactive Exploration of High-Dimensional Biological Data (VRMLGen)

### Chapter abstract

The analysis methods discussed in the previous chapters mostly rely on machine learning, network analysis and optimisation techniques, and provide results in the form of ranking tables or simple 2D graphical representations. However, as illustrated in the section “Comparative Evaluation of Clustering Methods” in chapter 4, creating low-dimensional representations directly from the original data is a further unsupervised analysis approach, enabling the experimenter to visually identify informative structures in the data.

Both for dimensionality reduction and 2D and 3D data visualisation, a multitude of software tools are already available, however, in particular for high-dimensional, noisy biological data with a wealth of additionally available annotation data, standard visualisation software often does not fully exploit the potential of flexibly combining different pre-processing and data transformation methods and providing interactive means to explore a data representation and interlink it with external data repositories.

Since web-based low-dimensional data visualisation is required in different analysis modules of the integrative framework presented here, this chapter will discuss a dedicated software package for 3D interactive data visualisation, *VRMLGen* (using material from the original publication [24]), which has been developed as a general purpose software package and is used in several components of the framework. *VRMLGen* creates 3D visualisations in common web-formats like the Virtual Reality Markup Language (VRML) and LiveGraphics3D, including 3D charts and bar plots, scatter plots with density estimation contour surfaces, visualisations of height maps, 3D object models and parametric functions. To maximise flexibility, the user can also access low-level plotting methods through a unified interface and freely group different function calls together to create new higher-level plotting methods.

Since *VRMLGen*'s functionality is not limited and specifically targeted towards integrative biological data analysis, but represents an important component of the integrative analysis

framework, this chapter will focus on providing a brief overview of VRMLGen's functions and features, and how they are used in the framework to facilitate the analysis of functional genomics data (for a more detailed discussion of VRMLGen, see [24]).

## 8.0.2 VRMLGen - Background and motivation

Low-dimensional data visualisations using dimensionality reduction techniques like *Principal Component Analysis* (PCA, figure 4.14), *Independent Component Analysis* (ICA, figure 4.15), *Isomap* (figure 4.16) and *Locally Linear Embedding* (LLE, figure 4.17) provide an intuitive means to identify underlying structures and patterns in a data set which would otherwise often remain undetected. For this reason, a wide variety of software tools exist to generate 3D visual data representations with interactive means to explore different 2D perspectives, analyse class membership and data density using colour schemes and contour surfaces, or compare different data sources using overlay-plots. The widely used statistical programming language R [353], for example, contains several software packages for 3D data analysis including a package to interconnect R with OpenGL [488] and various packages for the visualisation of multivariate data [489–494].

However, these tools are not tailored towards the analysis of biological data, in which the molecular entities (genes, proteins, and metabolites) are associated with numerous functional annotations, chromosomal/subcellular localisation information and further meta-data stored in different databases, which can typically not be interlinked with the graphic representation of the data. Moreover, although users can choose between many freely available offline tools to inspect 3D data on their own computer, currently available programming libraries for 3D statistical data plotting do not provide features to make interactive 3D visualisations directly viewable on the web.

For this purpose, the *VRMLGen* software package was developed for the R statistical data analysis environment to enable users to generate interactive web-based visualisations for 3D input data, enabling the user to annotate the data points with biological information and interlink them with public web-databases. *VRMLGen* visualises charts, graphs, bar plots and scatter plots, 3D meshes and parametric functions in two common web-formats, VRML (Virtual Reality Markup Language) and LiveGraphics3D. The software is used within the web-application *ArrayMining.net* (see chapter 5) to create such visualisations automatically when a new analysis task is submitted on the web-interface.

Although comparative analyses of unsupervised dimensionality reduction methods suggest that some techniques tend to provide superior results in relation to other methods [398], *ArrayMining* prevents the user from relying on a single reduction algorithm and instead enables a comparison between multiple methods. Specifically, on the *ArrayMining* clustering module, the user will obtain multiple low-dimensional VRML visualisations for each analysis task, including *Principal Component Analysis* (PCA), *Independent Component Analysis* (ICA), *Isomap* and *Locally Linear Embedding* (LLE). These visualisations can be viewed directly in the browser or downloaded on the bottom of the clustering results web page.

## 8.0.3 VRMLGen - Methods

The *VRMLGen* software package contains both higher-level plotting functions for creating common data visualisations like scatter plots, bar charts and mesh visualisations, and lower-level methods, enabling users to draw shapes and objects through a unified, format-independent interface.

Many higher level plotting tasks can be realized directly with a single function call of one of VRMLGen's three main plotting functions:

- the *cloud3d()* function for creating 3D scatter plots and contour surfaces,
- the *bar3d()* function for generating bar plots and height map visualisations, and
- the *mesh3d()* function for displaying parametric functions and 3D meshes.

Figure 8.1 provides an overview of the main plot types and functions available in the VRMLGen package.

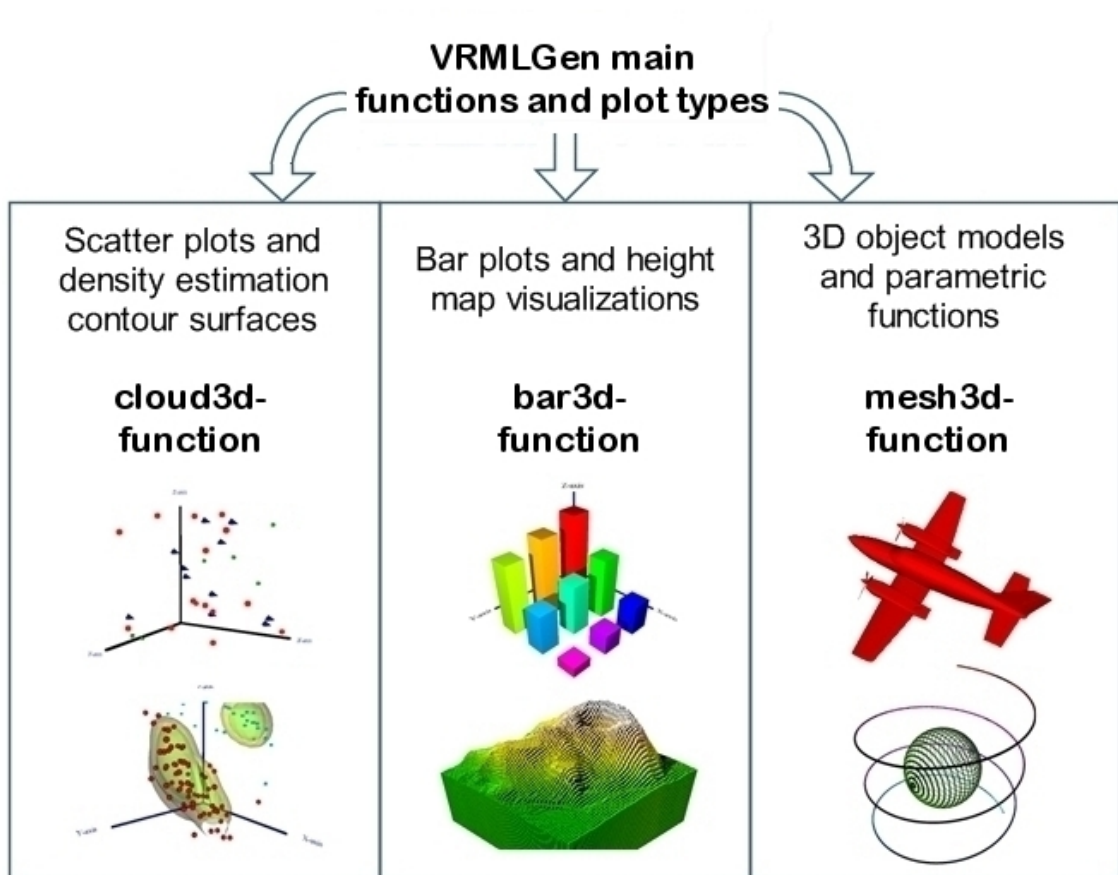


Figure 8.1: Overview of VRMLGen's main functions and features.

Alternatively, for users with more advanced computational skills, VRMLGen provides direct access to several lower level plotting functions (for drawing points (*points3d()*), lines (*lines3d()*), text (*text3d()*) and coordinate axes (*axis3d()*) through a format-independent interface, enabling users to switch between the VRML- and the Livegraphics3D-output format by changing a single parameter.

In addition to the functionality provided by these plotting methods, both low- and high-level function calls can be combined together by placing them between *open()* and *close()* statements (using *vrml.open()* or *lg3d.open()*, depending on the plot type). This modular design enables the user to build complex 3D scenes by sequentially adding new primitive objects, or instances of newly defined object groups.

In contrast to other plotting functions in the *R* statistical programming environment and other statistical

software tools, VRMLGen provides features related to the interactive exploration and HTML-embedding of the 3D data. For example, the *cloud3d* function allows users to specify biological “meta-labels” for data points in a 3D scene, i.e., gene or protein labels, or functional annotations are displayed when the user hovers the mouse over a data point in the 3D plot (available in all Javascript-enabled VRML-plugins). Moreover, hyperlinks can be added to each object in a 3D scene, interlinking data points corresponding to biological entities (genes, proteins and metabolites) with their corresponding database entries in online annotation databases (e.g. the ENSEMBL database for genes or SwissProt for proteins).

#### 8.0.4 VRMLGen - Example application for bioinformatics data visualisation

In the following paragraphs, some of VRMLGen’s basic and extended functionalities, including meta-labels and density estimation contour surfaces, are illustrated using an example dataset from a breast cancer microarray study in collaboration with the Queen’s Medical Centre (QMC) in Nottingham [19, 349–351] (see chapter 4 for a detailed description of the data).

VRMLGen’s *cloud3d* function to create 3D scatter plots allows the user to provide any type of numerical data as input that can be coerced to a numerical matrix with 3 columns (e.g. 3 vectors, a matrix, or a formula with 3 variables). Using microarray data after dimensionality reduction (with PCA, ICA, Isomap or LLE; see chapter 4) as input, the information from class labels and sample annotations can automatically be integrated into the generated visualisations. For example, class labels are automatically highlighted in the plots by different colours and point styles, and a class legend is generated above the Z-axis.

To illustrate the functionality using the QMC breast cancer dataset (and complement the VRMLGen visualisations shown in chapter 4), two diverse classes of samples were extracted from the data, corresponding to samples with two different tumour grades (tumour grade 1, the least aggressive form of the tumour, and tumour grade 3, the most severe stage of the tumour). The dimensionality reduction method Independent Component Analysis (ICA), whose visualisations tended to provide the best class separation in the comparison of reduction techniques in chapter 4, was then applied using the *fastICA* R software package [495] in combination with VRMLGen. Figure 8.2, left, shows the resulting VRMLGen visualisation, revealing that the samples from the two classes can be successfully separated in 3D space by means of the ICA data transformation. Similar visualisations of microarray data can be generated either directly with the VRMLGen software package or using only a few mouse-clicks on the ArrayMining.net clustering module to inspect the data as an interactive plot on the web.

Apart from the colour highlighted sample classes, generated automatically by VRMLGen when providing class labels, density estimation contour surfaces can optionally be added to the plot (Figure 8.2, right), by changing a single parameter setting. These contour surfaces represent regions of high data density in the plot (visualised by green and yellow surfaces, with green representing the highest data density). In the breast cancer example dataset, interestingly, most of the low-grade tumour samples appear to be clustered together in a high-density region of the plot, whereas the high-grade tumour samples are scattered more widely across the plot. This observation agrees well with earlier findings according to which gene expression levels in samples with strong tumour activity tend to display a higher variance than for example in cell samples after tumour resection (i.e. the surgical removal of the tumour) [496].

If the user wishes to include additional meta-information for the data (e.g. function annotation data) in a plot, or interlink each data point with an entry in a public web database, then the necessary information only needs to be specified in a vector of “meta-labels” or hyperlinks. After creating the visualisation, the user

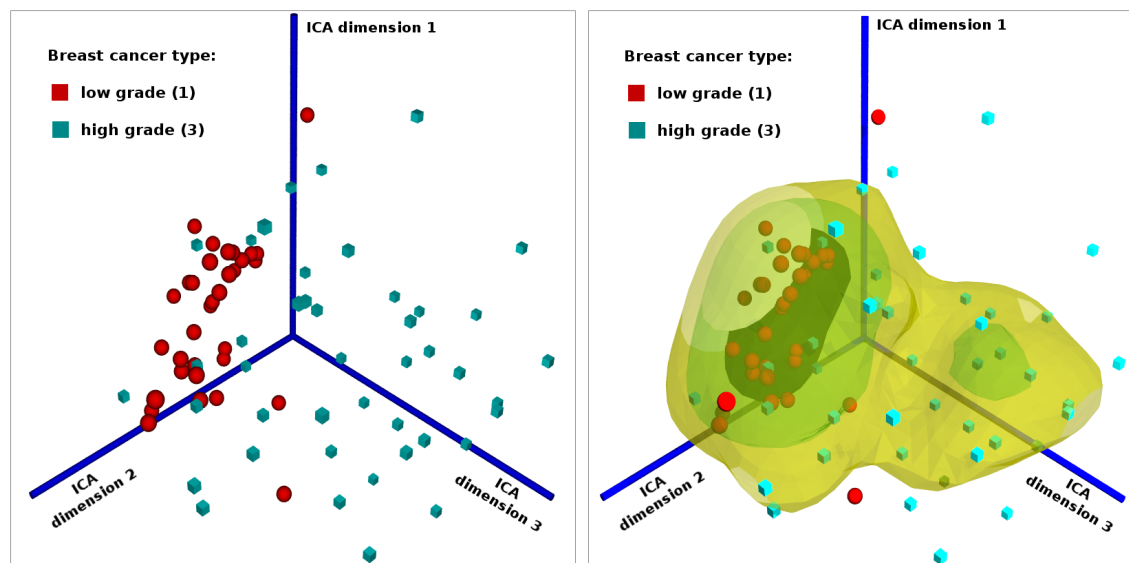


Figure 8.2: Left: Scatter plot visualisation of breast cancer microarray samples [19, 349–351] after dimensionality reduction with an Independent Component Analysis. Right: The same scatter plot with density estimation contour surfaces (yellow = high data density regions, green = region of highest density), see source code in Figure 8.3).

```
R> # VRMLGen - Example source code (R statistical programming language)

R> data("bc_dat")
R> data("bc_classes")

R> cloud3d(bc_dat, labels = paste("tumor grade",
+   as.numeric(bc_classes)), vml_showdensity = TRUE,
+   metalabels = paste("sample", 1:nrow(bc_dat)),
+   lab.axis = paste("ICA dimension", 1:3),
+   filename = "example2.wrl", htmlout = "example2.html")
```

Figure 8.3: VRMLGen Example source code: Generating a 3D VRML scatter plot for a breast cancer gene expression dataset obtained after dimensionality reduction with Independent Component Analysis (see Figure 8.2). First, the data, `bc_dat`, and the class labels, `bc_classes`, are loaded. Next, a plot with density estimation contour surfaces (`vml_showdensity`), meta-labels for the samples (`metalabels`) and user-defined axis labels (`lab.axis`) is generated, and the output is embedded in an HTML-file (`htmlout`). If the last parameter is not provided, only the VRML output will be generated.

can access the database entry for a data point by clicking on it, or view the meta-information as a tool-tip message above the Z-axis, when hovering the mouse over a data point.

To illustrate how complex plots with density estimation contour surfaces can be generated with a single function call, the R source code in figure 8.3 reveals how the visualisation for the breast dataset (see figure 8.2) can be generated (for other examples and more detailed explanations of VRMLGen's features and functions to generate other plot types like 3D meshes and parametric function visualisations, see [24]).

### 8.0.5 VRMLGen - Summary and conclusions

VRMLGen is a software package to generate 3D visualisations for interactive data exploration on the web, providing particular benefits for the analysis of high-dimensional biological data with additionally available

functional annotations, class labels and associations with external database entries. It contains functions for the creation of 3D scatter and bar plots, visualisations of meshes, parametric functions and height maps, as well as access to low-level plotting functions. Calls of different plotting functions can be grouped together to generate complex 3D scenes, and the user can program new higher-level plotting methods using the already existing functionality. Additional features, targeted specifically towards the analysis of biological data, include the possibility to assign hyperlinks and meta-labels to the data points, e.g. to interlink genes and proteins with corresponding online database entries. Moreover, regions of high data density can be highlighted by contour surfaces to assist the user in visually identifying cluster structures in a dataset. All outputs can be generated in two common web-formats, the VRML- and LiveGraphics3D-format. Apart from using the obtained 3D plots for direct visual inspection and to present the data on the web, the visualisation files can also serve as input for freeware rendering software to generate high-quality perspective plots for scientific publications.

The VRMLGen package is available as a standalone software package from the Comprehensive R Archive Network (CRAN, <http://cran.R-Project.org>) and from the Nottingham University Infobotics project repository (<http://www.infobotics.net>). Finally, the software package is employed within the ArrayMining web-server for microarray data analysis [18] to generate low-dimensional visualisations of unlabelled microarray data (see also the discussion of the ArrayMining clustering module in chapter 5).

In summary, the low-dimensional VRMLGen visualisations can often facilitate the interpretation of microarray data, allowing a user to visually and interactively explore the data and identify sample outliers and cluster groupings in the data, which are sometimes not detected by automatic bioinformatics methods.



## Chapter 9

# Main Biological Contributions

### Chapter abstract

Although algorithms, software tools and data analysis pipelines are the main deliverables of this doctoral project, the implemented analysis techniques have also been used to address current research questions in the biosciences with a focus on cancer biology.

This chapter will provide an overview of the main biological findings that were obtained by applying tools from the framework, including ArrayMining, TopoGSA, PathExpand and TSPP, to real-world problems based on recent data from collaborating institutions.

Following the structure of the previous chapters, the first section will present results derived from the integrative machine learning analysis of high-dimensional biological data, whereas the second part discusses results obtained from the analysis of general gene and protein lists, representing cellular pathways and processes.

### Breast cancer tumour sub-classification and marker gene identification

The study of heterogeneous diseases with complex genetic components and in particular cancer diseases, which are known to result from a variety of possible causes (e.g. hereditary risk factors, environmental influences, virus infections and spontaneous mutations [65, 66]) is a prime target for new high-throughput data analysis techniques, like gene and protein expression microarrays. Since deregulations of whole cellular signalling pathways and networks of interactions between multiple genes and proteins have been found to influence these diseases, systems biology approaches investigating entire genomes, transcriptomes, proteomes and metabolomes are suitable means to analyse these networks of associated biomolecules.

In collaboration with the department of Histopathology at the Queens Medical Centre in Nottingham, the integrative analysis modules from ArrayMining were used to analyse data from a large-scale microarray cohort study with samples from 128 breast cancer patients and 47,293 gene transcripts [19, 349–351] (see data description in chapter 4). The biological goal was to identify and experimentally validate tumour marker genes, which discriminate between two major breast cancer tumour subtypes, the luminal and the non-luminal subtype. These subtypes reflect whether certain protein receptors (in particular the oestrogen receptor, ER, and the human epidermal growth factor receptor 2, ErbB2) are expressed on the tumour cell surface, and influence the choice of the treatment for a patient and the clinical prognosis (e.g. the

risk for tumour relapse, preoperative chemotherapy response [497]). Importantly, an early and accurate diagnosis of the tumour type tends to improve the outcome of the therapy. More specifically, in recent years a sub-classification of the two major breast cancer categories into five molecular subtypes has become widely accepted: Luminal samples are divided into luminal A and B, and non-luminal samples into basal-like, ErbB2 overexpressing and normal-like samples [498–501]. Luminal A tumours have a relatively good prognosis in relation to luminal B tumours, since they express the hormone receptor ER at a higher level and can therefore be targeted by a therapy with the drug substance tamoxifen [502]. For ErbB2+ (overexpressing) breast cancers, the prognosis is poor, but a targeted treatment using the drug substances trastuzumab or lapatinib exists [503]. Similarly, for basal-like tumours, which lack the ER and ErbB2 receptors, the prognosis is poor, but poly(ADP-ribose) polymerase (PARP) inhibitors targeted against basal-like tumour cells with defects in DNA repair pathways are currently being developed [502]. Finally, for normal breast-like tumours, the biology is not yet well understood and the prognosis is similar to the basal-like tumours. Overall, as a first diagnostic step, the general categorisation of a sample into the luminal and non-luminal group is an important pre-requisite for the choice of an appropriate treatment (e.g. hormone, radiation, chemotherapy or surgical treatment).

To maximise the robustness of biomarker screening approaches for the identification of genes discriminating between the two breast cancer subtypes, a combination of an ensemble feature selection and ensemble learning approach with cross-validation based sub-sampling using the ArrayMining framework was applied. Specifically, three diverse feature selection methods, a combinatorial method (CFS [160]), a tree-based feature ranking approach (RFS [190]) and a classical univariate filter (PLSS [356]), were employed within an external leave-one-out cross-validation (LOOCV) procedure, in combination with four classification methods, the in-house developed evolutionary computation based machine learning system BioHEL, a linear C-SVM [346], Breiman's random forest classifier (RF) [191] and the nearest shrunken centroid classifier (PAM) [239] (see chapter 4 for details). The genes were ranked according to how often they were chosen among the top 30 features in the LOOCV across each of the three selection methods. The final list of top-ranked genes consisted of those features that were selected by at least two feature selection methods across all cross-validation cycles - three of these genes were in fact selected across all three selection approaches: The gene *RAS-like, estrogen-regulated, growth-inhibitor (RELG)*, identifier: *GI\_14249703-S*, *estrogen receptor 1 (ESR1)*, identifier: *GI\_4503602-S* and *potassium channel, subfamily K, member 15 (KCNK15)*, identifier: *GI\_16507967-S*. The complete ranking list can be found in table 4.11.

The *ESR1* gene, which encodes the oestrogen receptor (ER)  $\alpha$ , is already a well-known breast cancer marker gene. In luminal samples, the ER- $\alpha$  is known to be expressed in tumour cells (ER+, see above), whereas it is not expressed in basal-like samples (ER-, see the approach by Nielsen *et al.* for breast cancer subtype categorisation [504]). The oestrogen hormone is well-known to cause the growth of ER+ breast cancer cells, and some hormone therapies use anti-oestrogens as drugs against corresponding forms of breast cancer.

The second top-ranked gene, *RELG* (see gene expression box plot in figure 9.1), is again known to be associated with breast cancer, but had not been used before to separate the luminal and the non-luminal class as distinct prognostic subgroups. *RELG* is a GTP-binding protein and its expression has been reported to be decreased in aggressive ER negative subtypes [505]. Moreover, *in vitro* studies in which *RELG* expression was found to be induced in ER-responsive MCF-7 cells stimulated by estradiol and repressed by tamoxifen treatment confirm a potential tumour suppressive role of the gene. Therefore, this gene was investigated in detail as a potential tumour marker in a recently published pre-clinical study in collaboration with the Queen's Medical Centre in Nottingham [19]. The study validated the microarray results using immunohistochemistry on tissue microarrays (TMAs) containing 1,140 invasive breast cancers (see figure 9.1).

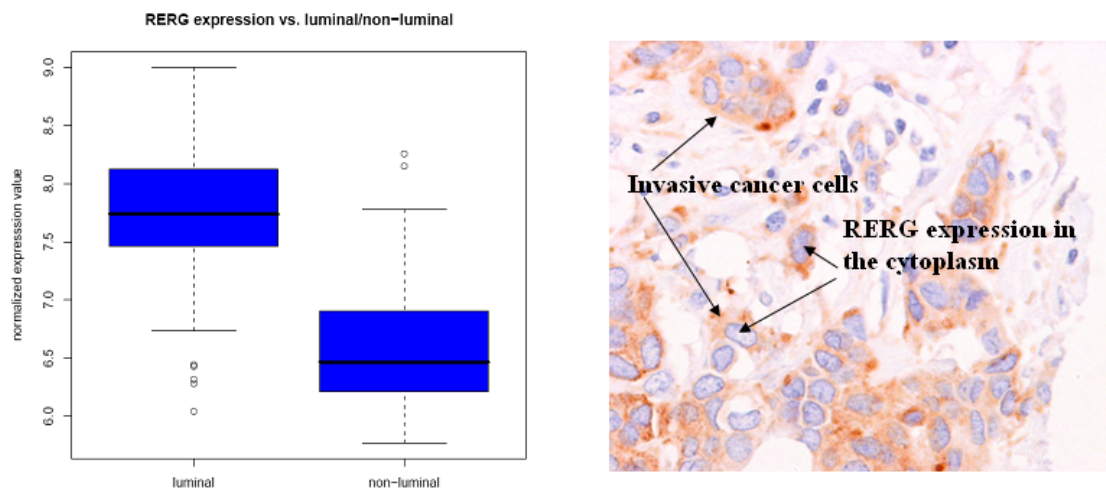


Figure 9.1: Left: Box plot revealing the gene *REG*'s potential as a new candidate marker of the ER-positive luminal-like breast cancer subtype. Right: Tissue microarrays of invasive breast cancer show strong *REG* expression. The image displays infiltrating and malignant tumour cells in breast tissue (invasive ductal carcinoma cells, IDC) after immunohistochemical staining of the *REG* protein using monoclonal antibodies (blue colour, see [506, 507] for details on the immunohistochemistry procedure). These regions stained in blue correspond to the cytoplasm of the tumour cells (which have a light orange/pink colour and include the blue cytoplasmic regions), revealing that the *REG* expression is localised to the cytoplasm. Moreover, by assessing the intensity of staining and the percentage of stained cells following immunohistochemistry using the histochemical score (H-score) [508], *REG* was found to be more strongly expressed in the cytoplasm of luminal cells than in non-luminal cells [19].

Specifically, the protein expression study revealed that *REG* expression was positively associated with several markers of luminal differentiation including ER, luminal cytokeratins (CK19, CK18, and CK7/8) and FOXA1 ( $p=0.013$ ), as well as with other clinical markers for good prognosis in breast cancer including small tumour size, low histological grade and positive expression of androgen receptor, nuclear BRCA1, FHIT and cell cycle inhibitors p27 and p21. An inverse association with *REG* expression was detected for the proliferation markers MIB1, P53 and EGFR. Moreover, strong *REG* expression was associated with longer breast cancer specific survival (BCSS) and distant metastasis free interval (DMFI) in the whole series as well as in the ER+ luminal group (independent of other prognostic variables) [19].

A role for the third top-ranked gene, *KCNK15* (also known as *TASK-5*), in oncogenesis is currently unknown, but the gene has been found to be silenced by hypermethylation of the promotor region in many tumours [509]. *KCNK15* encodes a two-pore potassium channel protein, which matches to findings for other ion channels, like the  $\text{Ca}^{2+}$  channel CACNA1G and the  $\text{Na}^{+}$  channel SLC5A8 with putative tumour suppressive function, which have already been reported to be hypermethylated in different cancers [510, 511]. Future investigations are planned to analyse this gene/protein in more detail.

### Identifying associations between cancer mutated genes, cellular pathways and different disease processes

As noted in chapter 2, multifactorial genetic diseases can often be interpreted as modulations of normal cellular pathway and process activities, where defects in different genes can have the same final effect. For example, the normal activity of the *Wnt*/ $\beta$ -catenin signalling pathway, a pathway associated with embryogenesis, can be disrupted by a deactivation of a protein known as *Adenomatous-polyposis-coli* (APC) protein, or by a mutation of the  $\beta$ -catenin protein that prevents its degradation, which can lead to the development

of the same form of colorectal cancer in both cases [512].

For this reason, the pathway-based analysis of high-throughput biomedical datasets and in particular research aimed at discovering disease-associated cellular processes is an effective means to improve the understanding of the molecular basis behind these diseases. In cooperation with the Spanish National Cancer Institute (CNIO), the software tool PathExpand from the framework (see chapter 6) was used to extend the definitions of various disease pathways by extracting information from molecular interaction data, and to analyse their associations with gene sets known to be mutated in different cancers. By extending the network representations of the original pathway and process definitions, novel putative candidate disease genes could be identified among the new genes/proteins added to the pathway definitions. For example, among five genes added to the Alzheimer's pathway from the KEGG database, three had already previously been implied in the disease (*TMED10*, *APH1B* and *PITX3*), and the remaining two, *METTL2B* and *MMP17*, were studied in more detail. Although for *MMP17*, a metalloproteinase protein, there was no other public experimental evidence implicating a role of the gene in Alzheimer's, six other members of the same protein family have been associated with the Alzheimer's disease (see HUGO navigator [464]). Similarly, the second gene, *METTL2B*, a methyltransferase-like protein, did not occur in any public disease associated datasets, but a member from the same protein family, *MMETL10*, had previously been associated with Alzheimer's disease in a case-control study [464]. For this reason, *MMP17* and *METTL2B* were proposed as new candidate disease genes in the PathExpand paper [22].

In the same study, extended cellular pathways were also used to study their enrichment in cancer mutated genes. Specifically, a large set of pancreas mutated genes had been obtained from pancreatic resequencing studies, and this set was tested both against the original cellular pathway definitions as well as against the extended pathways using a one-sided Fisher exact test. Several pathways, whose original definitions had already been enriched in cancer mutated genes received higher enrichment scores after the extension, due to an over-representation of these cancer mutated genes among the added genes. A particularly interesting case was the cancer-related *cell cycle G1/S check point process* from the BioCarta database, which already contained mutated genes in the original pathway/process definition. Seven proteins were added to this process, all of which were either transcription factors, kinases or other signal transduction regulators, and six of them are known to be involved in cell cycle regulation - all except the gene *TGIF2*. However, *TGIF2* is mutated in a pancreatic tumour and known to be amplified in some ovarian cancers. Thus, the evidence for the involvement of the seven predicted pathway members in the corresponding process and pancreatic cancer provides a starting point for further research to improve the understanding of the disease.

### Pairwise cellular pathway associations deregulated in genetic diseases

After having analysed diseases with genetic components on the level of cellular pathways, processes and complexes rather than at the single-gene or -protein level, a next logical step leads to the analysis of pairwise relations between pathways and processes, and how these relations change across different disease phenotypes.

Although the main motivation behind the development of the *Top-scoring pathways pairs* (TSPP) method [23] (see chapter 7) was to increase the robustness of microarray sample classification by replacing single gene predictors with pathway expression fingerprints and comparing relative expression values against each other rather than fitting continuous-valued threshold values, this method also enabled the identification of pairwise associations between pathways, whose genes change their relation of expression values across different disease conditions.

For example, the KEGG pathways *p53 signalling* and *purine metabolism* share only two genes/proteins, however, the TSPP method shows that the relation between the pathway expression fingerprints for these two processes changes across the sample classes in the prostate cancer microarray dataset by Singh *et al.* [348]. Indeed, when mapping the two pathways on a large human protein-protein interaction network [20], a multitude of direct protein interactions connect the two pathways tightly in the network (see chapter 7). Moreover, recent findings from the literature corroborate the association of the pathways with the disease: The inhibition of *de novo* purine synthesis by a drug was shown to also inhibit prostate cancer cell growth and increase the expression levels of p53 [487]. This is consistent with the inverse relation between gene expression levels in p53 signalling and purine metabolism pathways in prostate cancer detected by the TSPP method.

### Genome-wide transcription network analysis in seeds

The study of networks using bioinformatics approaches does not only involve the analysis of molecular interaction networks representing direct, physical interactions between functional molecules in a cell, but also networks of indirect functional relations inferred from high-dimensional biological data, e.g. gene co-expression data from microarray experiments.

In co-operation with the Division of Plant and Crop Sciences at Nottingham University, condition-dependent gene co-expression in seeds was analysed using 138 microarray samples from the plant model organism *Arabidopsis thaliana*, representing the two temporal conditions of dormancy (73 samples) and germination (65 samples). These conditions are associated with important agronomic and ecological traits, and improving the understanding of the gene regulatory network influencing these conditions can therefore also facilitate research aimed at enhancing seed performance in agriculture. Since classical gene co-expression data analysis is unsupervised, the condition-dependent analysis of this data represents a new type of analysis, which can identify gene associations specific to a certain developmental state.

In order to create a condition-specific gene co-expression network, the microarray data obtained from the 138 Nottingham Arabidopsis Stock Centre (NASC) arrays was first normalised using the Affymetrix GCOS/MAS5 method [352] with a trimmed mean target intensity value (TGT) of 100. Genes that were not expressed at least once above the background level in any of the samples were removed from the analysis, resulting in a pre-processed dataset with 14,092 genes. On this data, a weighted gene co-expression network analysis (WGCNA) was applied using the ArrayMining network analysis module [18] (see chapter 5). Moreover, a further unweighted network was computed by creating an edge between two genes if the absolute value of the Pearson correlation coefficient between their expression values exceeds a cut-off  $\rho > \tau$  where  $\tau$  was chosen as 0.75, a parameter choice providing a network closest to having a scale-free topology [122]. To enable researchers to inspect this data online using a queryable interface, the interactive network visualisation *SeedNet* was built (available online [35], see also figure 9.2). SeedNet enables users to zoom into the co-expression network, search and highlight genes of interest, forward chosen gene sets to other analysis services (e.g. to find overrepresented cis-regulatory elements, i.e. regions of DNA or RNA regulating the expression of genes on the same chromosome) and download images of the network representation. Moreover, the condition-specific up- and down-regulation of genes is highlighted by a colour coding in the network (blue background = up-regulated in germinating samples, red background = up-regulated in non-germinating samples; computed using the *Significance Analysis of Microarrays* (SAM) method [419], see figure 9.2).

The network analysis verified the existence of two distinct modules of state-dependent genetic interactions,

SeedNet - Genome-wide transcription network in seeds

**1)**

www.WIGIs.net

Save Graph as an Image (right click and "save as")

Reset PROMOMER ATHENA

**Reset** will reset the view of SeedNet  
**PROMOMER** will take all selected node AGIs to the Promomer tool at the BAR to find over-represented cis elements.  
**ATHENA** will take up to a maximum of 100 selected node AGIs to the ATHENA website to find over-represented cis elements.

**Please note:** Node labels will always be shown vertically centred in front of the corresponding selected nodes.  
 To see the node and its label at the same time, please zoom in with your mouse wheel or using the overview window.

**2)**

**3)**

**Selected Node Details**

**Search**

Example searches: [At3g24650](#), [abi](#), [exp](#), [stome](#), [multiple identifiers](#).

or insert your own search terms:

abi

Search

6 results

- Select All
- At5g42030 (ABI4)
- At3g24650 (ABI3)
- At4g26080 (ABI1)
- At5g57050 (ABI2)
- At2g40220 (ABI4)
- At2g36270 (ABI5)

**Appearance**

**Labels**

Please enter a gene name to highlight its first neighbours (examples: [At3g24650](#), [At5g13310](#)):   **4)**

Figure 9.2: The SeedNet network visualisation interface, enabling a user to find modules of co-expressed genes in dormant and germinating *Arabidopsis thaliana* seeds to better understand the transition between these phases. The interface consists of four windows: **1)** the *Detailed View* window to visualise and explore co-expressed genes, **2)** the *Overview* window to zoom into different regions of the network, **3)** the *Menu* window to search and highlight genes in the network, and **4)** the *Neighbourhood search* window to find the co-expression partners of a specified gene (see <http://vseed.notttingham.ac.uk> for details).

identified using the MCODE graph clustering software [336] and automatic graph layout generation, corresponding to the germination and dormancy phases. Specifically, genes associated with seed dormancy are up-regulated by the hormone abscisic acid (ABA) and down-regulated by the hormone gibberellic acid (GA), and located in a distinct region the network, whereas genes related to germination have the opposite regulation pattern and are clustered in a separate network region.

One of the main benefits of the new network model is that the accuracy by which regulators of germination can be predicted was improved from 22% by using genes within the SAM top-ranked non-germination (NG) list to 50% when considering high-degree “hub” genes within the dormancy region of the network. This suggests that the consideration of the network topology has an added predictive value for the inference of gene functions, complementing the results from a differential gene expression analysis.

In summary, information from the SeedNet co-expression network enables the prediction of novel regulator genes for seed germination and the identification of state-dependent sets of genetic interactions associated with germination and dormancy. The co-expression data and the network visualisation is publicly available at: <http://vseed.nottingham.ac.uk/>.

# Chapter 10

## Conclusions

### Chapter abstract

As part of this dissertation three types of new scientific results in the field of functional genomics data analysis have been presented:

- *Novel methodological results:* New ensemble/consensus algorithms and cross-domain analysis techniques, and new analysis pipelines were developed and implemented as publicly available software tools.
- *Comparative evaluation results:* A systematic comparative evaluation of supervised and unsupervised techniques for high-dimensional microarray data analysis, including ensemble and consensus methods, was conducted.
- *New biological results:* A new breast cancer biomarker was proposed as part of a pre-clinical study, and new candidate disease genes and extended pathway/process definitions were identified for different cancers and genetic diseases.

Below, the findings for each of these deliverables are summarised and remaining limitations and challenges are discussed. The final section will conclude the thesis with an outlook on future perspectives in the field of integrative bioinformatics and potential extensions of the work presented here.

### 10.1 General Summary and Discussion

**New methodological results:** The main deliverable of the dissertation was an integrative platform for the analysis of experimental gene and protein data, consisting both of new combinations between already known machine learning, cross-study normalisation, gene set and network analysis methods, i.e. new data analysis pipelines, as well as new algorithms, integrating multiple data types and analysis techniques.

The largest software project within this framework is *ArrayMining* [18], a web-server for microarray data analysis, extending classical machine learning methods and specialised gene expression data analysis methods by two types of integrative analysis: *Ensemble and consensus techniques* for single algorithmic problem types (within analysis modules), and *cross-domain analysis approaches* combining different problem types (interlinking different analysis modules). In addition to providing new data mining methods, the system



enables a user to set up *new integrative analysis pipelines* for microarray data, e.g. analysing combined data from two different studies obtained using cross-platform integration methods (Cross-study normalisation module), by extracting gene sets expression fingerprints representing cellular pathways and processes (Gene set analysis module), and using this data for a consensus clustering or an ensemble sample classification, comparing and combining multiple feature selection and prediction methods. In spite of the complexity of such analysis pipelines, they can be configured and run using only a few mouse clicks on a unified web-interface.

The results obtained with ArrayMining can also be interlinked with other novel integrative algorithms in the framework, including the methods *TopoGSA*, *EnrichNet*, *PathExpand* and *TSPP*.

*TopoGSA* [20] complements classical data mining techniques for investigating single genes and gene sets in experimental data by a topological analysis in molecular interaction networks. On the corresponding web-application, arbitrary gene sets for different model organisms (human, yeast, plant, worm and fly), e.g. sets of differentially expressed genes obtained from ArrayMining, can be mapped onto a large-scale protein interaction network (or other user-defined networks) and characterised with regard to several topological properties. This analysis enables the identification of genes with outstanding topological characteristics or the discovery of new discriminative patterns between genes which are up- or down-regulated in different sample classes. Moreover, uploaded gene sets can be compared to entire databases of pre-defined gene sets (KEGG, BioCarta, GO, InterPro and MetaCyc) representing cellular pathways and processes, e.g. to find similarities between experimentally derived and known disease-related gene sets.

The rich information content in molecular interaction networks is also exploited by the novel analysis methods *EnrichNet*, *PathExpand* and *TSPP*, included as modular components in the framework. While the web-application *EnrichNet* [21] uses network distance information and sub-network visualisations to improve the sensitivity and interpretability of gene set enrichment analysis results, *PathExpand* [22] goes one step further and extends the definitions of pathway-representing gene sets using a graph-theoretic analysis of sub-networks. By adding new genes/proteins that are tightly interconnected with a pathway to the corresponding pathway gene set, improved representations of cellular processes in terms of network compactness and connectivity are obtained, which can be used as input for pathway-based bioinformatics data analysis methods.

The *TSPP* algorithm [23] builds on both gene set definitions and molecular interactions as data sources in order to improve the supervised analysis of microarray data. More specifically, it complements classical single-gene based machine learning methods by a robust method to learn easily interpretable decision rules on the differential relation of pairs of cellular pathways (represented by gene set expression fingerprints). In addition to the sample classification results and ranking of differentially regulated pathway pairs provided by this algorithm, the top-ranked pathway pairs can be mapped onto molecular interaction data to identify subsets of significantly associated pathway-pairs, which are deregulated in certain biological conditions of interest.

Although most analysis modules in the framework are very generic and can be applied to biological data from different experimental data sources, the full benefit of emerging experimental approaches like RNA sequencing based gene expression profiling can only be exploited with dedicated analysis methods. For this purpose, the framework has been extended by the web-application RNAnalyze [33]. RNAnalyze performs a gene set analysis tailored to RNA sequencing data, exploiting the potential of multiple gene selection methods by combining their outputs into a majority vote selection, in a similar fashion as other ensemble learning methods within the framework. RNAnalyze, which is currently still an isolated analysis module

in the framework, will be interlinked with the other modules and external web-services in a similar manner as the related Gene Set Analysis module for microarray gene expression data in ArrayMining. More specifically, the user will have the possibility to forward the data to the ArrayMining Clustering and Class Assignment Analysis modules to identify sample clusters with differentially regulated gene sets and to use supervised feature selection and classification to learn predictive models from the data. Moreover, networks of gene sets with correlated expression values will be visualised on the Network Analysis module, and the identifiers for pathway-representing gene sets and single genes will be interlinked with external web-applications including GeneCards [26], DAVID [393] and the web-services for KEGG [31], Gene Ontology [30] and possibly other pathway and gene set data repositories.

Finally, in order to provide low-dimensional, visual representations of the results in all these analysis tools, the framework also contains a software package for interactive, web-based 3D visualisations of biological data, *VRMLGen* [24]. Different methods for creating 3D scatter plots, bar charts and mesh visualisations, in combination with the possibility to add biological annotations to the plots and density estimation contour surfaces, facilitate the exploration of complex data. Moreover, a more effective knowledge management can be achieved, by integrating these visual data representations into standard web-pages and directly inter-linking them with external data repositories.

Each of the above tools can be used both as a stand-alone software and in combination with other methods from the framework or several external web-portals and databases. Importantly, the methods in the framework have been interlinked in a manner that enables new types of analyses, e.g. combining an ensemble gene selection to identify differentially expressed genes (DEGs) on ArrayMining with a network topological analysis using TopoGSA to detect DEGs with outstanding topological properties in a protein interaction network. Since the module interfaces are standardised and accept gene/protein expression matrices, class labels and annotation data in a common format, the user is not confined to using the algorithms and applications in a pre-arranged manner, but can define new analysis pipelines consisting of modular combinations of framework tools and other external software.

**Comparative evaluation results:** To evaluate the benefit of ensemble, consensus and classical machine learning techniques for microarray analysis, several supervised and unsupervised methods, including novel approaches, were evaluated on real-world cancer datasets.

In summary, when comparing sample classification results for the in-house rule-based ensemble classifier BioHEL with the benchmark methods SVM, RF and PAM using external leave-one-out cross-validation, similar average accuracies were obtained on three datasets, in spite of the simplicity of BioHEL's if-then-else decision rules. The ensemble rule sets cannot only be used to obtain robust feature rankings but also to extract information on the expression value ranges of features associated with different outcome classes, which enable the user to identify putative tumour suppressor genes and oncogenes.

Moreover, the comparative evaluation showed that the application of a univariate method for feature selection equivalent to the F-score, can have advantages in comparison to combinatorial, embedded and wrapper selection approaches, when considering predictive performance and runtime efficiency in conjunction, rather than allowing significantly longer runtimes for the wrapper-based methods.

A comparative analysis of clustering methods highlighted not only the benefit of combining multiple methods into a consensus, but also showed that a higher robustness, and in terms of some validity indices also a better cluster separation, can be obtained by integrating additional annotation data into a gene-set based dimensionality reduction of the data. Although multiple biologically meaningful patterns can occur in the

data, the obtained clustering results, and in particular those obtained from consensus clustering and gene-set based dimensionality reduction, displayed a significantly higher agreement with an external, expert-based tumour grade categorisation than 10,000 random clusterings. Finally, when comparing different dimensionality reduction methods for creating 3D representations of the data (Principal Component Analysis, Independent Component Analysis, Locally Linear Embedding and Isomap), major differences were observed between the results of the different reduction methods, and overall, the Independent Component Analysis (ICA) provided the best separation of the different tumour grades.

In summary, although in accordance with the *No-free-lunch theorem* [513] no single algorithm in machine learning is superior to every other algorithm across all possible datasets, the comparative analysis shows that some methods tend to perform better across multiple diverse microarray datasets, and that ensemble and consensus techniques and the integration of additional biological information provide benefits both for supervised and unsupervised analysis of the data. Importantly, the idea of generating a consensus of multiple methods can also be applied to the validation of machine learning methods, and is particularly useful in clustering, where different cluster validity indices often provide different estimates of the optimal number of clusters. The comparative evaluation results indeed confirm that when combining multiple validity measures to a summary statistic, the obtained estimates are more robust across different clustering, standardisation and dimensionality reduction methods.

Importantly, the framework developed as part of this doctoral project allows external users to perform similar comparative analyses on their own data online, and to profit from the possibilities of combining multiple datasets and statistical techniques into an ensemble/consensus approach, an integrative cross-domain algorithm, or a new analysis pipeline.

**New biological results:** The biological results of the thesis have been discussed in detail in chapter 8, and will therefore only be summarised briefly here:

- **Discovery and validation of a novel breast cancer marker gene:** In collaboration with the Nottingham Queens Medical Centre, a gene expression analysis of 128 breast cancer samples using the ArrayMining software identified the gene *REG* as a putative marker for the discrimination between the clinically relevant categories of luminal and non-luminal samples. The utility of this marker was confirmed by a large-scale experimental validation using immunohistochemistry on tissue microarrays containing 1,140 invasive breast cancers [19].
- **Identification of candidate disease genes and improvement of disease pathway definitions:** With the PathExpand tool from the framework (see chapter 6), various disease pathways were analysed and extended by adding densely interconnected interaction partners in a large-scale protein interaction network. For an Alzheimer's pathway, this led to the discovery of two associated proteins, METTL2B and MMP17, which had previously not been linked to the neurodegenerative disease. Since several members of the corresponding protein families (metallopeptidases, methyltransferase-like proteins) are known to be associated with Alzheimer's, the two predicted pathway members were proposed as new candidate disease genes [22].

Moreover, by studying the enrichment of extended cellular pathways in genes mutated in pancreatic cancer, biologically insightful extensions of cancer-related pathways were obtained. The extended pathways were both enriched in pancreatic cancer mutated genes and had functional annotations similar to those of the original pathway members.

- **Identification of associations between pairs of cellular pathways and diseases:**

The Top-scoring pathway pairs (TSPP) method [23] identified pathway pairs that are differentially regulated across different disease conditions in microarray data. Apart from already known pathway-pair/disease associations, new functional relations were discovered and investigated in detail by mapping corresponding pathway pairs onto molecular interaction networks. For example, complementary information from the literature confirmed a proposed association between the KEGG pathways *p53 signaling pathway* and *Purine metabolism* in prostate cancer. Specifically, the inhibition of *de novo* purine synthesis was observed to inhibit prostate cancer cell growth and increase p53 expression levels [487]. Thus, in contrast to classical enrichment analysis methods for identifying functional associations between gene sets, the TSPP methods can also prioritize putative pathway-pair/disease associations and help to uncover de-regulated pathway-relations using microarray data.

- **Genome-wide transcription network analysis in seeds:**

Although bioinformatics methods for large-scale biological data analysis have many of their most important practical applications in biomedicine, a multitude of other applications exist in biotechnology and agriculture. To analyse the gene regulation in *Arabidopsis thaliana*, a model organism for crop plants, microarray samples representing the temporal conditions of dormancy and germination were studied in co-operation with the Division of Plant and Crop Sciences at Nottingham University. Using the ArrayMining network analysis module, a condition-specific, weighted gene co-expression network was constructed from the data, and an interactive network visualisation, SeedNet, was created and published online [35]. The network analysis identified two distinct modules of state-dependent genetic interactions reflecting the two different biological states and enabling the topology-based prediction of germination regulators with higher accuracy than previous methods. As an online gene regulation data repository and network visualisation tool, SeedNet will help to further improve the understanding of the gene regulatory network influencing the conditions of dormancy and germination, facilitating research aimed at improving seed performance in agriculture.

## 10.2 Outlook on Future Work

The framework for integrative biological data analysis presented in this dissertation was designed to cover some of the most wide-spread experimental techniques in functional genomics, and to be directly applicable to real-world bioscientific and biomedical problems, as illustrated by the wide range of collaborative projects employing the framework tools.

Prior to this doctoral project and during its implementation, other public software frameworks for analysing gene and protein microarray data and general gene and protein lists already provided access to a multitude of different analysis methods, but these tools were not interlinked to exploit the synergies of different analysis types like machine learning, network analysis and gene set based dimensionality reduction, or only enabled a fixed sequential combination of known analysis techniques. These existing analysis frameworks include *GEPAS* [403], *Expression Profiler* [404], *ASTERIAS* [405], *EzArray* [406], *CARMAweb* [407], *MAGMA* [408], *ArrayPipe* [409], *RACE* [410], *WebArray* [411] and *MIDAW* [412], as well as more generally applicable statistical programming languages like R [353] and Matlab [402]. While the dedicated analysis frameworks are easier to use in comparison to the statistical programming languages and contain both pre-implemented methods for data pre-processing (e.g. single-study normalisation, missing value

imputation and gene filtering) and higher level analysis (e.g. clustering, gene selection and annotation, classification, data visualisation and gene set enrichment analysis), they are less flexible and do not enable the user to set up complex analysis pipelines and exploit recent algorithmic advances provided by ensemble and consensus analysis techniques. The software framework presented here addresses these limitations by allowing users to benefit from the possibility to freely combine algorithms in parallel (using ensemble/consensus techniques) and sequentially (using modular combinations) to build new analysis pipelines, while keeping the interface simple and easy to use by applying automatic parameter selection mechanisms. Additionally, in contrast to existing integrative analysis frameworks, different analysis types are not just combined using already existing, domain-specific algorithms, but new cross-domain data analysis methods, operating on multiple, diverse data sources at the same time, have been developed and interlinked with other modules in the framework. Apart from their statistical advantages in terms of robustness and accuracy, the main benefit of these new integrative analysis techniques is that they provide new biological insights by enabling an interpretation of the data from a different perspective, e.g. by identifying associations between pathway deregulations and different disease states in microarray-based machine learning models (TSSP method) instead of only scoring the importance of single genes, by measuring the similarity of gene/protein sets in terms of topological properties (TopoGSA) and network distances (EnrichNet) rather than computing overlap-based enrichment scores (like the existing tools DAVID [393] and FatiGO [514], among others), and by re-defining cellular pathways to provide dense and compact network representations in molecular interaction networks (PathExpand). Importantly, both highly specialised algorithms and more general integrative analysis methods are required to fully exploit the information content in functional genomics datasets and ideally, these different types of approaches should also be interlinked. Thus, future versions of the integrative framework might also profit from the data exchange with recently developed specialized analysis tools, e.g. for biclustering of genes and samples [413], co-clustering of genes with similar functional annotations [414], inference of gene regulatory relationships [415] and cross-species clustering [416]. Although the framework already covers a wide range of data types, including gene/protein expression data, cellular pathway definitions, and many of the machine learning techniques in the framework are applicable to even further biological data sources (e.g. mass spectrometry data, nucleotide and amino acid sequence data, protein structural data, array comparative genomic hybridisation data (aCGH), clinical measurements, etc.), the full information content in other data sources can only be exploited by additionally considering dedicated analysis methods, adjusted to the specific properties of these data types (e.g. the types of noise sources, data distributions, and the feature and sample sizes).

Therefore, the framework presented here is designed to be modular and easily extensible by new analysis techniques for other types of input data. Specifically, a common, generic template is used for the web-interface, the storage and forwarding of the data, and the representation of the results in sortable tables and 2D and 3D visualisations.

However, before adding new analysis modules to the framework, future extensions could, in a first step, include mappings of other data types to the currently used gene and protein data, e.g. mapping the genes from a differential expression analysis onto gene mutation, alternative splicing, epigenetics and RNA/DNA and microRNA-mRNA interaction databases, or mapping proteins onto protein structural, protein-domain and functional annotation databases. These mappings would enable the user to investigate the results from the existing analysis modules in more detail using other interlinked web-services, e.g. to identify whether a differentially expressed gene in a cancer dataset is also known to have a hypermethylated promotor region in certain cancers or to contain mutations driving the tumour growth.

In a second extension step, new data types and analysis techniques could directly be integrated into the

framework. However, these extensions should not exceed the scope of the framework, and focus on wide-spread experimental methodologies in functional genomics, benefiting from synergies with already included data sources and methods. An important biological data type, which could be targeted by more dedicated analysis techniques to increase the utility of the framework, is mass spectrometry data, which can partly already be analysed using the current machine learning modules in ArrayMining, if the data has already been processed. Accordingly, by including mass spectrometry specific pre-processing steps into ArrayMining (i.e. raw data filtering, peak detection, peak grouping/alignment and retention time normalisation methods), the existing machine learning analysis modules could be used for a greater variety of purposes.


Alternative data sources, which are particularly useful for integrative analysis purposes, also include different types of systems biology network data, e.g. gene regulatory networks, protein-DNA interaction networks, protein domain networks, gene/protein functional association networks [439] and disease networks [515]. Moreover, in particular for microarray cancer data, the consideration of classical clinical measurements (e.g. tumour size, histological grade, status of angiogenesis, and status of lymphocytic infiltration, oestrogen receptor and progesterone receptor parameters) would provide a valuable extension of the framework. While additional annotation data for the *features* of a dataset can be integrated into an analysis using an approach that is also employed for gene set analysis, i.e. defining “meta-features” corresponding to summarised fingerprints of features with similar annotation data (see chapter 3 and 4), additional labels or numerical data for the *samples* can be used as additional outcome variables in a supervised analysis or as weight vectors within ensemble learning techniques, e.g. Feature Weighted Linear Stacking (FWLS) [516].

The focus on transcriptomics, proteomics and metabolomics data analysis could also be broadened by directly integrating analysis methods for further “omics” data types into the framework, e.g. to find homologous genes with known functions for differentially expressed genes with missing annotations (*genomics*), or to detect genes that are alternatively spliced under different conditions (*spliceomics*), as well as mutated and epigenetically silenced genes (*mutagenomics* and *epigenomics*). Importantly, the goal would not be to add already existing tools and web-services to the framework, but to develop new integrative data mining approaches that exploit the information sources and analysis types from different biological domains, e.g. by combining machine learning models trained on independent data sources to a meta-model, or using features from diverse datasets within a single model.

While this task of better exploiting the synergies of diverse “omics” data types is rather a long-term goal, improvements of the currently used ensemble learning techniques and metaheuristics for search space exploration specific to single data types can be achieved more quickly. For example, alternative evolutionary learning operators and niching techniques could be integrated into the existing optimisation methods (e.g. within the TSPP algorithm or the consensus clustering method), as well as alternative boosting and ensemble learning methods like variants of the *stacked generalisation* method [516] to combine different machine learning models. Moreover, some of the current implementations could be parallelised to increase the runtime (however, most algorithms in the framework already terminate in few minutes rather than hours or days).

In conclusion, the current integrative analysis framework already covers a wide range of data analysis problems in transcriptomics and proteomics, but also provides many opportunities for further extensions, both by interlinking the framework with external tools and web-services and by integrating analysis methods for other functional genomics data types to exploit synergies between different “-omics” disciplines.

As new generations of high-throughput experimental methodologies are becoming more sensitive, cheaper and more widespread in the academic community, integrative analysis methods will be more frequently



required to combine the incomplete information from different sources to obtain a more coherent picture of the biological systems of interest. Thus, many new opportunities are likely to emerge for bioinformatics data and algorithm integration approaches to support systems biology studies unravelling the processes taking place in different diseases, biotechnologically relevant microorganisms and agriculturally important plants.

The integrative framework presented here illustrates how ensemble, consensus and cross-domain integrative methods can provide novel biological insights, which are often not obtained when using domain-specific approaches and considering single algorithms and datasets independently. The promising results achieved so far and the almost infinite reservoir of unsolved problems in biology highlight the opportunities for integrative bioinformatics data analyses. The framework built in this doctoral project can therefore be useful both to tackle further specific biological problems and as a template for the development of new integrative analysis methods in the future.

# Bibliography

- [1] Bellman R. *Adaptive Control Processes*. Princeton University Press, NJ, USA, 1961.
- [2] Miller Jr R. *Simultaneous Statistical Inference*. McGraw-Hill Book Co., New York, NY, USA, 1966.
- [3] Benjamini Y and Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statist Soc Ser B Methodological*, 57(1):289, 1995.
- [4] Ding C and Peng H. Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol*, 3(2):185, 2005.
- [5] Yu L and Liu H. Redundancy based feature selection for microarray data. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 737–742. ACM New York, NY, USA, 2004.
- [6] Hastie T, Tibshirani R, and Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY, USA, 2001.
- [7] Braga-Neto U and Dougherty E. Bolstered error estimation. *Pattern Recognition*, 37(6):1267, 2004.
- [8] Dietterich TG. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pages 1–15. Springer-Verlag Berlin, Heidelberg, 2000.
- [9] Saeyns Y, Abeel T, and Peer Y. Robust Feature Selection Using Ensemble Feature Selection Techniques. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases-Part II*, pages 313–325. Springer-Verlag Berlin, Heidelberg, 2008.
- [10] Tan A and Gilbert D. Ensemble machine learning on gene expression data for cancer classification. *Appl Bioinformatics*, 2(3 Suppl):S75, 2003.
- [11] Monti S, Tamayo P, Mesirov J, and Golub T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52(1):91, 2003.
- [12] Swift S, Tucker A, Vinciotti V, Martin N, et al. Consensus clustering and functional interpretation of gene-expression data. *Genome Biol*, 5(11):R94, 2004.
- [13] Bacardit J and Krasnogor N. Empirical evaluation of ensemble techniques for a Pittsburgh Learning Classifier System. In *Revised Selected Papers of the 10th International Workshop on Learning Classifier Systems 2006*, pages 255–268. Springer-Verlag Berlin, Heidelberg, 2008.
- [14] Shabalina AA, Tjelmeland H, Fan C, Perou CM, et al. Merging two gene-expression studies via cross-platform normalization. *Bioinformatics*, 24(9):1154, 2008.
- [15] Wamat P, Eils R, and Brors B. Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics*, 6(1):265, 2005.
- [16] Walker W, Liao I, Gilbert D, Wong B, et al. Empirical Bayes accommodation of batch-effects in microarray data using identical replicate reference samples: application to RNA expression profiling of blood from Duchenne muscular dystrophy patients. *BMC Genomics*, 9(1):494, 2008.
- [17] Cope L, Garrett-Mayer E, Gabrielson E, and Parmigiani G. The integrative correlation coefficient: a measure of cross-study reproducibility for gene expression array data. *Johns Hopkins University Dept of Biostatistics Working Papers*, page 152, 2007.
- [18] Glaab E, Garibaldi J, and Krasnogor N. ArrayMining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization. *BMC Bioinformatics*, 10(1):358, 2009.
- [19] Habashy HO, Powe DG, Glaab E, Krasnogor N, et al. RERG (Ras-related and oestrogen-regulated growth-inhibitor) expression in breast cancer: A marker of ER-positive luminal-like subtype. *Breast Cancer Res Treat*, 128(2):315, 2011.
- [20] Glaab E, Baudot A, Krasnogor N, and Valencia A. TopoGSA: network topological gene set analysis. *Bioinformatics*, 26(9):1271, 2010.
- [21] EnrichNet: Network-based gene set enrichment analysis, 2010. <http://www.infobiotics.net/enrichnet>.
- [22] Glaab E, Baudot A, Krasnogor N, and Valencia A. Extending pathways and processes using molecular interaction networks to analyse cancer genome data. *BMC Bioinformatics RECOMB Computational Cancer Biology 2010*, 11(1):597, 2010.
- [23] Glaab E, Garibaldi J, and Krasnogor N. Learning pathway-based decision rules to classify microarray cancer samples. In D Schomburg and A Grote, editors, *German Conference on Bioinformatics 2010*, volume 173 of *Lecture Notes in Informatics*, pages 123–134. Gesellschaft fuer Informatik, 2010.
- [24] Glaab E, Garibaldi J, and Krasnogor N. vrm1gen: An R Package for 3D Data Visualization on the Web. *J Stat Software*, 36(8):1, 2010.



- [25] Wood I, Visscher P, and Mengersen K. Classification based upon gene expression data: bias and precision of error rates. *Bioinformatics*, 23(11):1363, 2007.
- [26] Rebhan M, Chalifa-Caspi V, Prilusky J, and Lancet D. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, 14(8):656, 1998.
- [27] Edgar R, Domrachev M, and Lash A. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 30(1):207, 2002.
- [28] Hubbard TJ, Aken BL, Ayling S, Ballester B, et al. Ensembl 2009. *Nucleic Acids Res*, 37(Database issue):D690, 2009.
- [29] Dennis Jr G, Sherman B, Hosack D, Yang J, et al. DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol*, 4(5):P3, 2003.
- [30] Ashburner M, Ball C, Blake J, Botstein D, et al. Gene Ontology: tool for the unification of biology. *Nat Genet*, 25(1):25, 2000.
- [31] Kanehisa M and Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27, 2000.
- [32] Nishimura D. BioCarta. *Biotech Software Internet Report*, 2(3):117, 2001.
- [33] RNAalyze: RNA sequencing data enrichment analysis, 2010. <http://bree.cs.nott.ac.uk/R-php-1/RNAseq>.
- [34] Futreal P, Coin L, Marshall M, Down T, et al. A census of human cancer genes. *Nat Rev Cancer*, 4(3):177, 2004.
- [35] Bassel GW, Lan H, Glaab E, Gibbs DJ, et al. A genome-wide network model capturing seed germination reveals co-ordinated regulation of plant cellular phase transitions. *Proc Natl Acad Sci USA*, 108(23):9709, 2011. <http://vseed.nottingham.ac.uk>.
- [36] Kanehisa M et al. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, 34(Database Issue):D354, 2006.
- [37] Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, et al. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*, 33(Database Issue):D428, 2005.
- [38] U.S. National Institute of Cancer. A to Z List of Cancers, 2011. <http://www.cancer.gov/cancertopics/types/alphalist>.
- [39] World Health Organization. Cancer, 2007. <http://www.who.int/mediacentre/factsheets/fs297/en>.
- [40] Cancer Research UK. Cancer incidence by age - UK statistics, 2011. <http://info.cancerresearchuk.org/cancerstats/incidence/age>.
- [41] Reik W and Walter J. Genomic imprinting: parental influence on the genome. *Nat Rev Genet*, 2(1):21, 2001.
- [42] Tayles N. Anemia, genetic diseases, and malaria in prehistoric mainland Southeast Asia. *Am J Phys Anthropol*, 101(1):11, 1996.
- [43] Allison A. Protection afforded by sickle-cell trait against subtertian malarial infection. *Br Med J*, 1(4857):290, 1954.
- [44] Hesdorffer C, Markowitz D, Ward M, and Bank A. Somatic gene therapy. *Hematol Oncol Clin North Am*, 5(3):423, 1991.
- [45] Velázquez F, Matson D, Guerrero M, Shults J, et al. Serum antibody as a marker of protection against natural rotavirus infection and disease. *J Infect Dis*, 182(6):1602, 2000.
- [46] Fahey J, Taylor J, Detels R, Hofmann B, et al. The prognostic value of cellular and serologic markers in infection with human immunodeficiency virus type 1. *N Engl J Med*, 322(3):166, 1990.
- [47] Chen P, Ratcliff G, Belle S, Cauley J, et al. Cognitive tests that best discriminate between presymptomatic AD and those who remain nondemented. *Neurology*, 55(12):1847, 2000.
- [48] Edström G. Rheumatoid arthritis and Still's disease in children a survey of 161 cases. *Arthritis Rheumatism*, 1(6):497, 1958.
- [49] Pyorala K, Backer G, Graham I, Goole-Wilson P, et al. Prevention of coronary heart disease in clinical practice. In *Annales de Cardiologie et d'Angéiologie*, volume 44, pages 379–388. Paris, L'Expansion scientifique française, 1995.
- [50] Aronson J. Biomarkers and surrogate endpoints. *Br J Clin Pharmacol*, 59(5):491, 2005.
- [51] Colburn W and Lee J. Biomarkers, validation and pharmacokinetic-pharmacodynamic modelling. *Clin Pharmacokinet*, 42(12):997, 2003.
- [52] Rarey M, Kramer B, Lengauer T, and Klebe G. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol*, 261(3):470, 1996.
- [53] Rarey M and Dixon J. Feature trees: a new molecular similarity measure based on tree matching. *J Comput Aided Mol Des*, 12(5):471, 1998.
- [54] Lemmen C, Lengauer T, and Klebe G. FLEXS: a method for fast flexible ligand superposition. *J Med Chem*, 41(23):4502, 1998.
- [55] Kubinyi H. *QSAR: Hansch analysis and related approaches*. Wiley-VCH, Weinheim, 1993.
- [56] Kim K, Greco G, and Novellino E. A critical review of recent CoMFA applications. *Perspect Drug Discov Des*, 12:257, 1998.
- [57] Shaw K, Woods C, and Mulholland A. *QM and QMMM Approaches to Evaluating Binding Affinities*. Wiley-VCH, Weinheim, 2008.
- [58] Pusztai L and Leyland-Jones B. Promises and caveats of in silico biomarker discovery. *Br J Cancer*, 99(3):385, 2008.
- [59] Dearden J. In silico prediction of ADMET properties: how far have we come? *Expert Opin Drug Metab Toxicol*, 3(5):635, 2007.
- [60] Beard D and Kushmerick M. Strong inference for systems biology. *PLoS Comput Biol*, 5(8):347, 2009.

- [61] van't Veer L, Dai H, van de Vijver M, and He Y. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530, 2002.
- [62] Heid C, Stevens J, Livak K, and Williams P. Real time quantitative PCR. *Genome Res*, 6(10):986, 1996.
- [63] Prader A, Labhart A, and Willi H. Ein Syndrom von Adipositas, Kleinwuchs, Kryptorchismus und Oligophrenie nach myatonieartigem Zustand im Neugeborenenalter. *Schweiz Med Wochenschr*, 86:1260, 1956.
- [64] Angelman H. Puppet Children A Report on Three Cases. *Dev Med Child Neurol*, 7(6):681, 1965.
- [65] Lichtenstein P, Holm N, Verkasalo P, Iliadou A, et al. Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med*, 343(2):78, 2000.
- [66] Crow J. The origins, patterns and implications of human spontaneous mutation. *Nat Rev Genet*, 1(1):40, 2000.
- [67] Bozinov D and Rahnenführer J. Unsupervised technique for robust target separation and analysis of DNA microarray spots through adaptive pixel clustering, 2002.
- [68] Ahmed A, Vias M, Iyer N, Caldas C, et al. Microarray segmentation methods significantly influence data precision. *Nucleic Acids Res*, 32(5):e50, 2004.
- [69] Bolstad B, Irizarry R, Astrand M, and Speed T. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185, 2003.
- [70] Smyth G. *Limma: linear models for microarray data*. Springer-Verlag Berlin, Heidelberg, 2005.
- [71] McClure J and Wit E. Post-normalization quality assessment visualization of microarray data. *Comp Funct Genomics*, 4:460, 2003.
- [72] Piper M, Daran-Lapujade P, Bro C, Regenber B, et al. Reproducibility of oligonucleotide microarray transcriptome analyses. *J Biol Chem*, 277(40):37001, 2002.
- [73] Szabo A, Perou C, Karaca M, Perreard L, et al. Statistical modeling for selecting housekeeper genes. *Genome Biol*, 5(8):R59, 2004.
- [74] Fan J and Niu Y. Selection and validation of normalization methods for c-DNA microarrays using within-array replications. *Bioinformatics*, 23(18):2391, 2007.
- [75] Cleveland W. LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *Am Stat*, 35, 1981.
- [76] Yang Y, Dudoit S, Luu P, Lin D, et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, 30(4):e15, 2002.
- [77] Buness A, Huber W, Steiner K, Sültmann H, et al. arrayMagic: two-colour cDNA microarray quality control and preprocessing. *Bioinformatics*, 21(4):554, 2005.
- [78] Workman C, Jensen L, Jarmer H, Berka R, et al. A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol*, 3:1, 2002.
- [79] Davis P. B-splines and geometric design. *SIAM News*, 29(5), 1996.
- [80] Huber W, von Heydebreck A, Sültmann H, Poustka A, et al. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18(1):96, 2002.
- [81] Durbin B, Hardin J, Hawkins D, and Rocke D. A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, 18:105, 2002.
- [82] Wu Z and Irizarry R. Pre-processing of oligonucleotide array data. *Nat Biotechnol*, 22(6):656, 2004.
- [83] Turro E, Bochkina N, Hein A, and Richardson S. BGX: a Bioconductor package for the Bayesian integrated analysis of Affymetrix GeneChips. *BMC Bioinformatics*, 8(1):439, 2007.
- [84] Pearson R, Liu X, Sanguinetti G, Milo M, et al. puma: a Bioconductor package for propagating uncertainty in microarray analysis. *BMC Bioinformatics*, 10(1):211, 2009.
- [85] Ono N, Suzuki S, Furusawa C, Agata T, et al. An improved physico-chemical model of hybridization on high-density oligonucleotide microarrays. *Bioinformatics*, 24(10):1278, May 2008.
- [86] Quackenbush J. Microarray data normalization and transformation. *Nat Genet*, 32:496, 2002.
- [87] Hoaglin D, Mosteller F, and Tukey J. *Understanding robust and exploratory data analysis*. Wiley, New York, NJ, USA, 1983.
- [88] Li C and Wong W. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc Natl Acad Sci USA*, 98(1):31, 2001.
- [89] Lazaridis E, Sinibaldi D, Bloom G, Mane S, et al. A simple method to improve probe set estimates from oligonucleotide arrays. *Math Biosci*, 176(1):53, 2002.
- [90] Bolstad B. *Low-level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization*. Ph.D. thesis, University of California, 2004.
- [91] Tukey J and Mosteller F. *Data Analysis and Regression*. Addison Wesley Reading MA USA, 1977.
- [92] Stark C, Breitkreutz B, Reguly T, Boucher L, et al. BioGRID: a general repository for interaction datasets. *Nucleic acids research*, 34(Database Issue):D535, 2006.
- [93] Mewes H, Heumann K, Kaps A, Mayer K, et al. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res*, 27(1):44, 1999.
- [94] Xenarios I, Salwinski L, Duan X, Higney P, et al. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res*, 30(1):303, 2002.

- [95] Chattr-Aryamontri A, Ceol A, Palazzi L, Nardelli G, et al. MINT: a Molecular INTERaction database. *Nucleic Acids Res*, 35(Database Issue):D572, 2007.
- [96] Peri S, Navarro J, Amanchy R, Kristiansen T, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*, 13(10):2363, 2003.
- [97] Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, et al. IntAct: an open source molecular interaction database. *Nucleic Acids Res*, 32(Database Issue):D452, 2004.
- [98] von Mering C, Krause R, Snel B, Cornell M, et al. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, 417(6887):399, 2002.
- [99] Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, et al. The HUPO PSI's molecular interaction format community standard for the representation of protein interaction data. *Nat Biotechnol*, 22(2):177, 2004.
- [100] Harris M, Clark J, Ireland A, Lomax J, et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*, 32(Database issue):D258, 2004.
- [101] Jain S and Bader G. An improved method for scoring protein-protein interactions using semantic similarity within the Gene Ontology. *BMC Bioinformatics*, 11(1):562, 2010.
- [102] Resnik P. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J Artif Intell Res*, 11(95):130, 1999.
- [103] Jiang JJ and Conrath DW. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *International Conference Research on Computational Linguistics*, pages 19–33. September 1997.
- [104] Dandekar T, Snel B, Huynen M, and Bork P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci*, 23(9):324, 1998.
- [105] Overbeek R, Fonstein M, D'Souza M, Pusch G, et al. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci USA*, 96(6):2896, 1999.
- [106] Enright A, Iliopoulos I, Kyrpides N, and Ouzounis C. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402(6757):86, 1999.
- [107] Marcotte E, Pellegrini M, Ng H, Rice D, et al. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428):751, 1999.
- [108] Suhre K. Inference of Gene Function Based on Gene Fusion Events: The Rosetta-Stone Method. *Methods Mol Biol*, 396:31, 2007.
- [109] Matthews LR, Vaglio P, Reboul J, Ge H, et al. Identification of Potential Interaction Networks Using Sequence-Based Searches for Conserved Protein-Protein Interactions or "Interologs". *Genome Res*, 11(12):2120, December 2001.
- [110] Pellegrini M, Marcotte E, Thompson M, Eisenberg D, et al. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA*, 96(8):4285, 1999.
- [111] Marcotte E, Xenarios I, Van Der Bliek A, and Eisenberg D. Localizing proteins in the cell from their phylogenetic profiles. *Proc Natl Acad Sci USA*, 97(22):12115, 2000.
- [112] Bhardwaj N and Lu H. Correlation between gene expression profiles and protein–protein interactions within and across genomes. *Bioinformatics*, 21(11):2730, 2005.
- [113] Huh W, Falvo J, Gerke L, Carroll A, et al. Global analysis of protein localization in budding yeast. *Nature*, 425(6959):686, 2003.
- [114] Pazos F and Valencia A. In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins Struct Funct Bioinf*, 47(2):219, 2002.
- [115] de Lichtenberg U, Jensen L, Brunak S, and Bork P. Dynamic complex formation during the yeast cell cycle. *Science*, 307(5710):724, 2005.
- [116] Chagoyen M, Carmona-Saez P, Gil C, Carazo J, et al. A literature-based similarity metric for biological processes. *BMC Bioinformatics*, 7(1):363, 2006.
- [117] Liu G, Li J, and Wong L. Assessing and predicting protein interactions using both local and global network topological metrics. *Genome Inform*, 21:253, 2008.
- [118] Topinka C and Shyu C. Predicting cancer interaction networks using text-mining and structure understanding. In *AMIA Annual Symposium Proceedings*, volume 2006, page 1123. American Medical Informatics Association, 2006.
- [119] Lage K, Karlberg E, Størling Z, Ólason P, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol*, 25(3):309, 2007.
- [120] Singh R, Park D, Xu J, Hosur R, et al. Struct2Net: a web service to predict protein–protein interactions using a structure-based approach. *Nucleic Acids Res*, 38(Web Server issue):W508, 2010.
- [121] Skrabanek L, Saini H, Bader G, and Enright A. Computational prediction of protein–protein interactions. *Mol Biotechnol*, 38(1):1, 2008.
- [122] Barabási A and Albert R. Emergence of scaling in random networks. *Science*, 286(5439):509, 1999.
- [123] Chen L, Liu H, and Friedman C. Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*, 21(2):248, 2005.
- [124] Maglott D, Ostell J, Pruitt KD, and Tatusova T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*, 35(Database issue), 2007.

- [125] Alibés A, Yankilevich P, et al. IDconverter and IDClight: Conversion and annotation of gene and protein IDs. *BMC Bioinformatics*, 8(1):9, 2007.
- [126] Bussey K, Kane D, Sunshine M, Narasimhan S, et al. MatchMiner: a tool for batch navigation among gene and gene product identifiers. *Genome Biol*, 4(4):R27, 2003.
- [127] Durinck S, Spellman P, Birney E, and Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc*, 4(8):1184, 2009.
- [128] Durinck S, Moreau Y, Kasprzyk A, Davis S, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439, 2005.
- [129] Fang H, Murphy K, Jin Y, Kim J, et al. Human gene name normalization using text matching with automatically extracted synonym dictionaries. In *Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis*, pages 41–48. Association for Computational Linguistics, 2006.
- [130] Schuemie M, Jelier R, and Kors J. Peregrine: Lightweight gene name normalization by dictionary lookup. In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, pages 131–133, 2007.
- [131] Wermter J, Tomanek K, and Hahn U. High-performance gene name normalization with GeNo. *Bioinformatics*, 25(6):815, 2009.
- [132] Morgan A, Lu Z, Wang X, Cohen A, et al. Overview of BioCreative II gene normalization. *Genome Biol*, 9(Suppl 2):S3, 2008.
- [133] Hakenberg J, Plake C, Royer L, Strobelt H, et al. Gene mention normalization and interaction extraction with context models and sentence motifs. *Genome Biol*, 9(Suppl 2):S14, 2008.
- [134] Hirschman L, Yeh A, Blaschke C, and Valencia A. Overview of BioCreative: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6(Suppl 1):S1, 2005.
- [135] Neves M, Carazo J, and Pascual-Montano A. Moara: a Java library for extracting and normalizing gene and protein mentions. *BMC Bioinformatics*, 11(1):157, 2010.
- [136] Crim J, McDonald R, and Pereira F. Automatically annotating documents with normalized gene lists. *BMC Bioinformatics*, 6(Suppl 1):S13, 2005.
- [137] McDonald R and Pereira F. Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinformatics*, 6(Suppl 1):S6, 2005.
- [138] Cohen A. Unsupervised gene/protein named entity normalization using automatically extracted dictionaries. In *Proceedings of the ACL-ISMB workshop on linking biological literature, ontologies and databases: Mining biological semantics*, pages 17–24. Association for Computational Linguistics, 2005.
- [139] Hanisch D, Fundel K, Mevissen H, Zimmer R, et al. ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics*, 6(Suppl 1):S14, 2005.
- [140] Torii M, Hu Z, Wu C, and Liu H. BioTagger-GM: a gene/protein name recognition system. *J Am Med Inform Assoc*, 16(2):247, 2009.
- [141] Lowe H and Barnett G. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *JAMA*, 271(14):1103, 1994.
- [142] Hastie T, Tibshirani R, and Walther G. Estimating the number of data clusters via the Gap statistic. *J Roy Stat Soc B*, 63:411, 2001.
- [143] van't Veer et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530, 2002.
- [144] Ma S, Song X, and Huang J. Supervised group Lasso with applications to microarray data analysis. *BMC Bioinformatics*, 8(1):60, 2007.
- [145] Tritchler D, Parkhomenko E, and Beyene J. Filtering genes for cluster and network analysis. *BMC Bioinformatics*, 10(1):193, 2009.
- [146] Zou H, Hastie T, and Tibshirani R. Sparse principal component analysis. *J Comp Graph Stat*, 15(2):265, 2006.
- [147] Blumer A, Ehrenfeucht A, Haussler D, and Warmuth M. Occam's razor. *Inf Process Lett*, 24(6):377, 1987.
- [148] Peterson D and Thaut M. Model and feature selection in microarray classification. *2004 CIBCB04 Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pages 56–60, 2004.
- [149] Cilibrasi R and Vitanyi P. Clustering by compression. *IEEE T Inform Theory*, 51(4):1523, 2005.
- [150] Lönnstedt I and Speed T. Replicated microarray data. *Stat Sin*, 12(1):31, 2002.
- [151] Smyth G. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3(1):3, 2004.
- [152] Bonferroni C. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8, 1936.
- [153] Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65, 1979.
- [154] Benjamini Y and Yekutieli D. The control of the false discovery rate in multiple testing under dependency. *Ann Stat*, pages 1165–1188, 2001.
- [155] Cortes C and Vapnik V. Support-vector networks. *Machine learning*, 20(3):273, 1995.
- [156] Kononenko I. Estimating attributes: Analysis and extensions of RELIEF. In *Machine Learning*, pages 171–182. Springer-Verlag Berlin, Heidelberg, 1994.

- [157] Kira K and Rendell L. The Feature Selection Problem: Traditional Methods and a New Algorithm. *Proceedings of the Tenth National Conference on Artificial Intelligence*, 1992.
- [158] Robnik-Šikonja M and Kononenko I. Theoretical and empirical analysis of ReliefF and RReliefF. *Machine learning*, 53(1):23, 2003.
- [159] Sun Y, Goodison S, Li J, Liu L, et al. Improved breast cancer prognosis through the combination of clinical and genetic markers. *Bioinformatics*, 23(1):30, 2007.
- [160] Hall M. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. In *Proc. Int. Conf. Mach. Learn.*, pages 359–366. Morgan Kaufmann, San Francisco, CA, 2000.
- [161] Hall M. Correlation-based feature selection for discrete and numeric class machine learning. *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 359–366, 2000.
- [162] Wu Y and Zhang A. Feature selection for classifying high-dimensional numerical data. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 251–258, 2004.
- [163] Zhang J and Deng H. Gene selection for classification of microarray data based on the Bayes error. *BMC Bioinformatics*, 8(1):370, 2007.
- [164] Fukunaga K and Kessell D. Nonparametric Bayes error estimation using unclassified samples. *IEEE T Inform Theory*, 19(4):434, 2002.
- [165] Hellman M and Raviv J. Probability of error, equivocation, and the Chernoff bound. *IEEE T Inform Theory*, 16(4):368, 1970.
- [166] Fano R. *Transmission of information*. M.I.T. Press New York, 1961.
- [167] Guo S, Lyu M, Lok T, and Kong H. Gene Selection Based on Mutual Information for the Classification of Multi-class Cancer. *Lect Notes Comput Sci*, 4115:454, 2006.
- [168] Witten I and Frank E. *Data Mining: Practical Machine Learning Tools and Techniques*. 2005.
- [169] Li M and Vitanyi P. *An introduction to Kolmogorov complexity and its applications*. Springer-Verlag, New York, NY, USA, 2008.
- [170] Vitanyi P, Balbach F, Cilibrasi R, and Li M. Normalized information distance. *Inform Theory Stat Learn*, pages 45–82, 2009.
- [171] Koller D and Sahami M. Toward optimal feature selection. In *Proc. Int. Conf. Mach. Learn.*, pages 284–292, 1996.
- [172] Subramani P, Sahu R, and Verma S. Feature selection using Haar wavelet power spectrum. *BMC Bioinformatics*, 7(1):432, 2006.
- [173] Liu Y. Wavelet feature selection for microarray data. In *Proceedings of the IEEE/NIH on Life Science Systems and Applications Workshop*, pages 205–208. Bethesda, MD, USA, 2007.
- [174] Nanni L and Lumini A. Wavelet selection for disease classification by dna microarray data. *Expert Syst Appl*, 38:990, January 2011.
- [175] Pudil P, Novovicová J, and Kittler J. Floating search methods in feature selection. *Pattern Recogn Lett*, 15(11):1119, 1994.
- [176] Skurichina M and Duin R. Combining Feature Subsets in Feature Selection. *Proceedings of the 6th International Workshop on Multiple Classifier Systems*, pages 165–175, 2005.
- [177] Deutsch JM. Evolutionary algorithms for finding optimal gene sets in microarray prediction. *Bioinformatics*, 19(1):45, 2003.
- [178] Freitas A. The principle of transformation between efficiency and effectiveness: towards a fair evaluation of the cost-effectiveness of KDD techniques. In *Proceedings of the First European Symposium on Principles of Data Mining and Knowledge Discovery*, pages 299–306. Springer-Verlag Berlin, Heidelberg, 1997.
- [179] Inza I. Gene selection by sequential search wrapper approaches in microarray cancer class prediction. *J Intelligent and Fuzzy Systems*, 12:25, 2002.
- [180] Li L, Pedersen L, Darden T, and Weinberg C. Computational Analysis of Leukemia Microarray Expression Data Using the GA/KNN Method. In *Methods of Microarray Data Analysis*, pages 81–95. Kluwer Academic Publishers, 2002.
- [181] Golub T, Slonim D, Tamayo P, Huard C, et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286:531, 1999.
- [182] Blanco R, Larranaga P, Inza I, and Sierra B. Selection of highly accurate genes for cancer classification by estimation of distribution algorithms. *Workshop of Bayesian Models in Medicine*, pages 29–34, 2001.
- [183] Bishop C and Nasrabadi N. *Pattern Recognition and Machine Learning*. *Journal of Electronic Imaging*, 16(4):9901, 2007.
- [184] Tibshirani R. Regression shrinkage and selection via the lasso. *J Roy Statist Soc Ser B*, 58:267, 1996.
- [185] Osborne M, Presnell B, and Turlach B. On the lasso and its dual. *J Comp Graph Stat*, 9:319, 2000.
- [186] Roth V. The generalized LASSO: a wrapper approach to gene selection for microarray data. Technical report, Dept. of Computer Science, University of Bonn, 2002.
- [187] Liu J, Iba H, and Ishizuka M. Selecting informative genes with parallel genetic algorithms in tissue classification. *Genome Inform*, 12:14, 2001.
- [188] Ooi C and Tan P. Genetic algorithms applied to multi-class prediction for the analysis of gene expression data. *Bioinformatics*, 19(1):37, 2003.
- [189] Guyon I, Weston J, Barnhill S, and Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46(1):389, 2002.

- [190] Breiman L. Manual – Setting Up, Using, and Understanding Random Forests V4.0, 2003. <ftp://ftp.stat.berkeley.edu/pub/users/breiman>.
- [191] Breiman L. Random forests. *Machine Learning*, 45(1):5, 2001.
- [192] Gini C. Measurement of Inequality of Incomes. *The Economic Journal*, 31(121):124, 1921.
- [193] Diaz-Uriarte R and Alvarez de Andres S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(1):1471, 2006.
- [194] Jiang H, Deng Y, Chen H, Tao L, et al. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC Bioinformatics*, 5(1):81, 2004.
- [195] Nicodemus K, Malley J, Strobl C, and Ziegler A. The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, 11(1):110, 2010.
- [196] Hoerl A and Kennard R. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80, 2000.
- [197] Bielza C, Robles V, and Larrañaga P. Regularized logistic regression without a penalty term: An application to cancer classification with microarray data. *Expert Syst Appl*, 38(5):5110, 2011.
- [198] Xing EP, Jordan MI, and Karp RM. Feature selection for high-dimensional genomic microarray data. In *Proc. Int. Conf. Mach. Learn.*, pages 601–608. Morgan Kaufmann, San Francisco, CA, 2001.
- [199] El Akadi A, Amine A, El Ouardighi A, and Aboutajdine D. A two-stage gene selection scheme utilizing MRMR filter and GA wrapper. *Knowl Inform Syst*, pages 1–14, 2010.
- [200] Leung Y and Hung Y. A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification. *IEEE ACM Trans Comput Biol Bioinform*, 7(1):108, 2010.
- [201] Raczynski L, Wozniak K, Rubel T, and Zaremba K. Application of Density Based Clustering to Microarray Data Analysis. *Int J Electron Telecommun*, 56(3):281, 2010.
- [202] Karypis G, Han E, and Kumar V. Chameleon: Hierarchical clustering using dynamic modeling. *IEEE Computer*, 32(8):68, 2002.
- [203] Hartigan J and Wong M. A K-means clustering algorithm. *JR Stat Soc Ser C*, 28:100, 1979.
- [204] MacQueen J. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:281, 1967.
- [205] Forgy E. Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications. *Biometrics*, 21:768, 1965.
- [206] Lloyd S. Least squares quantization in PCM. Special issue on quantization. *IEEE Trans Inform Theory*, 28:129, 1982.
- [207] Kaufman L and Rousseeuw P. *Finding groups in data: an introduction to cluster analysis*. Wiley-Interscience, New York, NY, USA, 1990.
- [208] Ng R and Han J. CLARANS: A Method for Clustering Objects for Spatial Data Mining. *IEEE T Knowl Data En*, 14:1003, 2002.
- [209] Kohonen T. *Self-Organizing Maps*. Springer-Verlag Berlin, Heidelberg, Berlin, 2001.
- [210] Bezdek J. *Pattern recognition with fuzzy objective function algorithms*. Kluwer Academic Publishers Norwell, MA, USA, 1981.
- [211] Stutz C. *Anwendungsspezifische Fuzzy-Clustermethoden*. Ph.D. thesis, Dissertation Fakultät für Informatik, Technische Universität München, Germany, 1999.
- [212] Ward J. Hierarchical grouping to optimize an objective function. *J Am Stat Assoc*, 58:236, 1963.
- [213] McQuitty L. Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data. *Educ Psychol Meas*, 26:825, 1966.
- [214] Macnaughton-Smith P. *Some Statistical and Other Numerical Techniques for Classifying Individuals*. Her Majesty's Stationery Office: London, 1965.
- [215] Ben-Dor A and Yakhini Z. Clustering gene expression patterns. In *Proceedings of the Third Annual International Conference on Computational molecular biology*, RECOMB'99, pages 33–42. ACM, New York, NY, USA, 1999.
- [216] Heller K and Ghahramani Z. Bayesian hierarchical clustering. In *Proc. Int. Conf. Mach. Learn.*, pages 297–304. ACM, 2005.
- [217] Savage R, Heller K, Xu Y, Ghahramani Z, et al. R/BHC: fast Bayesian hierarchical clustering for microarray data. *BMC Bioinformatics*, 10(1):242, 2009.
- [218] Kim E, Kim S, Ashlock D, and Nam D. MULTI-K: accurate classification of microarray subtypes using ensemble k-means clustering. *BMC Bioinformatics*, 10(1):260, 2009.
- [219] Fred A and Jain A. Combining multiple clusterings using evidence accumulation. *IEEE Trans Pattern Anal Mach Intell*, pages 835–850, 2005.
- [220] Fern X and Brodley C. Solving cluster ensemble problems by bipartite graph partitioning. In *Proceedings of the Twenty-First International Conference on Machine learning*, page 36. ACM, 2004.
- [221] Ghosh J, Strehl A, and Merugu S. A consensus framework for integrating distributed clusterings under limited knowledge sharing. In *Proceedings of the NSF Workshop on Next Generation Data Mining*, pages 99–108. 2002.
- [222] Iam-On N, Boongoen T, and Garrett S. LCE: a link-based cluster ensemble method for improved gene expression data analysis. *Bioinformatics*, 26(12):1513, 2010.
- [223] Iam-On N and Garrett S. Linkclue: A matlab package for link-based cluster ensembles. *J Stat Software*, 36(9):1, 2010.

- [224] Reuther P, Walter B, Ley M, Weber A, et al. Managing the quality of person names in DBLP. *Res Adv Tech Digit Libr*, pages 508–511, 2006.
- [225] Marinakis Y, Marinaki M, and Matsatsinis N. A hybrid clustering algorithm based on honey bees mating optimization and greedy randomized adaptive search procedure. *Learning and Intelligent Optimization*, pages 138–152, 2008.
- [226] Zhang A. *Advanced Analysis of Gene Expression Microarray Data*. World Scientific Publishing Co. Pte. Ltd., Singapore, 2006.
- [227] Rousseeuw P. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*, 20:53, 1987.
- [228] Dunn J. Well-separated clusters and optimal fuzzy partitions. *Cybernetics and Systems*, 4(1):95, 1974.
- [229] Hubert L and Schultz J. Quadratic assignment as a general data analysis strategy. *Br J Math Stat Psychol*, 29:190, 1976.
- [230] Milligan G and Cooper M. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159, 1985.
- [231] Brock G, Pihur V, Datta S, and Datta S. cValid, an R package for cluster validation. *J Stat Software*, 25(4):1, 2008.
- [232] Handl J, Knowles J, and Kell D. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201, 2005.
- [233] Hubert L and Arabie P. Comparing partitions. *Journal of Classification*, 2:193, 1985.
- [234] Tavazoie S, Hughes J, Campbell M, Cho R, et al. Systematic determination of genetic network architecture. *Nat Genet*, 22:281, 1999.
- [235] Boutin F and Hascoet M. Cluster validity indices for graph partitioning. *Proceedings of the Eighth International Conference on Information Visualisation*, pages 376–381, 2004.
- [236] Alon U, Barkai N, Notterman D, Gish K, et al. Broad patterns of gene expression revealed by clustering analysis of tumour and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci USA*, 96(12):6745, 1999.
- [237] Fisher R. The use of multiple measurements in taxonomic problems. *Ann Eugen*, 7:179, 1936.
- [238] Ancona N, Maglietta R, D'Addabbo A, Liuni S, et al. Regularized least squares cancer classifiers from DNA microarray data. *BMC Bioinformatics*, 6(Suppl 4):S2, 2005.
- [239] Tibshirani R, Hastie T, Narasimhan B, and Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci USA*, 99(10):6567, 2002.
- [240] Guo Y, Hastie T, and Tibshirani R. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, 8:86, 2007.
- [241] Tai F and Pan W. Incorporating prior knowledge of gene functional groups into regularized discriminant analysis of microarray data. *Bioinformatics*, 2007.
- [242] Theilhaber J, Connolly T, Roman-Roman S, Bushnell S, et al. Finding genes in the C2C12 osteogenic pathway by k-nearest-neighbor classification of expression data. *Genome Res*, 12(1):165, 2002.
- [243] Day W and Edelsbrunner H. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification*, 1(1):7, 1984.
- [244] Cleary J and Trigg L. K\*: An Instance-based Learner Using an Entropic Distance Measure. In *Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, California*, pages 108–114. Morgan Kaufmann, 1995.
- [245] Hastie T and Tibshirani R. Discriminant adaptive nearest neighbor classification. *IEEE Trans Pattern Anal Mach Intell*, 18(6):607, 2002.
- [246] Domeniconi C, Peng J, and Gunopulos D. Locally adaptive metric nearest-neighbor classification. *IEEE Trans Pattern Anal Mach Intell*, 24(9):1281, 2002.
- [247] Cohen I, Cozman F, Sebe N, Cirelo M, et al. Semisupervised learning of classifiers: Theory, algorithms, and their application to human-computer interaction. *IEEE Trans Pattern Anal Mach Intell*, 26(12):1553, 2004.
- [248] Driessens K, Reutemann P, Pfahringer B, and Leschi C. Using weighted nearest neighbor to benefit from unlabelled data. *Advances in Knowledge Discovery and Data Mining*, pages 60–69, 2006.
- [249] Hassan M, Hossain M, Bailey J, and Ramamohanarao K. Improving k-nearest neighbour classification with distance functions based on receiver operating characteristics. *Machine Learning and Knowledge Discovery in Databases*, pages 489–504, 2008.
- [250] Fujibuchi W and Kato T. Classification of heterogeneous microarray data by maximum entropy kernel. *BMC Bioinformatics*, 8(1):267, 2007.
- [251] Zou H and Hastie T. Regularization and variable selection via the elastic net. *J Roy Statist Soc Ser B Methodological*, 67(2):301, 2005.
- [252] Rosset S and Zhu J. Piecewise linear regularized solution paths. *Ann Stat*, 35(3):1012, 2007.
- [253] Liu Y. Active learning with support vector machine applied to gene expression data for cancer classification. *J Chem Inf Comput Sci*, 44(6):1936, 2004.
- [254] Sassano M. An empirical study of active learning with support vector machines for Japanese word segmentation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 505–512, 2001.
- [255] Tong S, Koller D, and Kaelbling L. Support Vector Machine Active Learning with Applications to Text Classification. *J Mach Learn Res*, 2:45, 2001.
- [256] Chu W, Ghahramani Z, Falciani F, and Wild D. Biomarker discovery in microarray gene expression data with Gaussian processes. *Bioinformatics*, 21(16):3385, 2005.

- [257] Cho J, Lee D, Park J, and Lee I. Gene selection and classification from microarray data using kernel machine. *FEBS Lett*, 571(1-3):93, 2004.
- [258] Rosenblatt F. *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*. Spartan Books, Washington, D.C., USA, 1962.
- [259] Schwarzer G, Vach W, and Schumacher M. On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology. *Stat Med*, 19(4):541, 2000.
- [260] Fariselli P, Olmea O, Valencia A, and Casadio R. Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations. *Proteins*, 45(S5):157, 2001.
- [261] Fariselli P and Casadio R. Prediction of disulfide connectivity in proteins. *Bioinformatics*, 17(10):957, 2001.
- [262] Fariselli P and Casadio R. HTP: a neural network-based method for predicting the topology of helical transmembrane domains in proteins. *Computer applications in the biosciences CABIOS*, 12(1):41, 1996.
- [263] Khan J, Wei J, Ringnér M, Saal L, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med*, 7(6):673, 2001.
- [264] Lancashire L, Rees R, and Ball G. Identification of gene transcript signatures predictive for estrogen receptor and lymph node status using a stepwise forward selection artificial neural network modelling approach. *Artif Intell Med*, 43(2):99, 2008.
- [265] Lancashire L, Powe D, Reis-Filho J, Rakha E, et al. A validated gene expression profile for detecting clinical outcome in breast cancer using artificial neural networks. *Breast Cancer Res Treat*, 120(1):83, 2010.
- [266] Ritchie M, White B, Parker J, Hahn L, et al. Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. *BMC Bioinformatics*, 4(1):28, 2003.
- [267] Ritchie M, Motsinger A, Bush W, Coffey C, et al. Genetic programming neural networks: a powerful bioinformatics tool for human genetics. *Applied Soft Computing*, 7(1):471, 2007.
- [268] Towell G, Shavlik J, and Noordewier M. Refinement of approximate domain theories by knowledge-based neural networks. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, pages 861–866. 1990.
- [269] Yang R, Wang Z, Heng P, and Leung K. Classification of heterogeneous fuzzy data by Choquet integral with fuzzy-valued integrand. *IEEE T Fuzzy Syst*, 15(5):931, 2007.
- [270] Duch W, Setiono R, and Zurada JM. Computational intelligence methods for rule-based data understanding. In *Proceedings of the IEEE*, pages 771–805. 2004.
- [271] Duch W, Adamczak R, and Grabczewski K. A new methodology of extraction, optimization and application of crisp and fuzzy logical rules. *IEEE Transactions on Neural Networks*, 12(2):277, 2001.
- [272] Vélez M, Sánchez O, Romero S, and Andújar J. A new methodology to improve interpretability in neuro-fuzzy TSK models. *Applied Soft Computing*, 10(2):578, 2010.
- [273] Eiamkanitchat N, Theera-Umpon N, and Auephanwiriyakul S. A novel neuro-fuzzy method for linguistic feature selection and rule-based classification. In *The 2nd International Conference on Computer and Automation Engineering*, volume 2, pages 247–252. 2010.
- [274] Webb G, Boughton J, and Wang Z. Not so naive bayes: Aggregating one-dependence estimators. *Machine Learning*, 58(1):5, 2005.
- [275] Bressan M and Vitria J. On the selection and classification of independent features. *IEEE Trans Pattern Anal Mach Intell*, 25(10):1312, 2003.
- [276] Fan L, Poh K, and Zhou P. Partition-conditional ICA for Bayesian classification of microarray data. *Expert Syst Appl*, 37:8188, 2010.
- [277] Miguel H et al. Network-based sparse Bayesian classification. *Pattern Recognit*, 44(4):886, 2011.
- [278] Minka T. *A family of algorithms for approximate Bayesian inference*. Ph.D. thesis, Massachusetts Institute of Technology, MA, USA, 2001.
- [279] Zhu Y, Shen X, and Pan W. Network-based support vector machine for classification of microarray samples. *BMC Bioinformatics*, 10(Suppl 1):S21, 2009.
- [280] Jacob L, Obozinski G, and Vert J. Group Lasso with overlap and graph Lasso. In *Proc. Int. Conf. Mach. Learn.*, pages 433–440. ACM, 2009.
- [281] Quinlan J. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Francisco, CA, USA, 1993.
- [282] Breiman L. *Classification and Regression Trees*. Chapman & Hall/CRC, Monterey, CA, USA, 1998.
- [283] Goodarzi M and Kangavari M. Using Fuzzy Decision Trees for Authentication via Keystroke Timings. In *Proceedings of the 4th International Conference on Intelligent Systems Design and Application*, pages 1–5. Budapest, Hungary, 2004.
- [284] Geman D et al. Classifying gene expression profiles from pairwise mRNA comparisons. *Stat Appl Genet Mol Biol*, 3:19, 2004.
- [285] Tan A, Naiman D, Xu L, Winslow R, et al. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, 21(20):3896, 2005.
- [286] Czajkowski M and Kretowski M. Novel Extension of k-TSP Algorithm for Microarray Classification. *New Frontiers in Applied Artificial Intelligence*, pages 456–465, 2008.
- [287] Yoon S and Kim S. k-Top Scoring Pair Algorithm for feature selection in SVM with applications to microarray data classification. *Soft Computing A Fusion of Foundations Methodologies and Applications*, 14(2):151, 2010.



- [288] Li J and Tang X. A new classification model with simple decision rule for discovering optimal feature gene pairs. *Comput Biol Med*, 37(11):1637, 2007.
- [289] Tsymbal A, Pechenizkiy M, and Cunningham P. Diversity in search strategies for ensemble feature selection. *Inform Fusion*, 6(1):83, 2005.
- [290] Breiman L. Bagging predictors. *Machine learning*, 24(2):123, 1996.
- [291] Efron B and Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, Monterey, CA, USA, 1 edition, 1994.
- [292] Valentini G, Muselli M, and Ruffino F. Cancer recognition with bagged ensembles of support vector machines. *Neurocomputing*, 56:461, 2004.
- [293] Amaratunga D, Cabrera J, and Lee Y. Enriched random forests. *Bioinformatics*, 24(18):2010, 2008.
- [294] Zhang H, Yu C, and Singer B. Cell and tumor classification using gene expression data: construction of forests. *Proc Natl Acad Sci USA*, 100(7):4168, 2003.
- [295] Chen X, Wang M, and Zhang H. The use of classification trees for bioinformatics. *Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery*, 1(1):55, 2011.
- [296] Zhang H and Wang M. Search for the smallest random forest. *Stat Interface*, 2(3):381, 2009.
- [297] Freund Y and Schapire R. Experiments with a new boosting algorithm. In *Proc. Int. Conf. Mach. Learn.*, pages 148–156. ACM, 1996.
- [298] Friedman J, Hastie T, and Tibshirani R. Additive logistic regression: a statistical view of boosting. *Ann Stat*, 28(2):337, 2000.
- [299] Demiriz A, Bennett K, and Shawe-Taylor J. Linear programming boosting via column generation. *Machine Learning*, 46(1):225, 2002.
- [300] Dettling M and Buhlmann P. Boosting for tumour classification with gene expression data. *Bioinformatics*, 19(9):1061, 2003.
- [301] Wolpert D. Stacked generalization. *Neural networks*, 5(2):241, 1992.
- [302] Bell R and Koren Y. Lessons from the Netflix prize challenge. *ACM SIGKDD Explorations Newsletter*, 9(2):75, 2007.
- [303] Warnat P et al. Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics*, 6(1):265, 2005.
- [304] Johnson W, Li C, and Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, 8:118, 2007.
- [305] Benito M, Parker J, Du Q, Wu J, et al. Adjustment of systematic microarray data biases. *Bioinformatics*, 20(1):105, 2004.
- [306] Marron J, Todd M, and Ahn J. Distance-weighted discrimination. *J Am Stat Assoc*, 102(480):1267, 2007.
- [307] Huang D, Sherman B, and Lempicki R. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucl Acids Res*, 37(1):1, 2009.
- [308] Subramanian A, Tamayo P, Mootha V, Mukherjee S, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*, 102(43):15545, 2005.
- [309] Backes C, Keller A, Kuentzer J, Kneissl B, et al. GeneTrail—advanced gene set enrichment analysis. *Nucleic Acids Res*, 35(suppl 2):W186, 2007.
- [310] Kim S and Volsky D. PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*, 6(1):144, 2005.
- [311] Luo W, Friedman M, Shedden K, Hankenson K, et al. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*, 10(1):161, 2009.
- [312] Al-Shahrour F, Arbiza L, Dopazo H, Huerta-Cepas J, et al. From genes to functional classes in the study of biological systems. *BMC Bioinformatics*, 8(1):114, 2007.
- [313] Lee H, Braynen W, Keshav K, and Pavlidis P. ErmineJ: tool for functional analysis of gene expression data sets. *BMC Bioinformatics*, 6(1):269, 2005.
- [314] Tu K, Yu H, and Zhu M. MEGO: gene functional module expression based on gene ontology. *Biotechniques*, 38:277, 2005.
- [315] Nam D, Kim S, Kim S, Yang S, et al. ADGO: analysis of differentially expressed gene sets using composite GO annotation. *Bioinformatics*, 22(18):2249, 2006.
- [316] Bauer S, Grossmann S, Vingron M, and Robinson P. Ontologizer 2.0—a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics*, 24(14):1650, 2008.
- [317] Carmona-Saez P, Chagoyen M, Tirado F, Carazo J, et al. GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol*, 8(1):R3, 2007.
- [318] Guo Z et al. Towards precise classification of cancers based on robust gene functional expression profiles. *BMC Bioinformatics*, 6(1):58, 2005.
- [319] Gagneur J, Krause R, Bouwmeester T, and Casari G. Modular decomposition of protein-protein interaction networks. *Genome Biol*, 5(8):R57, 2004.
- [320] Radicchi F, Castellano C, Cecconi F, Loreto V, et al. Defining and identifying communities in networks. *Proc Natl Acad Sci USA*, 101(9):2658, 2004.
- [321] Theocharidis A, Van Dongen S, Enright A, and Freeman T. Network visualization and analysis of gene expression data using BioLayout Express3D. *Nat Protoc*, 4(10):1535, 2009.
- [322] Cerami E, Demir E, Schultz N, Taylor B, et al. Automated network analysis identifies core pathways in glioblastoma. *PLoS ONE*, 5(2):e8918, 2010.

- [323] Ulitsky I, Maron-Katz A, Shavit S, Sagir D, et al. Expander: from expression microarrays to networks and functions. *Nat Protoc*, 5(2):303, 2010.
- [324] Jiang P and Singh M. SPICi: a fast clustering algorithm for large biological networks. *Bioinformatics*, 26(8):1105, 2010.
- [325] Vázquez C, Freyre-González J, Gosset G, Loza J, et al. Identification of network topological units coordinating the global expression response to glucose in *Bacillus subtilis* and its comparison to *Escherichia coli*. *BMC Microbiol*, 9(1):176, 2009.
- [326] Vellaichamy A, Dezső Z, JeBailey L, Chinnaiyan A, et al. "Topological Significance" Analysis of Gene Expression and Proteomic Profiles from Prostate Cancer Cells Reveals Key Mechanisms of Androgen Response. *PLoS ONE*, 5(6):e5364, 2010.
- [327] Ray M and Zhang W. Analysis of Alzheimer's disease severity across brain regions by topological analysis of gene co-expression networks. *BMC Syst Biol*, 4(1):136, 2010.
- [328] Chuang H, Lee E, Liu Y, Lee D, et al. Network-based classification of breast cancer metastasis. *Mol Syst Biol*, 3(1), 2007.
- [329] Rogers G, Moscato P, and Langston M. Graph algorithms for machine learning: a case-control study based on prostate cancer populations and high throughput transcriptomic data. *BMC Bioinformatics*, 11(Suppl 4):P21, 2010.
- [330] Newman M. Modularity and community structure in networks. *Proc Natl Acad Sci USA*, 103(23):8577, 2006.
- [331] Clauset A, Newman M, and Moore C. Finding community structure in very large networks. *Physical Review E*, 70(6):66111, 2004.
- [332] Pons P and Latapy M. Computing communities in large networks using random walks. *J Graph Algorithms Appl*, 10(2):191, 2006.
- [333] Reichardt J and Bornholdt S. Statistical mechanics of community detection. *Physical Review E*, 74(1):16110, 2006.
- [334] Pizzuti C. Community detection in social networks with genetic algorithms. In *Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation*, pages 1137–1138. ACM, 2008.
- [335] Li Z, Zhang S, Wang R, Zhang X, et al. Quantitative function for community detection. *Physical Review E*, 77(3):36109, 2008.
- [336] Bader G and Hogue C. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4(1):2, 2003.
- [337] Van Dongen S. *Graph clustering by flow simulation*. Ph.D. thesis, University of Utrecht, Netherlands, 2000.
- [338] Moschopoulos C, Pavlopoulos G, Schneider R, Likothanassis S, et al. GIBA: a clustering tool for detecting protein complexes. *BMC Bioinformatics*, 10(Suppl 6):S11, 2009.
- [339] Rivera C, Vakil R, and Bader J. NeMo: Network Module identification in Cytoscape. *BMC Bioinformatics*, 11(Suppl 1):S61, 2010.
- [340] Sun PG, Gao L, and Han SS. Identification of overlapping and non-overlapping community structure by fuzzy clustering in complex networks. *Inf Sci*, 18(6):1060, 2010.
- [341] Ahn Y, Bagrow J, and Lehmann S. Link communities reveal multiscale complexity in networks. *Nature*, 466(7307):761, 2010.
- [342] Zarei M and Samani K. Eigenvectors of network complement reveal community structure more accurately. *Phys Stat Mech Appl*, 388(8):1721, 2009.
- [343] Zhang B and Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*, 4(1):1128, 2005.
- [344] Newman M. Assortative mixing in networks. *Phys Rev Lett*, 89(20):208701, 2002.
- [345] Junker B and Schreiber F. *Analysis of biological networks*. John Wiley & Sons, Hoboken, New Jersey, USA, 2008.
- [346] Vapnik V and Chervonenkis A. *Theory of Pattern Recognition*. Nauka, Moscow, 1974.
- [347] Shipp M, Ross K, Tamayo P, Weng A, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med*, 8(1):68, 2002.
- [348] Singh D, Febbo P, Ross K, Jackson D, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203, 2002.
- [349] Chin S, Teschendorff A, Marioni J, Wang Y, et al. High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. *Genome Biol*, 8(10):R215, 2007.
- [350] Naderi A, Teschendorff A, Barbosa-Morais N, Pinder S, et al. A gene-expression signature to predict survival in breast cancer across independent data sets. *Oncogene*, 26(10):1507, 2006.
- [351] Zhang H, Rakha E, Ball G, Spiteri I, et al. The proteins FABP7 and OATP2 are associated with the basal phenotype and patient outcome in human breast cancer. *Breast Cancer Res Treat*, 121(1):41, 2010.
- [352] Affymetrix. *Affymetrix Microarray Suite User Guide, Version 5*, 2001.
- [353] Ihaka R and Gentleman R. R: A language for data analysis and graphics. *J Comput Graph Stat*, 5(3):299, 1996.
- [354] Wu Z and Irizarry R. Stochastic Models Inspired by Hybridization Theory for Short Oligonucleotide Arrays. *J Comput Biol*, 12(6):882, 2005.
- [355] Hastie T and Tibshirani R. *Generalized Additive Models*. Chapman & Hall/CRC, Monterey, CA, USA, 1990.
- [356] Boulesteix A and Strimmer K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief Bioinform*, 8(1):32, 2007.
- [357] Zhang C, Lu X, and Zhang X. Significance of gene ranking for classification of microarray samples. *IEEE ACM Trans Comput Biol Bioinform*, 3(3):312, July 2006.

- [358] Li W and Yang Y. How many genes are needed for a discriminant microarray data analysis? In *Methods of Microarray Data Analysis*, pages 137–150. Kluwer Academic, Boston, USA, 2001.
- [359] Wolfinger R, Gibson G, Wolfinger E, Bennett L, et al. Assessing Gene Significance from cDNA Microarray Expression Data via Mixed Models. *J Comput Biol*, 8(6):625, 2001.
- [360] Wold H. Soft modeling by latent variables: the nonlinear iterative partial least squares approach. In *Perspectives in Probability and Statistics*, pages 117–142. 1975.
- [361] Boulesteix A. PLS Dimension Reduction for Classification with Microarray Data. *Stat Appl Genet Mol Biol*, 3(1):33, 2004.
- [362] de Jong S. SIMPLS: An alternative approach to partial least squares regression. *Chemom Intell Lab Syst*, 18:251, 1993.
- [363] Wessels L, Reinders M, Hart A, Veenman C, et al. A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics*, 21(19):3755, 2005.
- [364] Guyon I, Gunn S, Ben-Hur A, and Dror G. Result analysis of the NIPS 2003 feature selection challenge. *Adv Neural Inf Process Syst*, 17:545, 2005.
- [365] Chen Y and Lin C. Combining SVMs with Various Feature Selection Strategies. *Studies in fuzziness and soft computing*, 207:315, 2006.
- [366] Stout M, Bacardit J, Hirst J, and Krasnogor N. Prediction of Recursive Convex Hull Class Assignments for Protein Residues. *Bioinformatics*, 24(7):916, 2008.
- [367] Bacardit J and Krasnogor N. Fast Rule Representation for Continuous Attributes in Genetics-Based Machine Learning. In *Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation*, pages 2039–2046. ACM Press, New York, NY, USA, 2008.
- [368] Bacardit J, Stout M, Hirst JD, Valencia A, et al. Automated alphabet reduction for protein datasets. *BMC Bioinformatics*, 10(1):6, 2009.
- [369] Bacardit J, Burke E, and Krasnogor N. Improving the scalability of rule-based evolutionary learning. *Memetic Computing*, 1(1):55, 2009.
- [370] Giraldez R, Aguilar-Ruiz J, and Riquelme J. Natural Coding: A more efficient representation for evolutionary learning. *Lect Notes Comput Sci GECCO*, 2723:979, 2003.
- [371] Venturini G. Sia: A supervised inductive algorithm with genetic search for learning attributes based concepts. In *Proceedings of the European Conference on Machine Learning*, pages 280–296. Springer-Verlag, London, UK, 1993.
- [372] Rivest R. Learning decision lists. *Machine Learning*, 2(3):229, 1987.
- [373] Bacardit J. Pittsburgh Genetics-Based Machine Learning in the Data Mining era: Representations, generalization, and run-time. *Doctoral dissertation Ramon Llull University Barcelona Catalonia Spain*, 2004.
- [374] Rissanen J. A Universal Prior for Integers and Estimation by Minimum Description Length. *Ann Stat*, 11(2):416, 1983.
- [375] De Jong KA and Spears WM. Learning concept classification rules using genetic algorithms. In *In Proceedings of the International Joint Conference on Artificial Intelligence*, pages 651–656. Morgan Kaufmann, 1991.
- [376] Dimitriadou E, Hornik K, Leisch F, Meyer D, et al. Misc functions of the department of statistics (e1071), TU Wien, 2005. R-Package e1071 version 1.5-19.
- [377] Chang CC and Lin CJ. *LIBSVM: a library for support vector machines*, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [378] Furey T, Cristianini N, Duffy N, Bednarski D, et al. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10):906, 2000.
- [379] Hsu C, Chang C, and Lin C. A practical guide to support vector classification, 2008. <http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>.
- [380] Ambrose C and McLachlan G. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci USA*, 99(10):6562, 2002.
- [381] Bacardit J and Krasnogor N. BioHEL: Bioinformatics-oriented Hierarchical Evolutionary Learning, 2006. Nottingham eprints, University of Nottingham.
- [382] Liu J and Zhou H. Tumor classification based on gene microarray data and hybrid learning method. In *Proc. Int. Conf. on Machine Learning and Cybernetics*, pages 2275–2280. 2003.
- [383] Braga-Neto U and Dougherty E. Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 20(3):374, 2004.
- [384] Goh L, Kasabov N, and Song Q. A novel feature selection method to improve classification of gene expression data. In *Second Asia-Pacific Bioinformatics Conference*, pages 161–166. ACS, 2004.
- [385] Lecoche M and Hess K. An empirical study of univariate and genetic algorithm-based feature selection in binary classification with microarray data. *Cancer Inform*, 2:313, 2007.
- [386] Hu Y and Kasabov N. Ontology-based framework for personalized diagnosis and prognosis of cancer based on gene expression data. In M Ishikawa, K Doya, H Miyamoto, and T Yamakawa, editors, *Neural Information Processing*, pages 846–855. Springer-Verlag Berlin, Heidelberg, 2008.
- [387] Shen L and Tan E. Dimension Reduction-Based Penalized Logistic Regression for Cancer Classification Using Microarray Data. *IEEE ACM Trans Comput Biol Bioinformatics*, 2(2):166, 2005.

- [388] Paul T and Iba H. Extraction of informative genes from microarray data. In *Proceedings of the 7th Annual Conference on Genetic and Evolutionary Computation*, pages 453–460. ACM Press, 2005.
- [389] Chu W, Ghahramani Z, Falciani F, and Wild D. Biomarker discovery in microarray gene expression data with Gaussian processes. *Bioinformatics*, 21(16):3385, 2005.
- [390] Conover WJ. *Practical Nonparametric Statistics*. John Wiley, New York, NY, USA, 1971.
- [391] Demšar J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J Mach Learn Res*, 7:1, 2006.
- [392] Alexe G et al. Breast cancer prognosis by combinatorial analysis of gene expression data. *Breast Cancer Res*, 8(4):R41, 2006.
- [393] Dennis Jr G, Sherman B, Hosack D, Yang J, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol*, 4(9):R60, 2003.
- [394] Hu P, Greenwood C, and Beyene J. Integrating Affymetrix microarray data sets using probe-level test statistic for predicting prostate cancer. In *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pages 63–70. 2006.
- [395] Gerhold D, Jensen R, and Gullans S. Better therapeutics through microarrays. *Nat Genet*, 32(Suppl.):547, 2002.
- [396] Belacel N, Cuperlovic-Culf M, and Ouellette R. Molecular method for diagnosis of prostate cancer, September 13 2006. US Patent App. 11/519,892.
- [397] Bolstad B, Irizarry R, Gautier L, and Wu Z. Pre-processing high-density oligonucleotide arrays. In R Gentleman, V Carey, W Huber, R Irizarry, and S Dudoit, editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pages 15–16. Springer-Verlag Berlin, Heidelberg, 2005.
- [398] Bartenhagen C, Klein H, Ruckert C, Jiang X, et al. Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data. *BMC Bioinformatics*, 11(1):567, 2010.
- [399] Matsumoto M and Nishimura T. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, 8(1):3, 1998.
- [400] Zervakis M, Blazadonakis M, Tsiliki G, Danilidou V, et al. Outcome prediction based on microarray analysis: a critical perspective on methods. *BMC Bioinformatics*, 10(1):53, 2009.
- [401] Irizarry R, Bolstad B, Collin F, Cope L, et al. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res*, 31(4):e15, 2003.
- [402] The MathWorks Inc., Natick, MA, USA. *Matlab*, 1998.
- [403] Tarraga J, Medina I, Carbonell J, Huerta-Cepas J, et al. GEPAS, a web-based tool for microarray data analysis and interpretation. *Nucleic Acids Res*, 31(13):3461, 2008.
- [404] Kapushesky M, Kemmeren P, Culhane A, Durinck S, et al. Expression Profiler: next generation—an online platform for analysis of microarray data. *Nucleic Acids Res*, 32(Web Server Issue):W465, 2004.
- [405] Díaz-Uriarte R, Alibés A, Morrissey E, Cañada A, et al. Asterias: integrated analysis of expression and aCGH data using an open-source, web-based, parallelized software suite. *Nucleic Acids Res*, 35(Web Server issue):W75, 2007.
- [406] Zhu Y, Zhu Y, and Xu W. EzArray: A web-based highly automated Affymetrix expression array data management and analysis system. *BMC Bioinformatics*, 9(1):46, 2008.
- [407] Rainer J, Sanchez-Cabo F, Stocker G, Sturn A, et al. CARMAweb: comprehensive R- and bioconductor-based web service for microarray data analysis. *Nucleic Acids Res*, 34(Web Server issue):W498, 2006.
- [408] Rehrauer H, Zoller S, and Schlapbach R. MAGMA: analysis of two-channel microarrays made easy. *Nucleic Acids Res*, 35(Web Server issue):W86, 2007.
- [409] Hokamp K, Roche F, Acab M, Rousseau M, et al. ArrayPipe: a flexible processing pipeline for microarray data. *Nucleic Acids Res*, 32(Web Server Issue):W457, 2004.
- [410] Psarros M, Heber S, Sick M, Thoppae G, et al. RACE: remote analysis computation for gene expression data. *Nucleic Acids Res*, 33(Web Server Issue):W638, 2005.
- [411] Xia X, McClelland M, and Wang Y. WebArray: an online platform for microarray data analysis. *BMC Bioinformatics*, 6(1):306, 2005.
- [412] Romualdi C, Vitulo N, Favero M, and Lanfranchi G. MIDAW: a web tool for statistical analysis of microarray data. *Nucleic Acids Res*, 33(Web Server Issue):W644, 2005.
- [413] Wu C, Fu Y, Murali T, and Kasif S. Gene expression module discovery using Gibbs sampling. *Genome Inform*, 15(1):239, 2004.
- [414] Lee J, Sinkovits R, Mock D, Rab E, et al. Components of the antigen processing and presentation pathway revealed by gene expression microarray analysis following B cell antigen receptor (BCR) stimulation. *BMC Bioinformatics*, 7(1):237, 2006.
- [415] Aburatani S, Goto K, Saito S, Toh H, et al. ASIAN: a web server for inferring a regulatory network framework from gene expression profiles. *Nucleic Acids Res*, 33(Web Server Issue):W659, 2005.
- [416] Lu Y, He X, and Zhong S. Cross-species microarray analysis with the OSCAR system suggests an INSR -> Pax6 -> NQO1 neuro-protective pathway in aging and Alzheimer's disease. *Nucleic Acids Res*, 35(Web Server issue):W105, 2007.
- [417] Shabalina A, Tjelmeland H, Fan C, Perou C, et al. Merging two gene-expression studies via cross-platform normalization. *Bioinformatics*, 24(9):1154, 2008.
- [418] Gentleman R, Carey V, Bates D, Bolstad B, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10):R80, 2004.

- [419] Tusher V, Tibshirani R, and Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA*, 98(9):5116, 2001.
- [420] Armstrong S, Staunton J, Silverman L, Pieters R, et al. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet*, 30(1):41, 2001.
- [421] Chipman H and Tibshirani R. Hybrid hierarchical clustering with applications to microarray data. *Biostatistics*, 7(2):286, 2006.
- [422] Herrero J, Valencia A, and Dopazo J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, 17(2):126, 2001.
- [423] Kirkpatrick S, Jr CG, and Vecchi M. Optimization by Simulated Annealing. *Science*, 220:671, 1983.
- [424] Szu H. Fast simulated annealing. In *AIP Conference Proceedings*, volume 151, page 420. 1986.
- [425] de Vicente J, Lanchares J, and Hermida R. Placement by thermodynamic simulated annealing. *Phys Lett A*, 317(5-6):415, 2003.
- [426] Ingber L. Simulated annealing: Practice versus theory. *Math Comput Model*, 18(11):29, 1993.
- [427] Bacardit J, Stout M, Hirst J, and Krasnogor N. Data mining in proteomics with learning classifier systems. In L Bull, E Bernado Mansilla, and J Holmes, editors, *Learning Classifier Systems in Data Mining*, pages 17–46. Springer-Verlag Berlin, Heidelberg, 2008.
- [428] Cohen J et al. A coefficient of agreement for nominal scales. *Educ Psychol Meas*, 20(1):37, 1960.
- [429] Huberty CJ. *Applied Discriminant Analysis*. John Wiley, New York, NY, USA, 1994.
- [430] Zhu Z, Ong Y, and Dash M. Markov blanket-embedded genetic algorithm for gene selection. *Pattern Recognit*, 40(11):3236, 2007.
- [431] Fruchterman T and Reingold E. Graph drawing by force-directed placement. *Software Practice and Experience*, 21(11):1129, 1991.
- [432] Schmuhl M. Graphopt, 2010. <http://www.schmuhl.org/graphopt>.
- [433] Martin S. Drl, 2010. <http://www.cs.sandia.gov/~smartin/software.htm>.
- [434] Kamada T and Kawai S. An algorithm for drawing general undirected graphs. *Information processing letters*, 31(12):7, 1989.
- [435] Shannon P, Markiel A, Ozier O, Baliga N, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498, 2003.
- [436] Martinez R, Pasquier C, and Pasquier N. GenMiner: Mining Informative Association Rules from Genomic Data. In *Proceedings of the 2007 IEEE International Conference on Bioinformatics and Biomedicine*, pages 15–22. 2007.
- [437] Abatangelo L et al. Comparative study of gene set enrichment methods. *BMC Bioinformatics*, 10(1):275, 2009.
- [438] Krallinger M et al. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol*, 9(Suppl 2):S8, 2008.
- [439] Snel B et al. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res*, 28(18):3442, 2000.
- [440] Jenssen T et al. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet*, 28(1):21, 2001.
- [441] Watts D and Strogatz S. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440, 1998.
- [442] Bonacich P and Lloyd P. Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, 23(3):191, 2001.
- [443] Apweiler R, Attwood T, Bairoch A, Bateman A, et al. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res*, 29(1):37, 2001.
- [444] Jonsson P and Bates P. Global topological features of cancer proteins in the human interactome. *Bioinformatics*, 22(18):2291, 2006.
- [445] Vogelstein B and Kinzler K. Cancer genes and the pathways they control. *Nat Med*, 10(8):789, 2004.
- [446] Xenarios I et al. DIP: the database of interacting proteins. *Nucleic Acids Res*, 28(1):289, 2000.
- [447] Bader G, Donaldson I, Wolting C, Ouellette B, et al. BIND—the biomolecular interaction network database. *Nucleic Acids Res*, 29(1):242, 2001.
- [448] Peri S et al. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res*, 32(Database Issue):D497, 2004.
- [449] Olmea O, Rost B, and Valencia A. Effective use of sequence correlation and conservation in fold recognition. *J Mol Biol*, 293(5):1221, 1999.
- [450] Baserga R, Peruzzi F, and Reiss K. The IGF-1 receptor in cancer biology. *Int J Cancer*, 107(6):873, 2003.
- [451] Haluska P, Carboni J, Loegering D, Lee F, et al. In vitro and in vivo antitumor effects of the dual insulin-like growth factor-I/insulin receptor inhibitor, BMS-554417. *Cancer Res*, 66(1):362, 2006.
- [452] Houldsworth J, Mathew S, Rao P, Dyomina K, et al. REL proto-oncogene is frequently amplified in extranodal diffuse large cell lymphoma. *Blood*, 87(1):25, 1996.
- [453] Pico A, Kelder T, Van Iersel M, Hanspers K, et al. WikiPathways: pathway editing for the people. *PLoS Biol*, 6(7), 2008.
- [454] Life Technologies. Invitrogen iPath, 2011. <http://escience.invitrogen.com/ipath>.

- [455] Natarajan M, Lin K, Hsueh R, Sternweis P, et al. A global analysis of cross-talk in a mammalian cellular signalling network. *Nat Cell Biol*, 8(6):571, 2006.
- [456] Kelley R and Ideker T. Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol*, 23(5):561, 2005.
- [457] Ulitsky I and Shamir R. Pathway redundancy and protein essentiality revealed in the *Saccharomyces cerevisiae* interaction networks. *Mol Syst Biol*, 3(1), 2007.
- [458] Ma X, Tarone A, and Li W. Mapping genetically compensatory pathways from synthetic lethal interactions in yeast. *PLoS ONE*, 3(4):e1922, 2008.
- [459] Brady A, Maxwell K, Daniels N, and Cowen L. Fault tolerance in protein interaction networks: Stable bipartite subgraphs and redundant pathways. *PLoS ONE*, 4(4):e5364, 2009.
- [460] Cerami E, Demir E, Schultz N, Taylor BS, et al. Automated network analysis identifies core pathways in glioblastoma. *PLoS ONE*, 5(2):e8918, 2010.
- [461] Nitsch D, Tranchevent L, Thienpont B, Thorrez L, et al. Network analysis of differential expression for the identification of disease-causing genes. *PLoS ONE*, 4(5):e5526, 2009.
- [462] Aerts S, Lambrechts D, Maity S, Van Loo P, et al. Gene prioritization through genomic data fusion. *Nat Biotechnol*, 24(5):537, 2006.
- [463] Limviphuvadh V, Tanaka S, Goto S, Ueda K, et al. The commonality of protein interaction networks determined in neurodegenerative disorders (NDDs). *Bioinformatics*, 23(16):2129, 2007.
- [464] Yu W, Clyne M, Khoury M, and Gwinn M. Phenopedia and Genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations. *Bioinformatics*, 26(1):145, 2010.
- [465] Ravetti M and Moscato P. Identification of a 5-protein biomarker molecular signature for predicting Alzheimer's disease. *PLoS ONE*, 3(9):e3111, 2008.
- [466] Moscato P, Berretta R, Hourani M, Mendes A, et al. Genes related with alzheimers disease: A comparison of evolutionary search, statistical and integer programming approaches. In F Rothlauf, J Branke, S Cagnoni, DW Corne, R Drechsler, Y Jin, P Machado, E Marchiori, J Romero, GD Smith, and G Squillero, editors, *Applications on Evolutionary Computing*, volume 3449 of *Lect. Notes Comput. Sci.*, pages 84–94. Springer-Verlag Berlin, Heidelberg, 2005.
- [467] Moscato P, Gomez-Ravetti M, Rosso O, and Berretta R. Towards a multimarker molecular signature that identifies biomarkers that correlated with cognitive decline in Alzheimer's disease: The entropic perspective. *Alzheimers and Dementia*, 6(4):S530, 2010.
- [468] Zheng H, Wei D, Zhang R, Wang C, et al. Screening for New Agonists Against Alzheimers Disease. *Med Chem*, 3(5):488, 2007.
- [469] Wood L, Parsons D, Jones S, Lin J, et al. The Genomic Landscapes of Human Breast and colorectal Cancers. *Science*, 318(5853):1108, 2007.
- [470] Jones S, Zhang X, Parsons D, Lin J, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*, 321(5897):1801, 2008.
- [471] Parsons D, Jones S, Zhang X, Lin J, et al. An integrated genomic analysis of human glioblastoma multiforme. *Science*, 321(5897):1807, 2008.
- [472] Cheng H, Gao Q, Jiang M, Ma Y, et al. Molecular cloning and characterization of a novel human protein phosphatase, LMW-DSP3. *Int J Biochem Cell Biol*, 35(2):226, 2003.
- [473] Melhuish T, Gallo C, and Wotton D. TGIF2 interacts with histone deacetylase 1 and represses transcription. *J Biol Chem*, 276(34):32109, 2001.
- [474] Guo Y, Hastie T, and Tibshirani R. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, 8(1):86, 2007.
- [475] Alexe G et al. Breast cancer prognosis by combinatorial analysis of gene expression data. *Breast Cancer Res*, 8(4):R41, 2006.
- [476] Subramanian A et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*, 102(43):15545, 2005.
- [477] Efron B and Tibshirani R. On testing the significance of sets of genes. *Ann Appl Stat*, 1(1):107, 2007.
- [478] Goeman J et al. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93, 2004.
- [479] Jirapech-Umpai T and Aitken S. Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. *BMC Bioinformatics*, 6(1):148, 2005.
- [480] Stattin P and Kaaks R. Prostate cancer, insulin, and androgen deprivation therapy. *Br J Cancer*, 89(9):1814, 2003.
- [481] Hsing A et al. Prostate cancer risk and serum levels of insulin and leptin: a population-based study. *J Natl Cancer Inst*, 93(10):783, 2001.
- [482] Qiang Y, Endo Y, Rubin J, and Rudikoff S. Wnt signaling in B-cell neoplasia. *Oncogene*, 22(10):1536, 2003.
- [483] Lustig B and Behrens J. The Wnt signaling pathway and its role in tumour development. *J Cancer Res Clin Oncol*, 129(4):199, 2003.
- [484] Lin W, Hsueh H, and Chen J. Power and sample size estimation in microarray studies. *BMC Bioinformatics*, 11(1):48, 2010.
- [485] Churchill G. Fundamentals of experimental design for cDNA microarrays. *Nat Genet*, 32:490, 2002.

- [486] Van Hulse J, Khoshgoftaar T, Napolitano A, and Wald R. Feature Selection with High-Dimensional Imbalanced Data. In *IEEE International Conference on Data Mining Workshops*, pages 507–514. 2009.
- [487] Obajimi O, Keen J, and Melera P. Inhibition of de novo purine synthesis in human prostate cells results in ATP depletion, AMPK activation and induces senescence. *Prostate*, 69(11):1206, 2009.
- [488] Adler D and Murdoch D. *rgl: 3D Visualization Device System (OpenGL)*, 2010. R package version 0.91.
- [489] Feng D and Tierney L. *misc3d: Miscellaneous 3D Plots*, 2010. R package version 0.7-0.
- [490] Feng D and Tierney L. Computing and Displaying Isosurfaces in R. *J Stat Software*, 28(1):1, 2008.
- [491] Sarkar D. *lattice: Multivariate Data Visualization with R*. Springer Verlag, New York, NY, USA, 2008.
- [492] Sarkar D. *lattice: Lattice Graphics*, 2010. R package version 0.18-8.
- [493] Ligges U and Maechler M. *scatterplot3d: 3D Scatter Plot*, 2010. R package version 0.3-30.
- [494] Ligges U and Maechler M. scatterplot3d – an r package for visualizing multivariate data. *J Stat Soft*, 8(11):1, 2003.
- [495] Marchini JL, Heaton C, and Ripley BD. *fastICA: FastICA Algorithms to Perform ICA and Projection Pursuit*, 2010. R package version 1.1-11.
- [496] Blackhall F, Pintilie M, Wigle D, Jurisica I, et al. Stability and heterogeneity of expression profiles in lung cancer specimens harvested following surgical resection. *Neoplasia*, 6(6):761, 2004.
- [497] Rouzier R, Perou C, Symmans W, Ibrahim N, et al. Breast cancer molecular subtypes respond differently to preoperative chemotherapy. *Clin Cancer Res*, 11(16):5678, 2005.
- [498] Perou C, Sørlie T, Eisen M, van de Rijn M, et al. Molecular portraits of human breast tumours. *Nature*, 406(6797):747, 2000.
- [499] Sørlie T, Perou C, Tibshirani R, Aas T, et al. Gene expression patterns of breast carcinomas distinguish tumour subclasses with clinical implications. *Proc Natl Acad Sci USA*, 98(19):10869, 2001.
- [500] Sørlie T, Tibshirani R, Parker J, Hastie T, et al. Repeated observation of breast tumour subtypes in independent gene expression data sets. *Proc Natl Acad Sci USA*, 100(14):8418, 2003.
- [501] Bertucci F, Finetti P, Rougemont J, Charafe-Jauffret E, et al. Gene expression profiling identifies molecular subtypes of inflammatory breast cancer. *Cancer Res*, 65(6):2170, 2005.
- [502] Lønning P. Tailored targeted therapy for all: a realistic and worthwhile objective? *Breast Cancer Res*, 11(Suppl 3):S7, 2009.
- [503] Longo R, Torino F, and Gasparini G. Targeted therapy of breast cancer. *Curr Pharm Des*, 13(5):497, 2007.
- [504] Nielsen T, Hsu F, Jensen K, Cheang M, et al. Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. *Clin Cancer Res*, 10(16):5367, 2004.
- [505] Finlin B, Gau C, Murphy G, Shao H, et al. RERG is a novel ras-related, estrogen-regulated and growth-inhibitory gene in breast cancer. *J Biol Chem*, 276(45):42259, 2001.
- [506] Habashy H, Powe D, Rakha E, Ball G, et al. Forkhead-box a1 (foxa1) expression in breast cancer and its prognostic significance. *European Journal of Cancer*, 44(11):1541, 2008.
- [507] Abd El-Rehim D, Ball G, Pinder S, Rakha E, et al. High-throughput protein expression analysis using tissue microarray technology of a large well-characterised series identifies biologically distinct classes of breast cancer confirming recent cdna expression analyses. *International journal of cancer*, 116(3):340, 2005.
- [508] McCarty Jr K, Miller L, Cox E, Konrath J, et al. Estrogen receptor analyses: correction of biochemical and immunohistochemical methods using monoclonal antireceptor antibodies. *Archives of pathology laboratory medicine*, 109(8):716, 1985.
- [509] Shu J, Jelinek J, Chang H, Shen L, et al. Silencing of bidirectional promoters by DNA methylation in tumorigenesis. *Cancer Res*, 66(10):5077, 2006.
- [510] Li H, Myeroff L, Smiraglia D, Romero M, et al. SLC5A8, a sodium transporter, is a tumour suppressor gene silenced by methylation in human colon aberrant crypt foci and cancers. *Proc Natl Acad Sci USA*, 100(14):8412, 2003.
- [511] Toyota M, Ho C, Ohe-Toyota M, Baylin S, et al. Inactivation of CACNA1G, a T-type calcium channel gene, by aberrant methylation of its 5' CpG island in human tumours. *Cancer Res*, 59(18):4535, 1999.
- [512] Markowitz S and Bertagnolli M. Molecular basis of colorectal cancer. *N Engl J Med*, 361(25):2449, 2009.
- [513] Wolpert D. The lack of a priori distinctions between learning algorithms. *Neural Comput*, 8(7):1341, 1996.
- [514] Al-Shahrour F, Diaz-Uriarte R, and Dopazo J. Fatigo: a web tool for finding significant associations of gene ontology terms with groups of genes. *Bioinformatics*, 20(4):578, 2004.
- [515] Goh K, Cusick M, Valle D, Childs B, et al. The human disease network. *Proc Natl Acad Sci USA*, 104(21):8685, 2007.
- [516] Sill J, Takacs G, Mackey L, and Lin D. Feature-Weighted Linear Stacking. *Arxiv preprint arXiv09110460*, 2009.
- [517] Bachmann K. Genotyping and phenotyping the cytochrome p-450 enzymes. *Am J Ther*, 9(4):309, 2002.
- [518] Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence*, volume 14, pages 1137–1145. Citeseer, 1995.
- [519] Shaffer J. Multiple hypothesis testing. *Annu Rev Psychol*, 46(1):561, 1995.
- [520] Cardoso J. Infomax and maximum likelihood for blind source separation. *IEEE Signal Process Lett*, 4(4):112, 2002.

# Appendix

## 11.1 Glossary

All terms in the glossary appear underlined when they first occur in the thesis.

### 11.1.1 Biological terminology

- *Gene*: A gene is a stretch of DNA encoding functional biomolecules and corresponding to a unit of heredity in living organisms.
- *Allele*: An allele is an alternative form of a gene occupying a given position on a specific chromosome. Individuals inherit two alleles for each gene, one from each parent.
- *Genotype/phenotype*: The genotype is an organism's full gene-based hereditary information (excluding only epigenetic information, see below). The phenotype corresponds to the observable properties of an organism, i.e. the expressed hereditary information resulting in a specific morphology, development or behaviour.
- *Autosomes/gonosomes*: Sex-linked chromosomes are called gonosomes, whereas all other chromosomes are called autosomes. In the human species the gonosomes are the chromosomes X and Y.
- *Genome*: The genome is the entirety of an organism's hereditary information that is encoded in DNA or in RNA for some viruses. It includes both genetic and non-coding nucleotide sequences, but does not include epigenetic information (see *Epigenome* below).
- *Epigenome*: The epigenome is the entirety of an organism's hereditary information encoded by other structures than genomic DNA and RNA sequences. Epigenetic inheritance mechanisms include DNA methylation and chromatin remodelling (e.g. histone acetylation/de-acetylation), RNA signalling and certain transcription factor activity.
- *Transcriptome*: The transcriptome is the entirety of all RNA molecules (transcribed from the genes in a genome) in a certain cell type under well-defined conditions. In contrast to the genome, which is mostly static for a given cell line, the transcriptome can change in response to alterations in environmental conditions.
- *Proteome/Metabolome*: The proteome/metabolome is the entirety of proteins/metabolites in a given cell type under well-defined conditions. Similar to the transcriptome, the proteome/metabolome can vary in response to changes in environmental conditions.



- *Functional Genomics*: Functional genomics is a discipline studying the large-scale, genome-wide experimental data obtained from genomics and proteomics projects to investigate gene/protein functions and interactions. In contrast to genomics and proteomics, functional genomics focuses on dynamic functions and processes like transcription, translation, protein-protein interactions, rather than the mostly static genome or a static fingerprint of a proteome under well-defined environmental conditions.
- *Biomarker*: A biomarker (or biological marker) is a biological substance or process used as an indicator of a biological state. The most common examples are disease- and drug-related biomarkers, for the monitoring, diagnosis and prognosis of disease conditions and drug effects.
- *Polymerase-chain reaction (PCR)*: The PCR is an experimental technique in molecular biology to amplify DNA fragments by several order magnitudes, generating exact copies of the original template DNA. The procedure consists of the repeated application of three different temperature changes (cycles), including a denaturation step (separating the double-stranded template DNA into single strands), an annealing step (attaching small primer fragments to the single strands) and an extension/elongation step (synthesising a new DNA strand using the primer, the single-stranded template and heat-resistant DNA polymerase enzymes).
- *qPCR*: The quantitative real-time PCR (also Q-PCR, qrt-PCR) is an extension of the PCR method, which does not only amplify the target DNA, but also quantifies its original abundance. The quantification is achieved by measuring the amount of fluorescence emitted when scanning fluorescent dyes which intercalate with the double-stranded template DNA and become detectable after a certain replication cycle. The quantification is typically more precise in comparison to high-throughput quantification methods (see *Microarray* below), and enables both relative and absolute quantification of the original nucleic acid abundances for different samples.
- *Microarray*: A microarray is a miniaturized chip which allows experimenters to make thousands of biomolecular measurements in parallel using only a small amount of probe material. Depending on the type of the probe material, microarrays can be employed for different purposes, e.g. transcriptomics (DNA microarray) or proteomics (Protein microarray) data analysis. In this thesis, the term “microarray” will mostly be used to refer to DNA gene expression microarrays, rather than SNP arrays (see below) and protein arrays (the latter will always be referred to by the full name).
- *SNP, “snips”*: Single-nucleotide polymorphisms (SNPs) are DNA sequence variations affecting a single nucleotide in a genome, but occurring more frequently in a population than mutations (as a rule of thumb, if the frequency of occurrence of the variant is greater than 1%, and there are at least two variants, the variant is considered as a polymorphism and not as a mutation [517]).
- *Target DNA/probe DNA*: On a DNA microarray, the probe DNA corresponds to fragments of DNA which are immobilised on the chip surface and typically represent genes from single species’ genome. The target DNA corresponds to the nucleic acid sequences in the biological samples to be analysed, which are labelled with a fluorescent dye, and hybridised with the probe DNA on the chip to measure the relative abundance of corresponding fragments within the target DNA using a laser scanner and a CCD chip detector.
- *Probe replicates*: On a DNA microarray, the probe DNA representing a specific gene does typically not only occur once on the chip, but multiple times, to obtain a more accurate quantification of transcript abundance in a biological sample by summarising the measurements for all replicates.

- *Homology*: Homologous traits are inherited characteristics of organisms derived from a common ancestor, i.e. resulting from a divergent evolutionary process as opposed to analogous traits corresponding to similarities between organisms resulting from convergent evolution. Homology can be analysed using bioinformatics methods to detect significant similarities in DNA or protein sequences (typically employing sequence alignment and hierarchical clustering techniques).
- *Orthology/paralogy*: Homologous sequences are *orthologous* if they were separated by a speciation into two separate species (sequences with similar function, but occurring in different organisms). Vice-versa, sequences are *paralogous* if they were separated by a gene duplication event (homologous sequences with different functions, but occurring in the same organism).
- *Yeast-2-Hybrid (Y2H)*: Y2H is a molecular biology screening methodology to discover protein-protein and protein-DNA interactions by testing for direct physical interactions between the corresponding molecules. The idea behind the method is to fuse one of the proteins to be tested with the activating domain (AD) of a transcription factor (TF) and the other with the corresponding binding domain (BD), and since these domains work in a modular fashion, an interaction of the target proteins is likely to activate the transcription of a reporter gene by bringing these TF domains close together. In the case of a protein-DNA interaction (also called One-Hybrid), a single fusion protein is used, in which the AD is directly linked to the BD. Importantly, the Y2H method is error-prone, and can provide both false positive and false negative results.
- *Tandem affinity purification (TAP)*: TAP is an alternative technique for identifying protein-protein interactions (PPIs) and protein complexes. A fusion protein with a designed end tag, the TAP tag, is created, so that it binds to beads coated with Immunoglobulin G (IgG) antibody molecules. This enables the extraction of this protein in combination with its binding partners using a washing procedure with two affinity columns. The binding partners can then be identified using other experimental techniques, e.g. mass spectrometry or SDS-PAGE. Importantly, since proteins might also bind indirectly to the target protein, e.g. by only binding to one of its binding partners, TAP can only identify indirect PPIs and protein complexes, but not direct binary interactions like Y2H. Moreover, adding a tag to a protein might also negatively affect the binding of some potential interaction partners, hence, this method is also error-prone.

### 11.1.2 Statistical terminology

- *Classification and regression*: Classification techniques in statistics are methods that predict categorical, nominal outputs (class labels) for an observation, and regression techniques are methods that predict numerical, continuous-valued outputs. Depending on whether the predictions are made using training data with known or unknown outputs, these approaches are also called *supervised* (data with known outputs), *semi-supervised* (both data with known and unknown outputs) or *unsupervised* (data with unknown outputs) classification and regression methods.
- *sensitivity/specificity*: In the result of a binary classification, the sensitivity is the proportion of positively labelled samples which are correctly identified (sensitivity = true positives / (true positives + false negatives)), and the specificity is the proportion of correctly identified negative samples (specificity = true negatives / (true negatives + false positives)).
- *Receiver Operating Characteristic (ROC)*: The ROC curve is a plot of the true positive rate (= sensitivity) against the false positive rate (= 1 - specificity) of a binary classification model with varying

discrimination threshold applied to a test dataset with known class labels. This plot assists the experimenter in comparing the performance of different classification methods and in choosing a suitable discrimination threshold to obtain a good balance between sensitivity and specificity. Moreover, the area under the ROC curve (AUROC) is a frequently used performance measure for machine learning methods, and in contrast to the average classification accuracy, it accounts for biased predictions resulting from class imbalances in the training data.

- *“Type 1” and “type 2” errors*: In statistical hypothesis testing and binary classification models two types of errors can be made. Type 1 errors are false positives or incorrect rejections of a null hypothesis, whereas type 2 errors are false negatives or failed rejections of a null hypothesis that is false.
- *P-value*: The p-value in statistical significance testing is the probability of obtaining a value for a test statistic which is at least as extreme as the actually observed value, assuming that the null hypothesis is true (i.e. the probability for observing the value or a more extreme value by chance and not because the null hypothesis is false). The lower the p-value, the less likely is the observed value if the null hypothesis is true, and therefore the more significant the observation is.
- *Cross-validation (CV)*: Cross-validation is a statistical method to evaluate the performance of a predictive model on training data, by partitioning the samples in the data into  $n$  approx. equal-sized subsets and using each combination of  $n - 1$  subsets to train a predictive model, which is then evaluated on the left-out subset. The average accuracy or average area under the ROC curve across all cross-validation cycles provides nearly unbiased performance estimates [518], which tends to be closer to the performances obtained on large external test sets than the often overly optimistic performance estimates obtained when applying the model on the training data (training accuracy or training AUROC). If  $n$  is equal to the total number of samples, the method is also called leave-one-out CV, or LOOCV.
- *Data fusion*: Data fusion techniques are approaches which combine data from different sources, while at the same time reducing the total amount of data, e.g. by removing redundant information, exploiting synergies and replacing low-confidence data with high-confidence data (in contrast to the more general term *data integration*, which also refers to integrative methods without a reduction step). The assumption behind data fusion is that the individual input data sources contain incomplete but complementary information, so that some of the information gaps in the single data sources can be filled by aggregating the information from all inputs.
- *False-discovery rate (FDR), q-value*: The FDR is the expected proportion of false discoveries (type 1 errors) among all statistically significant hypotheses identified by a hypothesis test. It is used to adjust a hypothesis test in a multiple testing scenario. Given a number of false positives  $\#fp$  and a number of all positives  $\#ap$ , the FDR can be written as:  $FDR = E[\#fp/\#ap]$ . The *q-value* for a specific hypothesis test is the minimum FDR threshold at which the test would be regarded as significant.
- *Familywise error rate (FWER)*: The FWER is the probability of making one or more false discoveries (type 1 errors) among all hypotheses in a multiple testing scenario. Given a number of false positives  $\#fp$ , the FWER can be written as:  $FWER = Pr(\#fp > 0) = 1 - Pr(\#fp = 0)$ . Like the FDR, the FWER is used to adjust hypothesis tests for multiple comparisons, and has a stronger control of type 1 errors, but less statistical power than the FDR method [519]. FWER methods tend to be more conservative than FDR approaches, because the FDR can be rewritten as:  $FDR = E[\#fp/\#ap | \#fp > 0] Pr(\#fp > 0) = E[\#fp/\#ap | \#fp > 0] \cdot FWER$ , and given that  $\#fp/\#ap \leq 1$  it follows that FDR

$\leq$  = FWER.

- *Bagging*: Bagging or “bootstrap aggregating” is a special model averaging and ensemble technique in machine learning to improve classification and regression models in terms of accuracy and robustness. Given a training dataset with  $n$  samples, bagging generates  $m$  new training datasets of size  $n' \leq n$ , by sampling uniformly from the original data with replacement (generating so-called bootstrap samples). Finally,  $m$  machine learning models are fitted using the bootstrap samples and their prediction results are combined using averaging (for regression) or majority voting (for classification).
- *Boosting*: Boosting is a further model averaging and ensemble technique in machine learning to improve classification and regression models by combining weak learning algorithms (classifying samples only slightly better than a random classifier) into a more robust and accurate ensemble model. There are many specific boosting algorithms, but most of them iteratively add weighted weak learners to an additive ensemble model, where the weight depends on the weak learners accuracy in the classification of weighted samples (the sample weights typically increase for samples that were misclassified in previous iterations of the algorithm, to force future learners to improve the classification for these samples). Examples for boosting algorithms are AdaBoost [297], LogitBoost [298] and Linear Programming Boosting (LPBoost) [299], among others.
- *Principal Component Analysis (PCA)*: PCA is a dimensionality reduction and orthogonal data transformation method used to convert a set of potentially correlated variables into a (smaller or equal-sized) set of uncorrelated, derived variables termed “principal components” (PCs). The PCs correspond to the eigenvectors of the data covariance matrix and can therefore be computed by the eigenvalue decomposition of the covariance matrix, or alternatively, using singular value decomposition (SVD) of the original data matrix. The PCA decomposition has the special property, that the first PC and every succeeding PC covers as much of the data variance as possible, so that low-dimensional representations can cover a large degree of the variance (and thus a large degree of the information content) in the data, in spite of a potentially very high original dimensionality of the data.
- *Independent Component Analysis (ICA)*: ICA is a dimensionality reduction and data transformation method used to convert a multivariate signal into additive components, assuming that the original signal consists of a mixture of mutually independent non-Gaussian source signals (the *independent components*). To identify the independent components, the statistical independence of the estimated components is maximised, and ICA algorithms for this purpose differ mainly in the definition of independence, e.g. minimising the mutual information between the components, or maximising the non-Gaussianity of the components (whose mixture is assumed to have become more Gaussian according to the central limit theorem). Popular algorithms for ICA include fastICA [495] and infomax [520] among others.
- *Expectation maximisation (EM) algorithm*: The EM algorithm is an iterative approach for finding maximum likelihood estimates for parameters in statistical models which depend on unobserved latent variables. The generic algorithm iteratively repeats two steps: An expectation step (E-step), which computes the expectation of the log-likelihood given current estimates for the latent variables, and a maximisation step (M-step), which computes parameters maximising the expected log-likelihood found on the E-step (initially, the parameters are often set to random values). Next, these parameter estimates are used to determine the distribution of the latent variables in the following E-step, and the procedure continues until the parameter estimates converge.

## 11.2 Example flowcharts for new integrative analysis pipelines

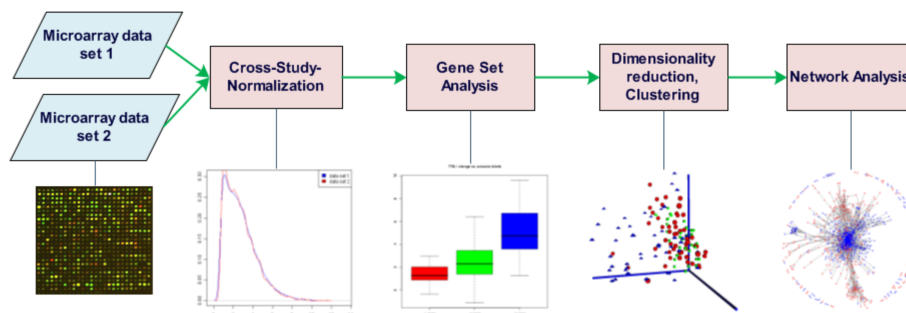


Figure 11.1: Example flowchart illustrating combinations of analysis modules within the web-application ArrayMining. Data from two microarray studies is combined using cross-study normalisation (output: density plot), gene sets representing cellular pathways are extracted (output: box plots for differentially regulated gene sets), the pathway expression matrix is clustered (output: low-dimensional cluster representations) and a network of co-expressed pathway expression fingerprints is computed (output: network visualisation).

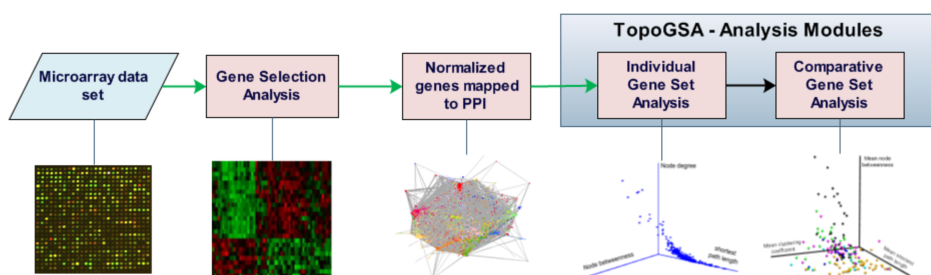


Figure 11.2: Example flowchart illustrating a cross-domain combination between modules from ArrayMining and TopoGSA. Differentially expressed genes from a microarray study are identified using the ArrayMining Gene Selection Analysis module (output: heat map of differentially expressed genes) and mapped to a protein-protein interaction (PPI) network for a network topological analysis with TopoGSA, revealing outstanding topological properties of single genes (output: Individual Gene Set Analysis plot) and similarities between the selected genes and known gene sets from functional annotation databases (output: Comparative Gene Set Analysis plot).

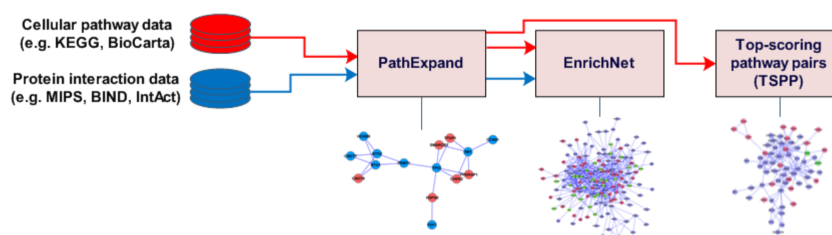


Figure 11.3: Example flowchart illustrating a cross-domain modular combination between PathExpand, EnrichNet and the Top-scoring pathway pairs method (TSPP). Protein interaction data and cellular pathway data is used by PathExpand to create extended pathways with compact network representations (output: new pathways), which are investigated for functional associations with known and experimentally derived gene sets using EnrichNet (output: similarity ranking and network visualisations) and used for sample classification within TSPP (output: classification model and network representations).