# Network-Based Biomarker Discovery:

## Development of Prognostic Biomarkers for Personalized Medicine by Integrating Data and Prior Knowledge

**Dissertation**

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

**Yupeng Cun**

aus

Yunnan, China

Bonn, 2014

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter: Prof. Dr. Holger Fröhlich, Universität Bonn

2. Gutachter: Prof. Dr. Armin B. Cremers, Universität Bonn

Tag der Promotion:

Erscheinungsjahr:

Abstract

Advances in genome science and technology offer a deeper understanding of biology while at the same time improving the practice of medicine. The expression profiling of some diseases, such as cancer, allows for identifying marker genes, which could be able to diagnose a disease or predict future disease outcomes. Marker genes (biomarkers) are selected by scoring how well their expression levels can discriminate between different classes of disease or between groups of patients with different clinical outcome (e.g. therapy response, survival time, etc.). A current challenge is to identify new markers that are directly related to the underlying disease mechanism.

During the last years, an increasing number of tools have been developed to derive biomarkers from gene expression data. These methods typically involve machine learning approaches, like support vector machines, decision trees, neural networks or linear discriminant analysis. Currently, a general problem is that biomarker gene signatures have a low reproducibility and are difficult to interpret biologically. It has been shown that robustness, stability and biological interpretability of biomarker gene signatures can be significantly improved by incorporating biological knowledge, such as protein-protein interaction networks.

In this thesis, we first compared a collection of published gene selection methods, of which some include network information. Our results show that incorporating prior knowledge of network information into gene selection method in general does not significantly improve classification accuracy, but greatly enhances the interpretability of gene signatures compared to classical algorithms. In a next step we developed a new method, called stSVM, which integrates both, network information as well as gene and microRNA expression profiles, into one classifier. This new approach not only shows superior prediction performance, but also stability and interpretability of selected features. An open source software, called netClass, was developed for implementing the proposed feature selection algorithm.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

"A mathematician is a device for turning coffee into theorems."

– *Paul Erdős.*

## 1.1 Personalized medicine

In the past decades, the topic "personalized medicine" has gained much attention, and it is defined as "A form of medicine that uses information about a person's genes, proteins, and environment to prevent, diagnose, and treat disease" [NCI13]. Personalized medicine is a rapidly advancing filed of health care which became relevant after the completion of the Human genome projects [LLB+01, VAM+01, MMG13]. Genomic variations (such as mutations in the BRCA1 gene) can lead to an increased risk to develop disease like cancer. Knowledge of these genomic variations is therefore useful for risk assessment and optimal therapy design. The principle aim of personalized medicine is to optimize medical, and to

a new era: prescribe the right drug combinations for right patient groups at the right dosage.

Personalized medicine lies in the intersection between three domains: personal genomics, pharmacogenomics and medicine (Figure 1.1). Personal genomics deals with the wealth of genomic information of individuals, For this purpose omics data of various kinds is employed, such as whole genome sequence, transcriptomics, proteomics, metabolomics, epigenomics, etc.. Pharmacogenomics is using genomics variabilities as biomarkers that determine or predict response to drugs. The final goal of personalized medicine is to transfer the knowledges from personal genomics and Pharmacogenomics to practice. Genome-based information amplify our understanding of diseases mechanisms. Moreover, molecular information enables to implement refined treatment schemes and to assess an individual's risk for a specified disease. It potentially also allows for preventing the outbreak of diseases via early prevention

Personalized medicine aims to tailor patients to individualize care based on patient's molecular profiles [GW08]. Personalized medicine take the discovery in biomedicine research to facilitate highly precise health care. One of the major goals is to identify reliable molecular biomarkers that predict a patient's response to therapy, including potential adverse effects, in order to avoid ineffective treatment and to reduce drug side-effects and associated costs. A biomarker, or biological marker, is a marker for a biological state, which is measured and evaluated as an indicator for a specific status of a biological processes, pathogenic process, or pharmacologic response to a therapeutic intervention (Biomarker Definitions Working Group, 2001). In the context of biomedical research, biomarkers

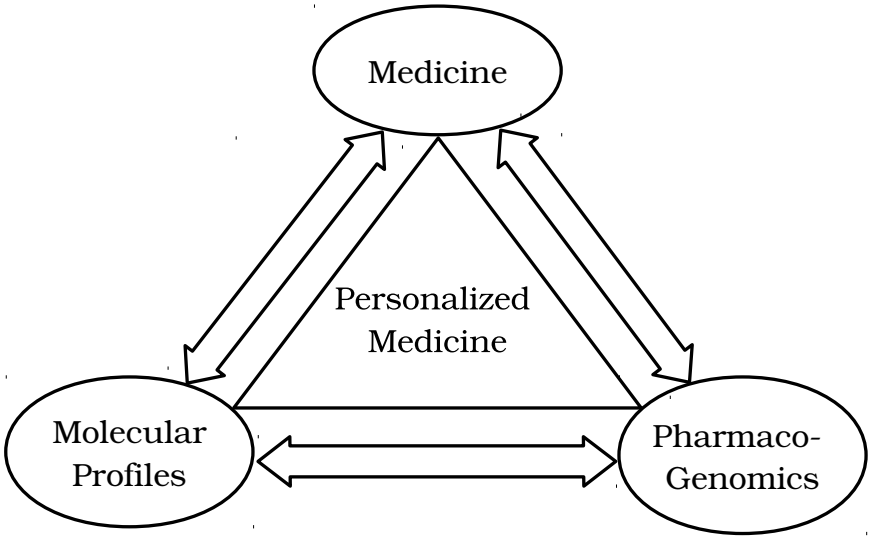



Figure 1.1: Personalized medicine. Adopt from [FCD+11].

can be further distinguished into predictive (allowing to forecast a patient's respond to a clinical treatment), prognostic (allowing to forecast a patient's disease outcome regardless of the type of treatment). Predictive, stable and interpretable gene signatures are generally seen as an important step towards a better personalized medicine. During the last decade various methods have been proposed for that purpose. Such approaches highly depend on molecular profiles of patients, which could be used to precisely predict patient's risk of disease. Detailed overviews on current successful approaches in personalized medicine are given by [GW08, GW+09].

A famous example for a personalized cancer therapy is the application of Cetuximab. Cetuximab binds to the EGF receptor and, consequently, prevents activation of the downstream signaling pathway, thus inhibiting cell proliferation [VCKH+09, VMR+08]. However, it has been found that

Cetuximab can only work, if the K-RAS gene is not mutated [VMR+08, BCR+12]. Testing patients for mutations of this gene in the European Union is thus prescribed before application of Cetuximab to prevent a costly and ultimately ineffective therapy. Another example is the anti-cancer drug Trastuzumab, which is only effective in patients that express highly the human epidermal growth factor (HER2) at the cell surface, to which the antibody binds [Hud07]. Recently, next generation sequencing (NGS) has been applied to cancer patients to identify new markers, and to uncover the mechanisms of therapy resistance or sensitivity. Prognostic or diagnostic biomarker signatures (mostly from gene expression data, but more recently also from other data types, such as microRNA, methylation patterns or copy number alterations) have been derived in numerous publications for various disease entities. One of the best known ones is a 70-gene signature for breast cancer prognosis (mammaprint) by [vtVDvdV+02], which has gained FDA approval.

Nowadays modern high-throughput technologies allow for screening of massive amounts of omic-type data, and so one goal is to associate such data with a patient's clinical prognosis or with the membership to a certain disease subtype. Based on omics data it has been even possible to identify novel disease subtypes. For example, based on gene expression profiles, five subtypes of breast cancer have been identified [SPT+01]. In 2006, NIH launched The Cancer Genome Atlas (TCGA) for deciphering the genomics and epigenomic landmark of more than 20 cancers. And later, another big world-wide collaborating project, the International Cancer Genome Consortium (ICGC), was started with the goal of charactering the molecular profiles of 50 cancers with larger tumor samples. The samples in these study accompanied with relevant clinical features generate molecular profiles which contain genomic variations, transcrip-

tome microRNAs profiles and epigenomics methylation profiles. Most data from these two projects are available to public access and could someday advance clinical practice.

## 1.2 Machine learning approaches for biomarker discovery

Recent fast progress in high-throughput technologies have led to an dramatic increase of the potential data to find meaningful causalities of disease mechanisms, including interactions between gene-gene and gene-environment. In order to mine such large data collections, efficient machine learning methods are required. Similarly, there is an urgent need for integration of different kinds of available molecular data of the same patient to improve and find robust biomarkers for clinical outcomes, which would translate personalized medicine into reality.

Machine learning is about the design of a system that can learn from experience and could be expressed as "How can we program systems to automatically learn and to improve with experience?" according to Tom MitChell [Mit97]. Statistical learning theory gives a theoretical framework for machine learning, which arises from statistics and functional analysis. Machine learning has led to lots of applications of text mining, computer vision, natural language process, artificial intelligence and bioinformatics [Vap00, HTF08, MRT12].

Statistical learning provides tools to understand data and these tools are general classified into two categories: supervised and unsupervised

learning. The most significant difference between unsupervised and supervised learning is that class labels of input data are available or not. In supervised learning, the goal is to train a model that can predict an outcome (e.g. a class label) based on available input data; in unsupervised learning, the goal is to discover patterns from the input data without class information. Figure 1.2 depicts the learning process of unsupervised and supervised learning. In Figure 1.2, unsupervised learning discovers two classes, $a$ and $b$ from the input data; supervised learning learning uses input data with two known classes, $a$ and $b$, to make predictions for data with unknown labels.

Supervised learning generates a function that maps inputs to desired outputs (also called labels which are often defined by human experts labeling the training examples). For example, in a classification problem, the learner approximates a function mapping a vector on to classes by looking at input-output examples of the function. Unsupervised learning models a set of inputs, based on similarity. Here, labels are not known during training. In this thesis, we mainly focus on supervised learning. For example, we have a gene expression profiles of cancer patients with two categories (early and late relapse), and by using these dataset, a model is trained to make prognosis for an individual patient.

To address the construction of biomarker signatures, one typically uses supervised machine learning methods together with algorithms for variable / feature selection. The microarray technology nowadays enables measurement of tens of thousands of transcripts at the same time, whereas the sample size is typically in the order of 100 - 300 patients. This not only imposes high challenges for the interpretation of such data, but

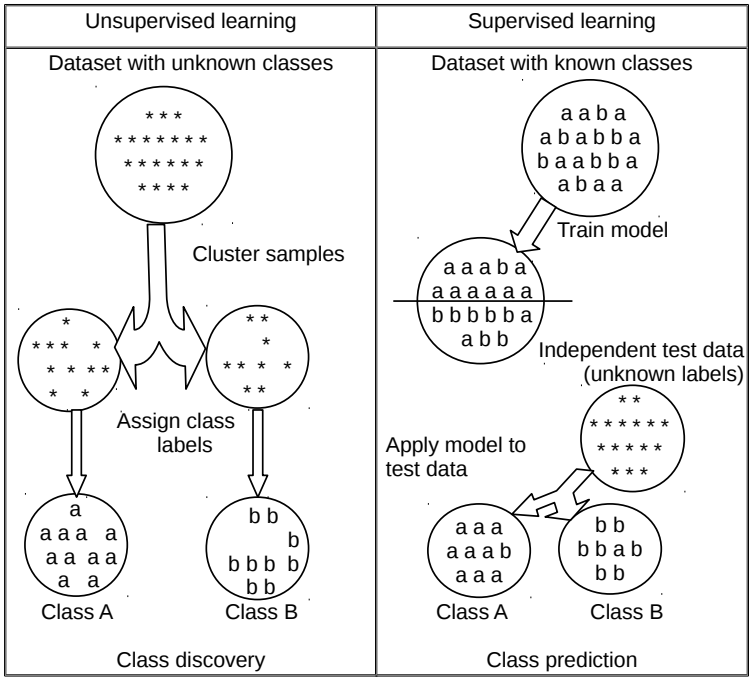



Figure 1.2: Unsupervised and supervised learning. Adapted from [RG02].

also for robust and stable statistical procedures, which are needed to detect those genes, which are truly correlated with the clinical phenotype. In this context, it should be mentioned that typical machine learning algorithms operating with far more variables / features than samples are prone to the so-called "over-fitting" phenomenon: The classifier or Cox regressor can perfectly explain the data used for model construction, but fails in making good predictions on new test data [DHS01, HTF08]. Therefore algorithms and statistical procedures for efficient reduction and selection of relevant features of the data are crucial.

Well known algorithms for this purpose are PAM [THNC02], SVM-RFE [GWBV02a], Random Forests [DUdA06a] or statistical tests, like SAM [TTC01], in combination with conventional machine learning methods (e.g. Support Vector Machines, $k$-nearest neighbor ($k$-NN), Linear dis-

criminant analysis (LDA), logistic regression, ...). An excellent overview about these algorithms can be found in [HTF08]. Moreover, several modifications of Support Vector Machines (SVMs) for embedding gene selection into this algorithm have been proposed [WZZ08, ZALP06, BTLB11]. For associating gene expression or other high dimensional experimental or clinical data with patient survival times, typically Cox regression or variations thereof (multivariate penalized Cox regression) are employed [Goe10, BS09].

However, retrieved gene signatures are often not reproducible in the sense that inclusion or exclusion of a few patients can lead to quite different sets of selected genes. Moreover, these sets are often difficult to interpret in a biological way [Gön09]. For that reason, more recently a number of approaches have been proposed, which try to integrate prior biological knowledge on canonical pathways or protein-protein interactions into gene selection algorithms. The general hope is not only to make biomarker signatures more stable, but also more interpretable in a biological sense. This is seen as a key to making gene signatures a standard tool in clinical diagnosis [BZK11].

## 1.3 Sources of biological knowledge

A very important source of biological knowledge about individual genes regarding cellular components, involvement into biological processes and molecular functions can be obtained from the Gene Ontology database [ABB+00]. Another important aspect of biological information include molecular interactions which can be categorized into protein-protein in-

teractions (PPI), metabolic pathways, signaling pathways and gene regulatory networks.

A protein-protein interaction means that two or more proteins bind together to carry out their biological function. Interactions betweens proteins are important for most molecular processes, and play a central role in a living cells. Protein-protein interactions (PPIs) as well as canonical pathways can be retrieved easily in a computer readable format from databases, such as KEGG [KAG+08], HPRD [PKP09], PathwayCommons [CGD+11] or others. These databases contain collections of protein interactions that have been reported in the literature. In this thesis, I mainly focus on interaction networks from KEGG and PathwayCommons.

Gene regulatory networks represent interactions between transcriptional regulators (e.g. transcription factors, miRNAs) and their regulated target genes [BGL11]. In this thesis, I mainly focus on miRNA-target gene networks.

Integration of biological knowledge, specifically from protein-protein interaction networks and canonical pathways, is widely accepted as an important step to make biomarker signature discovery from high dimensional data more robust, stable and interpretable. Consequently there is an increasing amount of methodologies for this purpose. What has to be mentioned, however, is that usually these interactions have been observed under differing biological conditions and cell types. Thus a purely literature based network reconstruction will suffer from a lack of specificity with respect to the cell or tissue type under study. Moreover, false interactions can be frequently observed due to technological limitations, which are, for instance, imposed by genome scale two-hybrid or co-precipitation screens. Hence, confidence measures for interactions are

of high value [CKZ+07, GPF+11]. On the other hand it is widely believed that only a fraction of the true interactome is known so far. Despite these limitations network reconstructions have turned out to provide valuable hypotheses for biomarker signature discovery. In Section 2.7, I give a general overview about these approaches and grouped them into categories.

## 1.4 Contribution of this thesis

This thesis is motivated by the employment of feature selection methods in prognostic / diagnostic biomarker discovery. The main contributions is the development of a method that allows to integrate in one classification model :

1. biological knowledge in form of protein-protein interactions;
2. different molecular data entities, namely miRNA and mRNA expression data.

I also performed a comprehensive study on current state of art feature selection methods, which employed prior information or not.

The outline of this thesis is as follows:

In Chapter 2, some basics of molecular biology, current techniques for molecular profiling are explained. Afterwards, classification methods for high dimension data are presented together with feature selection methods. Support vector machines illustrate the problem of binary data classification. The methods for classification model assessment and selection

are also described in this section. Finally, I give a overview on current network-based approaches for gene selection.

In Chapter 3, I investigate whether network-based approach provide an advantage compared to classical approaches. I compared fourteen published gene selection methods (eight methods were network-based approaches) on six public breast cancer datasets with respect to prediction accuracy, gene selection stability and the biological interpretability of gene signatures. Incorporating prior knowledge of network information into gene selection method in general did not significantly improve classification accuracy, but could greatly enhance the interpretability of gene signatures compared to classical algorithms.

In Chapter 4, a new algorithm is proposed to integrating network information as well as mRNA and miRNA expression into one classifier. This is done by smoothing t-statistics of individual genes or miRNAs over the structure of a combined PPI and miRNA-target gene network. A permutation test is conducted to select features in a highly consistent manner, and then a SVM is employed to train a classifier. The method shows an improved on prediction performance, stability and interpretability of selected features compared to RRFE, netRank [JBF+10, WKK+12].

In Chapter 5, I describe my open source software, netClass, for network-based based feature selection. netClass implements several network-based classifiers algorithms, which are used in Chapter 3 and Chapter 4, in the R programming language and is freely available on the CRAN repository at http://cran.r-project.org.

In Chapter 6, I summarize my results on network-based biomarker discovery algorithms. Moreover, possible future research directions are pointed out.

# Chapter 2

# Background

"***E***very answer given on principle of experience begets a fresh question."

– Immanuel Kant.

THIS chapter focuses on two topics: the first part aims to give a brief overview about molecular biology and biomarker discovery. The second part introduces methodologies for high-dimensional data classification. The methodology part gives an overview about current classical classification methods for high dimensional data, with emphasis on support vector machines methods. Network-based feature selection methods are also introduced. A review about these methods has been published in ***Biology*** [CF12a].

## 2.1 Basic molecular biology

Modern molecular biology has remarkable impacted on our understanding of disease, their causes and transmissibility. A cell is the smallest

basic building block of life which contains the complete genetics information [HL03]. Deoxyribonucleic acid (DNA) is in most organisms, expect for some viruses, the carrier of the hereditary information. DNA has a double helical structure, which is encodes the genetic information via four nucleotides: guanine (G), adenine (A), thymine (T), and cytosine (C).

Genes are genetic information-bearing sections of DNA that divide a long DNA sequence into different functional units. A chromosomes is a piece of DNA that organized DNA, protein and RNA in a cell. Chromosomes are folded in the nucleus for locating most of the DNA in cell, and different chromosome associates with certain proteins. The genome of an organism is compiled of all complement of DNA. The genes commonly contains two parts: a coding part and a regulatory part. The coding part specifies a protein's amino acid sequence and the regulatory part controls when and where the protein is translated. Transcription is a segment of DNA copied into RNA, and translation is a process of the transcribed RNAs create to proteins.

Usually, cells use three complex steps to convert the DNA codes to proteins (Figure 2.1). DNA replication from itself via complementary matching rule, which means A convert to G and C to T, is the first step. The second step is transcription into a single-stranded ribonucleic acid (RNA). RNA is large enzyme that translated from DNA and composed by four nucleotides: guanine (G), adenine (A), uracil (U), and cytosine (C). The initial RNA split into smaller message RNA (mRNA) polymerase in eukaryotic cells. mRNA will transported to the cytoplasm in the next step. Ribosome is a complex assemble of RNA and protein that will motivate the translation process. The mRNA sequence translate amino acids of protein via the universal genetic code to finish the third step. This pro-

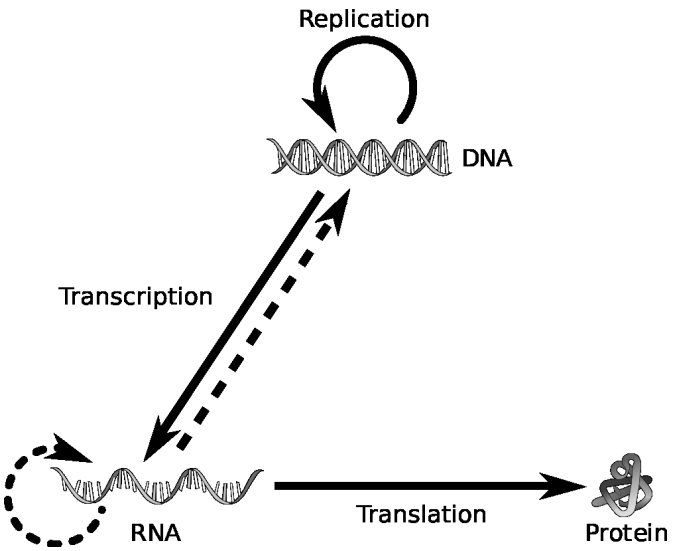



Figure 2.1: Central dogma in molecular biology. The dogma of molecular biology is an explanation for how information flow works in biological system. The solid arrow are flows occur in all cell: DNA replicate from itself; DNA transcript into RNA; RNA translate to protein. The dotted arrow show flows are occasionally occur. Images from [BEC+12] under free copy license CC-BY-SA.

cess is refereed to the central dogma in molecular biology.

The genotype is the summary of the genetic information provided by all genes in a cell. Phenotype is the observation of an organism's characteristics or traits. Phenotypes result from an organism's genotype as well as environmental stimulation. In genetics, mutation is defined as the nucleotide sequence changes in the replication process of an organism's genome. Mutation is the source of evolutionary novelty that may produce harmful or beneficial changes in the phenotype of an organism [CC+10]. Somatic mutation is a change in the genetic sequence that is not inheritable, i.e. occurs during life time, for example due to environmental factors. Genetic and epigenetic alterations can affect a gene's function and thus also indirectly phenotypes.

## 2.2 Cancer is a genetic disease

A review by Vogelstein and Kinzler [VK04] states that "The revolution in cancer research can be summed up in a single sentence: cancer is , in essence, a genetic disease". Modern technologies in the area of genome research allowed for significant advances in cancer research [Wei07, SCF09, GL13, VPV+13]. Changes in genes/genome can be use for tracing human disease. These variations can cause abnormal transformation of living cells into malignant neoplasms which overcome the normal cell pathway to uncontrolled process.

The complex process of the change of normal into cancer cells is called tumorgenesis. Tumorgenesis is also sometimes called tumorigenesis, tumor progression, carcinogenesis or oncogenesis. Some genes can completely or partially reduce the risk of tumorgenesis. These genes are called tumor suppressors. A lot of experimental effort has been undertaken to find such cancer-related genes, for example the Catalogue of Somatic Mutations in Cancer (COSMIC) database [FBB+11]. From 1970s on several oncogenes (such as SRC and BCR-ABL1 fusion gene) and tumor suppressors (such as TP53, RB) have been discovered. Later studies showed that these genes operate in canonical signaling pathways. Information about such pathways can be found in public databases, such as KEGG. [KAG+08].

Current high-throughput biotechnologies have promoted our understanding of the molecular nature of tumors. Such reteaches require to unravel the genetics variations at different molecular levels. For example, we can depict the mutation landscape via whole genome sequencing with as many samples as possible, or measure the mRNA expression profiles

of most knowns genes with different conditions. Such exhaustive measurements of molecular profiles often are often called genome-wide techniques.

## 2.3 Gene expression profiles

Gene expression is a process by which a gene's hereditary information is transcribed into RNA in the cell, which is most fundamental process by which the genotype influences the phenotype. The genetic information stored in DNA will be "interpreted" via gene expression. Expression of genes includes two steps: first is the transcription of genomic information into messenger RNA (mRNA) and then translated to protein; the second step is the translation of mRNA into proteins. Measurements of the expression of mRNA level of given genes in a tissue is widely used in biomedicine. RNAs which are not translated to protein are non-coding RNAs which may influence gene expression via post-translational regulation. They are also potential biomarkers.

In the past decades, gene expression profiling has been widely used via microarray chips that simultaneously measures the activity of thousands of genes. The transcriptome of a set of patients is widely used for measuring the biological phenomena and for discovering patterns that potentially provide insights into disease mechanisms. Moreover, gene expression profiles are used to identify diagnostic, prognostic and therapeutic biomarkers. One of the first studies on gene expression profiling showed that breast cancers could be clustered into distinct subtypes based on gene expression patterns [PSE+00]. A few years later a very successful

study identified a 70-gene signatures for breast cancer prognosis prognostic by using supervised learning [VDVHvV$^+$02, vtVDvdV$^+$02].

Apart from transcriptomics and interactomics, other omic approaches, such as genome-wide copy number variation (CNV), single nucleotide polymorphisms (SNPs), DNA methylation (epigenomics) etc, also have been widely used in oncology research. A comprehensive review on omic approaches can be found in [GW08].

The Gene Expression Omnibus (GEO) by the National Center for Biotechnology Informatics (NCBI) and ArrayExpress by the European Bioinformatics Institute (EBI) are two major public gene expression profile databases. Microarray and other types of high-throughput omics data are freely open for public download and use by the scientific community.

## 2.4 MicroRNA expression profiles

MicroRNAs (miRNAs) are small non-coding RNA molecules which were first found in Caenorhabditis elegans [LFA93]. miRNAs usually contains around twenty nucleotide-long single strand RNA molecular and serve as master regulators of gene expression via sequence-specific fashion [CR07]. miRNAs target mRNAs through fractional complementarity with their seed-specific sequence, and then insufficient mRNA translation and stability will decrease protein expression level. Their alteration in tumor have important tumor-genesis consequences. Over-expressed miRNAs in tumor lead to down-regulation of tumor suppressors or oncogenes and thus influence cancer development (see Figure 2.2).

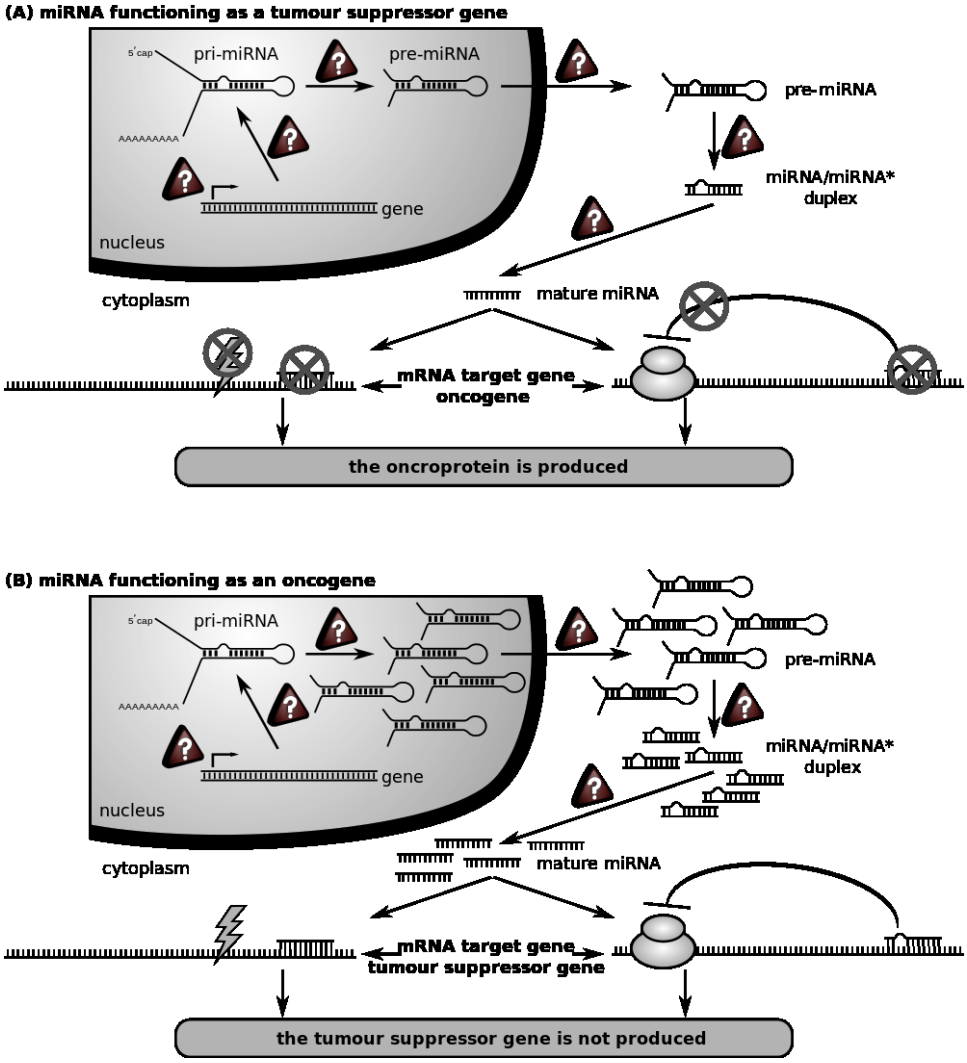



Figure 2.2: miRNA in a cancer cell. Any abnormal in miRNA expression can lead to the target protein improper translated. (A) The decrease expression or loss of a tumor suppressor miRNA leads to an abnormal high translation level of the target oncoprotein. (B) The enhance or overexpression of a oncogene miRNA leads to sweep of tumor suppressor protein. Image from [BEC+12] and adopted form [EKS06] under free copy license CC-BY-SA.

Current studies of miRNA expression profiles of cancer patients have revealed that miRNAs can server as biomarkers [LGM+05, GM12]. For example, low expression of miR-324a results in a poor survival prognosis in non-small cell lung cancer (NSCLC) [VVK+11], and the miRNA-200 family (miR-200a, miR-200b, miR-200c, miR-141 and miR-429) are down regulated during the tumor progression of breast cancer [GBP+08].

The miRBase database is a database for collecting published miRNA sequences and a major warehouse for miRNA related annotation information [GJSvDE08]. All miRNAs in miRBase are mapped to their genomic locations. The repeated and annotated transcripts of miRNA sequences are described. The latest miRBase has 24521 hairpin sequences in over 140 species, and 30424 mature sequences. The growth and development of the database provides a powerful prior tools for omics data integration.

## 2.5 Microarray technology

The revolution in biotechnologies has advanced our understanding of in vivo cellular functional process via in vitro DNA technologies [Mar11]. Microarray technology appeared in 1995, it is based on the principle of complementary hybridization of nucleotide sequences [BH02, Hel02]. The DNA microarray technology, which is also termed as DNA chip or biochip, provides microscopic sensor tools to quantify the genome-wide mRNA or miRNA expression on a tiny slide. Microarray technology has been widely applied to biological and medical research to find biomarker in many diseases [GW08].

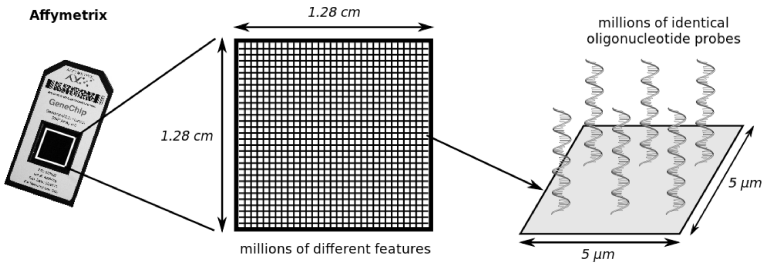



Figure 2.3: Chip designs of Affymatrix. The chip carries about 6.5 millions features. Each feature is composed by millions of identical oligonucleotide probes. Image from [BEC+12] and adopted from [DWWTM06] under free copy license CC-BY-SA.

The core principle of microarray is hybridization between two DNA strands, and the complementary property of nucleotide sequences target specifically pair with each other via forming hydrogen bonds between complementary pairs. Each probe (DNA, RNA or Protein) attaches to a fixed slide and has a specific chip, such as glass and silicon [SMS99]. Any given sequence can be assigned to the probes, so microarrays have been developed for genome, transcriptome and proteome profiling. For example, SNP array and array-comparative genomic hybridization (aCGH) are used to measure genome-wide SNPs and CNVs. In this thesis, we mainly focus on microarrays for mRNA expression profiling.

In transcriptomics, the Affymatrix GeneChip® is widely used (Figure 2.3, [DWWTM06]). The chip can measure about 6.5 million featured in a single experiment. The amount of features on a chip has quickly increased

over time due to the progress in microarray production process. Agilent, Nimblegen and Illumia also provide microarray products that are widely used. The Affymatrix GeneChip® technique produce light intensities which are proportional to the transcript level.

After scanning the microarray probes, signal light density are transferred to an image. Higher intensities of spots usually refers to the higher expression level. The expression value of genes or probes can be extracted from the image. A background correction has to be used to remove background noises, and normalization removes the spatial effect on the array or variance between samples. The normalized expression profiles represent a gene expression matrix which is can be further used for statistical analysis and inference. The workflow of expression profiling for miRNA, SNP, aCGH is similar.

Widely used normalization methods are Factor Analysis for Robust Microarray Summarization (FARMS) [HCO06] and Robust Multi-array Average (RMA) [BAAS03]. Finding new methods for effective and robust normalization remains a very active area in current high-throughput data analysis.

The microarray technology produces measurements of tens of thousands of transcripts at the same time, whereas the sample size is typically in the order of 50 - 300 patients. Hence, classical statistical methods, such as ordinary least squares regression, are not applicable.

## 2.6 Methods for high dimension data classification

### 2.6.1 Pattern discovery in gene expression data

Pattern recognition is concerned with developing system that learn to solve a given problems using input data, represented as a matrix of samples times features [HTF08]. These problems include clustering that grouping feature by their similarity; classification that predict the label to a given instance. These two problems corresponding to unsupervised and supervised learning as described in Section 1.2. In this thesis, I mainly focus on classification problems.

### 2.6.2 Classification methods

For high dimension omic data classification, one typically uses supervised machine learning methods together with feature selection algorithms. This is, because omics data has typically far more features (p) than samples (n). This not only imposes high challenges for the interpretation of such data, but also for robust and stable statistical procedures, which are needed to detect those genes that are truly correlated with the clinical phenotype. Well known algorithms for data classification are $k$-NN, LDA, Logistic regression. An detail overview about these algorithms can be found in [HTF08]. In this thesis, I mainly focus on SVMs as classification methods.

The goal of predictive models is to infer a rule to predict the response $Y = \{-1, 1\}$ with given data $X$. For example, Logistic regression (LR) is

a classical probabilistic classification model that describes the possibility that $X$ belongs to a particular class: $Pr(Y = 1|X)$. Logistic regression model $Pr(Y = 1|X)$ using the logistic function:

$$Pr(Y = -1|X) = \frac{exp(\beta_0 + \beta_1 X)}{1 + exp(\beta_0 + \beta_1 X)}$$

$$Pr(Y = 1|X) = \frac{1}{1 + exp(\beta_0 + \beta_1 X)}$$

where $\beta_0 \, and \, \beta_1$ are two unknown coefficients of regression model, which can be estimating by maximizing the likelihood function:

$$L(\beta_0, \beta_1) = \prod_{i:\, y_i = 1} Pr(y_i = 1|x_i) \prod_{i:\, y_i = 0} (1 - Pr(y_i = 1|x_i)).$$

Logistic regression is a classical method for supervised learning, and particular efficient when sample size exceeds the number of variables.

### 2.6.3 Support vector machines

#### Introduction

SVMs is a series of supervised learning methods that produce a separating hyperplane for classification or regression problems. The term "SVM" refers to SVM classification in this thesis. This problem can be summarized as following: Given input data

$$\{(x_1, y_1), ..., (x_n, y_n)\} \in X \times Y \tag{2.1}$$

, where usually $\{x_i\} \in \mathbb{R}^p$ are input vectors and $Y = \{\pm 1\}$ are binary labels. This particular case is called binary pattern recognition or two

classes classification.

SVM learning is based on ideas from statistical learning theory [Vap00]. The main idea of SVMs is to construct a discriminative hyperplane by maximizing the so-called margin between the two classes (see below). If this is not possible in the original input space the so-called kernel trick can be used to implicitly map the data into a higher dimensional space. SVMs are widely used for classification problems in computational biology due to their ability to deal with high-dimensional data in an elegant and efficient manner [STV04, BHOS$^+$08].

**Hard margin SVMs**

Given a training data $D = \{(x_1, y_1), ..., (x_n, y_n)\}$ with $\{x_i\} \in \mathbb{R}^p$ and $y_i \in \{\pm 1\}$. A hyperplane is defined by

$$\{x : \ f(x) = w^T x + b = 0\}, \tag{2.2}$$

where $\{w_i : i = 1, ..., n\}$ is a unique coefficient vectors, and $b$ is bias term. A classification rule for the data$\{x_i\}$ introduced by $g(x)$ is

$$g(x) = sign(f(x)) = sign(w^T x + b). \tag{2.3}$$

where sign function is defined as

$$sign(a) = \begin{cases} 1, & if \quad a > 0 \\ -1, & otherwise. \end{cases}$$

If the training data are separable, the hyperplane of linear boundary classifies the data into one or two classes. From the geometrical point of view, $f(x)$ in Equation (2.2) corresponds to the signed distance of the given point $x$ to the separating hyperplane $f(x) = w^T x + b = 0$ (see page 418 in [HTF08]). We must have $y_i f(x_i) \geq 1, \; for \; all \; i = 1, 2, ..., n$.

There are a lots of separating hyperplanes satisfying Equation (2.2). The hyperplane with the maximum margin among these hyperplanes is selected as the optimal separating hyperplane (see Fig. 2.4). In Figure 2.4, the optimal margin between line $a$ and line $c$ equals to $2M = \frac{2}{\|w\|}$. The optimal separating hyperplane is determined as following procedure. $|f(x)|/\|w\|$ is the the geometric distance form training points $x$ to the hyperplane. Then training data must satisfy

$$\frac{y_k f(x_k)}{\|w\|} \geq \delta, \quad for \; k = 1, ..., p, \tag{2.4}$$

where $\delta$ is parameter for margin. A following constraint was introduced:

$$\delta \|w\| = 1 \tag{2.5}$$

to find the optimal separating hyperplane for Equation (2.4), we have to look for the $\|\omega\|$ with minimum that satisfies:

$$f(x) = w^T x + b = c, \quad for \; -1 < c < 1. \tag{2.6}$$

We construct a optimal separating hyperplane for Equation (2.6) by solving the following optimization problem:

$$minimize \quad \tau(w, b) = \frac{1}{2} \|w\|^2, \tag{2.7}$$

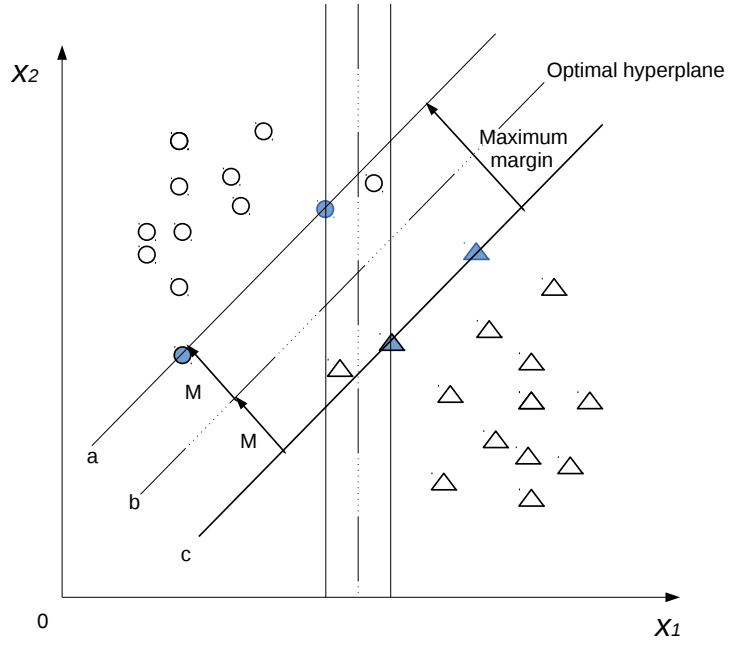



Figure 2.4: Optimal hyperplane for in a two dimensional data space. Image adopted from [Abe10].

$$subject\ to \quad y_i(w^T x + b) \geq 1, \quad for\, i = 1, ..., n. \tag{2.8}$$

The function $\tau$ in Equation (2.7) is termed as objective function and Equation (2.8) is termed as inequality constraints. This is a constrained optimization problem. The $\|w\|^2$ guarantee the optimization of Equation (2.7) to be a convex problem that can be solved by quadratic programming. Equivalently one can convert Equation (2.7) and Equation (2.8) to the so-called dual problem. This can be done by introducing Lagrange multipliers $\alpha_i > 0$:

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{n} \alpha_i \{y_i(w^T x + b) - 1\}. \tag{2.9}$$

The maximization of the Lagrangian $L$ leads to same solution than the

minimization of Equation (2.7) with respect to constraints Equation (2.8. This is true due to the convexity of the optimization problem. According to the Karush-Kuhn-Tucker (KKT) theorem the solution has to fulfill the saddle point conditions:

$$\frac{\partial L(w,b,\alpha)}{\partial b} = 0,\ and\ \frac{\partial L(w,b,\alpha)}{\partial w} = 0. \tag{2.10}$$

Furthermore at the saddle point it has to hold that:

$$\alpha_i\{y_i(w^T x + b) - 1\} \ = 0, \qquad \alpha_i \geq 0,\ for\ i = 1,\ ...,\ n. \tag{2.11}$$

In Equation (2.11), either $\alpha_i$ or $\{y_i(w^T x + b) - 1\}$ have to equal $0$. Thus, if $\alpha_i > 0$ then $y(w^T x + b) = 1$. In case that $\alpha_i$, the training points that $y_i(w^T x + b) = 1$ are called support vectors (SVs). They lie exactly on the margins (see the the filled circles on margin $a$ and the triangles on the margin $c$, 2.4). Solving Equation (2.10) leads to

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \tag{2.12}$$

and

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i. \tag{2.13}$$

By replacing Equation (2.12) and Equation (2.13) into the Lagrangian Equation (2.9), the following dual optimization problem is obtained:

$$maximize \quad W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,\,j=1}^{n} \alpha_i \alpha_j y_i y_j \left\langle x_i, x_j \right\rangle, \tag{2.14}$$

$$subject\ to \quad \sum_{i=1}^{n} \alpha_i y_i = 0\ and\ \alpha_i \geq 0,\ for\ i = 1,\ ...,\ n. \tag{2.15}$$

So the decision function in Equation (2.3) can be written as

$$g(x) = sign(\sum_{i \in S} \alpha_i y_i \langle x_i, x_j \rangle + b). \tag{2.16}$$

**Soft margin SVMs**

We described hard margin SVMs for the linear separable case, but the hard-margin SVMs is unsolvable when the training data is linear non-separable. In order to solve this problem, [CV95] extend hard margin to soft margin SVMs by introducing a set of slack variables,

$$\xi_i > 0, \ for\ i = 1, ..., p, \tag{2.17}$$

and the separation constraints in Equation (2.6) are relaxed to

$$y_i(w^T x + b) \geq 1 - \xi_i, \quad for\, i = 1, \ ..., \ p. \tag{2.18}$$

To avoid insignificant solution of all slack variables $\xi_i$, a penalty on $\xi_i$ is needed in the objective function (see Fig. 2.5). With respect to this consideration, a term $\sum_i^n \xi_i$ is introduced into Equation (2.7) for the linear non-separable case, termed as soft margin SVMs:

$$\begin{aligned} minimize \quad & \tau(w, \xi) = \tfrac{1}{2} \|w\|^2 - \tfrac{C}{q} \textstyle\sum_{i=1}^{p} \xi_i^q \\ subject\, to \quad & y_i(w^T x + b) \geq 1 - \xi_i, \ \xi_i > 0, \quad for\, i = 1, ..., p, \end{aligned} \tag{2.19}$$

where $C > 0$ is the cost parameter that balances between maximizing the margin and minimizing the classification error, and $C = \infty$ refers to the

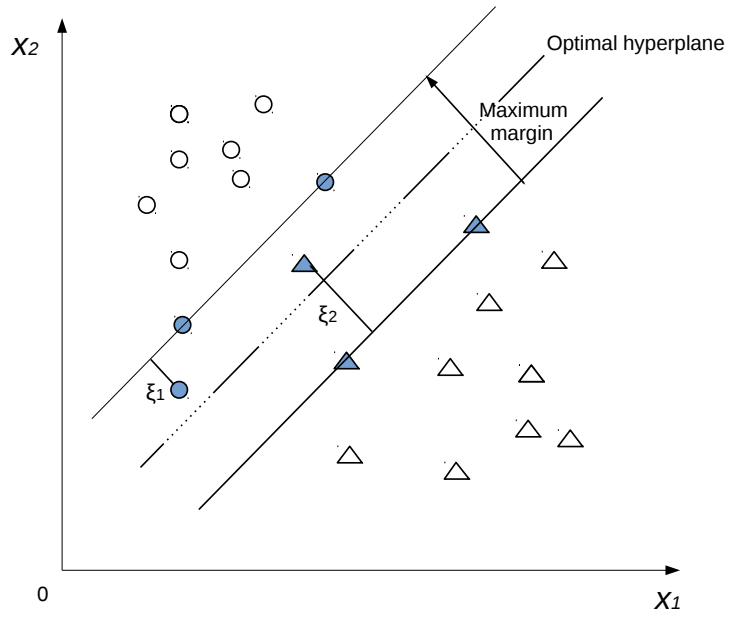



Figure 2.5: Soft margin SVM for the linear non-separable case. Image adopted from [Abe10].

linear separable case. If $\xi_i = 0$, there is no margin error for the corresponding point, and a non-zero $\xi_i$ relates to a fractional margin error. $q$ is the parameter for norm on $\xi_i$. The optimization problem in Equation (2.19) is similar to linear separable case. By defining the Lagrange multipliers $\alpha$ and $\beta$ , the Lagrange function with respect to the optimization problem in Equation (2.19) is:

$$L(w,b,\alpha,\beta) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\{y_i(w^T x + b) - 1 - \xi_i\} - \sum_{i=1}^{n}\beta_i\xi_i. \quad (2.20)$$

In order to find the optimal solution, we employ the Karush-Kuhn-Tucker (KKT) complementarity conditions to solve Equation (2.20):

$$\frac{\partial L(w,b,\xi,\alpha,\beta)}{\partial \xi} = 0,\ \frac{\partial L(w,b,\xi,\alpha,\beta)}{\partial b} = 0,\ \frac{\partial L(w,b,\xi,\alpha,\beta)}{\partial w} = 0 \quad (2.21)$$

$$\begin{aligned} \alpha_i\{y_i(w^T x + b) - 1 + \xi\} = 0, \\ \beta_i \xi_i = 0, \qquad\qquad for\ i = 1, ..., n. \\ \alpha_i \geq 0, \alpha_i \geq 0, \alpha_i \geq 0, \end{aligned} \tag{2.22}$$

Equation (2.21) can be reduced to

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i., \tag{2.23}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \tag{2.24}$$

$$\alpha_i + \beta_i = C, \quad for\ i = 1, ..., p. \tag{2.25}$$

And then, by replacing Equation (2.25 - 2.19), the Lagrangian dual problem can be written as:

$$maximize \quad W(\alpha) = \sum_{i=1}^{p} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \left\langle x_i, x_j \right\rangle \tag{2.26}$$

$$subject\ to \quad \sum_{i=1}^{p} \alpha_i y_i = 0,\ C > \alpha_i > 0, \quad for\ i = 1, ..., n. \tag{2.27}$$

Compared to hard margin SVMs, soft margin SVMs are more flexible due to the constraint C on $\alpha_i$. From Equation (2.23) to Equation (2.25), $\alpha_i$ can be categorized into three case: 1) $\alpha_i = 0$ leads to $\xi_i = 0$, and then the correspond $x_i$ is correctly classified; 2) if $C > \alpha_i > 0$, the corresponding $x_i$ is termed as in-bound support vector; 3) if $\alpha_i = C$, the corresponding $x_i$ is termed as bound support vector. In the third case, $x_i$ is correctly classified when $0 < \xi_i < 1$ and not correctly classified when $\xi_i \geq 1$.

The decision function of soft margin SVMs is defined by

$$g(x) = sign(\sum_{i \in S} \alpha_i y_i \left\langle x_i, x_j \right\rangle, + b), \tag{2.28}$$

here $S$ is a series index of the support vectors to guarantee that only support vectors are summarized. Given an new data without labels, the data is classified to

$$\begin{cases} 1, \ if \ f(x) > 0, \\ -1, \ if \ f(x) < 0. \end{cases} \tag{2.29}$$

When $f(x) = 0$, there is no unique decision possible.

**Kernel methods for non-linear SVMs**

The hard and soft margin SVMs find linear separating boundaries between the training data. In case of low dimensional data, a linear separating hyperplane may not exist. A way out is to convert the original input space to high dimensional feature space in which a linear separating SVM hyperplane can be constructed (see Figure 2.6). A kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ can be thought of as a special similarity measure between objects $x \in \mathcal{X}$ ($\mathcal{X}$ being the input space), which fulfills additional mathematical requirements, namely symmetry (i.e. $k(x, y) = k(y, x)$ for all $x, y \in \mathcal{X}$) and positive semi-definiteness (i.e. $k(x, y) = \langle \phi(x), \phi(y) \rangle$ for all $x, y$, where $\langle \cdot \rangle$ denotes the dot product in a Hilbert space $\mathcal{H}$ and $\phi : \mathcal{X} \to \mathcal{H}$ is some arbitrary function mapping objects from input space to the (possibly higher dimensional) Hilbert space $\mathcal{H}$ [SS02].

By employing a mapping function $\phi$ , the discriminant function Equation (2.2) can be written as:

$$f(x) = \langle w, \phi(x) \rangle + b \tag{2.30}$$

By using the kernel trick, the dual problem of $L_1$ soft margin SVMs in feature space is

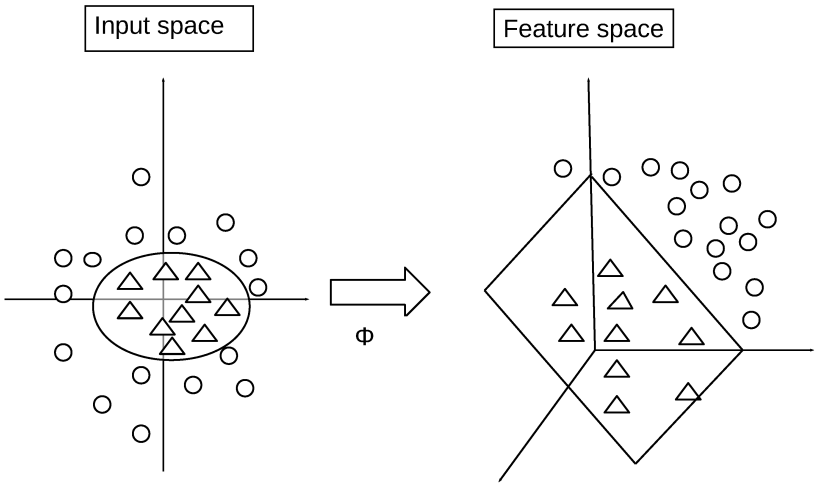



Figure 2.6: Example of kernel methods to mapped input data into feature space. Image adopted from [SS02].

$$\begin{aligned} maximize \quad & W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{pn} \alpha_i \alpha_j y_i y_j K(x_i, x_j), \\ subject\,to \quad & \sum_{i=1}^{n} \alpha_i y_i = 0,\ C > \alpha_i > 0, \quad for\, i = 1,\ ...,\ p. \end{aligned} \tag{2.31}$$

where $k(x_i, x_j) =< \phi(x_i), \phi(x_j) >$. That means kernel function k implicitly defines the map $\phi$. This is the so called kernel trick. That means $\phi$ has never to be defined explicitly as long as k is known. The following kernel functions are frequently used in SVMs:

- the linear kernel: $k(x, x') = \langle x, x' \rangle$,
- the polynomial kernel: $k(x, x') = \langle x, x' \rangle^{degree}$,
- the Radial Basis Function (RBF) kernel: $k(x, x') = exp(-\sigma \left\| x - x' \right\|^2)$.

Another popular used kernel is diffusion kernel. Diffusion kernel is also called graph kernel and defines a similarity measure between nodes in a

graph. Since the diffusion kernel is a valid kernel which corresponds at the same time to a dot product in some Hilbert space [KL02]. Suppose we are given an undirected graph $G$ with adjacency matrix $A$ and diagonal degree matrix $D$. If node $i$ connect to node $j$, $A_{i,j} = 1$, otherwise $A_{i,j} = 0$. $D_{i,i} = \sum_{j \in G} A_{i,j}$. The diffusion kernel matrix is defined as

$$K_D = exp(-\beta L), \tag{2.32}$$

where $L = D - A$ is of the graph Laplacian $L$ and $exp(-\beta * \Lambda) = Diag[e^{-\beta * \lambda_1}$, ..., $e^{-\beta * \lambda_n}]$. $\lambda_1, ...\lambda_n$ are the eigenvalues of $L$. The parameter $\beta$ control the degree of diffusion and a kernel with stronger off-diagonal effects when $\beta$ increase [KL02]. We will discuss the use of diffusion kernels in Chapter 4. Diffusion kernel can be computed as:

$$K_D = U exp(-\beta * \Lambda) U^T, \tag{2.33}$$

where $U$ is the matrix with columns being the eigenvectors of $L$. Another method for computing kernels from graph structures is $p_{step}$ random walk kernel:

$$K_{RWK} = (aI - L)^{p_{step}}, \tag{2.34}$$

where $a$ and $p_{step}$ are two positive integer parameter. Random walks tends to ramble about to their original state. In case of $a = 2$ and $p_{step} = 1$ , $K_{RWK} = 2I - L$, that converts the off-diagonal dissimilarities in $L$ to off-diagonal similarities.

### 2.6.4 Feature selection

SVMs are powerful tools for pattern classification, but have the major disadvantage that all input variables / features are used during the training

process. Especially in high dimension data, redundant and irrelevant features would inappropriately add noise to the construction of a separating hyperplane. Moreover, this would make it difficult to investigate whether specific features or feature group are related to class membership or not. Feature selection methods aim to select a specific subgroup from all features based on feature selection criteria. Moreover, the classifier using only the subset of relevant features should perform better than the one using all features.

How to choose the relevant feature sets is an important issue in statistical learning. [BL97] defined the relevance of a feature, with respect to the class label, as follows: a feature $S_i$ is relevant to label $c$ when the removal of $S_i$ will influence the classification results with respect to label $c$. Generally, feature selection methods help to improve prediction performance by dimension reduction and thus make computation faster. Usually, feature selection methods can be categorized into three classes: filter, wrapper and embedded methods [GE03, SIL07]. The work-flow of these methods are show in Figure 2.7.

***Filter methods*** use the relevance of a feature via a defined selection criterion. Then the selected features are used to train a classification algorithm (see Figure 2.7). Each feature is assigned a score by filter algorithm, such as student t-statistics or Wilcoxon sum-rank statistic, and the low-scoring features are filtered out. This procedure is fast and flexible because the feature selection procedure is independent of the classifier model. The student t-statistics, $\chi^2$ test, Markov blanket filter (MKF) [KS96], correlation-based feature selection (CFS) [Hal99, YL03]. are frequently used techniques to filter features. However, many filter methods ignore dependencies among features. Moreover, most filter feature selec-

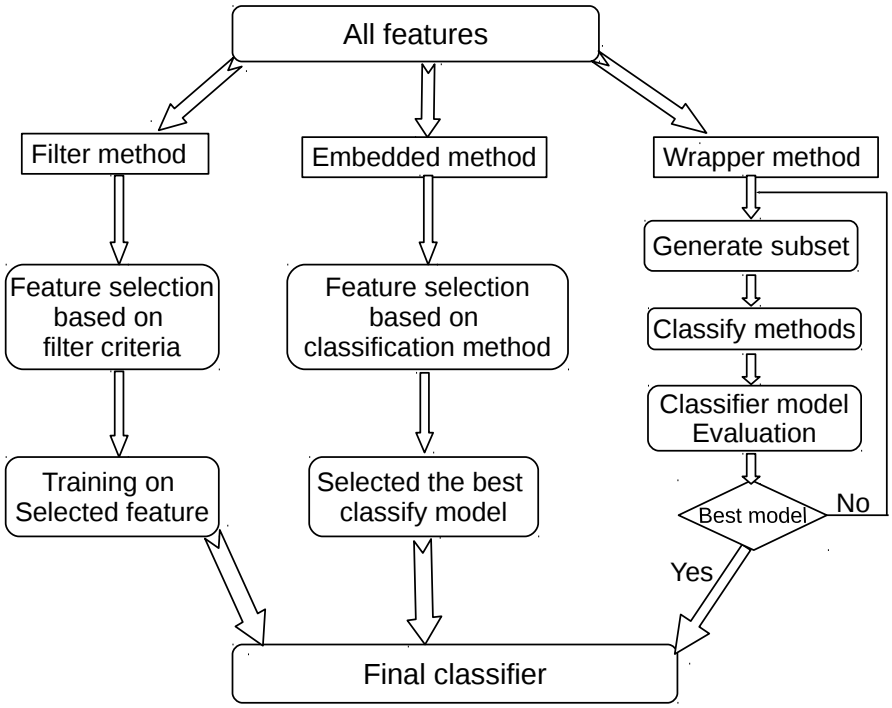



Figure 2.7: Work-flow of three feature selection methods.

tion algorithms need a threshold above which a feature is selected, which is arbitrary (see Table 1 in [SIL07]).

***Wrapper methods*** search an optimal subset of features by evaluating the prediction performance of the classifier model (see Figure 2.7). Each selected subset is thus evaluated by classifier model, and thus highly depends on classifier algorithm itself. A search algorithm is "wrapped" around the classifier algorithm during finding the best subset among all features. Heuristic methods are employed to guide the search in high dimensional feature space. A main drawback of the wrapper methods is that they are computationally intensive. An example of wrapper methods is the recursive feature elimination (RFE) algorithm for support vector machines [GWBV02b]. RFE is based on the following steps:

1. Train a SVM.
2. Rank features based on $w_i^2$ coefficient of the hyperplane.
3. Eliminate the feature with lowest ranking score from the training data
4. If more than one feature is left, then go to step 1; otherwise stop.

***Embedded methods*** search through feature space during the optimization of the classifier. Thus they usually achieve better computational performance compared to wrapper methods (see Figure 2.7). Embedded methods include random forests [DUDA06b], penalized logistic regression [MH05] and penalized SVMs [ZRHT04]. A detailed review about recent developments in penalized feature selection as embedded methods for high dimensional omics data classification is given in Ma et al. [MH08].

In some cases, different feature selection methods also can work together with aims to build a better classifier model. Apart from these approaches, ensemble feature selection methods are also popular in machine learning that use one or all three feature selection mechanisms to achieve a better model for classification [SIL07, AHVdP+10].

**Penalization methods for SVMs**

The technique described in this section is an extension of the standard SVMs by using penalty functions that allow for feature selection. Given

$f(x) = h(x)^T w + b$ with a linear separable input space, the soft margin optimization for linear SVMs can be described in the "loss + penalty" form:

$$\underset{w,b}{minimize} \quad \sum_{i=1}^{n} [1 - y_i f(x_i)]_+ + \frac{\lambda}{2} \|w\|^2 , \tag{2.35}$$

where $[1-y_i f(x_i)]_+ = max(1-y_i f(x_i), 0)$ is the so-called Hinge loss function (that means a function penalizing training errors in a defined way), and $\frac{\lambda}{2} \|w\|^2$ is the so-called penalty function. The solution of Equation (2.35) and Equation (2.19) is same when$\lambda = 1/C$. Equation (2.35) convert the SVMs to a problem of regularized function estimation, where coefficients $w$ are shrunken towards zero. The concept of penalized / regularized function estimation is very general. Apart from the $L_2$ penalty for coefficients $\omega$ described above one can consider general $L_q - norm$ penalties.

The $L_q - norm$ penalty has form:

$$L_q(w) = \left( \sum_{j=1}^{p} |w_j|^q \right)^{1/q} . \tag{2.36}$$

Several forms of such penalty are known in literature [Abe10, HTF08]:

- $L_0(w) = (\sum_{i=1}^{p} I(w_j \neq 0)$,
- $L_1(w) = \sum_{i=1}^{p} |w_j|$, (LASSO),
- $L_2(w) = \sum_{i=1}^{p} |w_j|^2$, (RIDGE).

The $L_p - norm$ family can be interpreted as a soft threshold penalty when $q \leq 1$ [BM98]. This leads to the consequence that many of the coefficients

in w become exactly 0. The corresponding input variables thus have no influence on the decision function and are practically discarded. With the $L_2$ penalty the situation is different. In this case many of the coefficients in w become small, but not exactly 0. Hence, the solution is not sparse in terms of used input variables / features. For q < 1 the optimization problem (Equation 2.35) becomes non-convex. $L_1$ penalty is continues and sparse, but has limits:

1. $L_1$ penalty selects at maximum $n$ features if $p > n$ cases;
2. In case of a group of highly correlated features the $L_1$ penalty arbitrarily picks one of them. In contrast, the $L_2$ penalty would distribute non-zero weights among them. Correlations among features are specifically observed for gene expression microarray data.

In order to overcome the limits of $L_1$ penalty, Zou and Hastie [Hui05] proposed the elastic net penalty that integrate the $L_2$ to the $L_1$ penalty to one combined penalty:

$$pen_{en} = \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2, \tag{2.37}$$

where the $\lambda_1$ and $\lambda_2$ are constant parameters that balance the cost between $L_2$ to the $L_1$ penalty. So the elastic net penalty combines sparseness properties of the $L_1$ penalty with the property of the $L_2$ penalty to distribute non-zero weights between highly correlated features. The elastic net penalty is thus expected to be more robust in cases, where one has high dimensional data with significant correlations between features

[WZZ08, LL08, BTLB11]. Apart from the $L_q$ and elastic net penalties there exist also other penalty schemes. Smooth clipped absolute deviation penalty (SCAD, [ZALP06]) is a non-convex penalty function:

$$pen_\lambda = \sum_{j=1}^{p} p_\lambda(\omega_j), \tag{2.38}$$

where

$$p_\lambda(\omega_j) = \begin{cases} \lambda|\omega_i| & if\ |\omega_i| \leq \lambda, \\ -\frac{|\omega_i|^2 - 2a\lambda|\omega_i| + \lambda^2}{2(a-1)} & if\ \lambda < |\omega_i| \leq a\lambda, \\ \frac{(a+1)\lambda^2}{2} & if\ |\omega_i| > a\lambda, \end{cases}$$

where $\omega$ are coefficients defined by hyperplanes of SVM and $a > 2$ and $\lambda > 0$ are tuning parameters.

### 2.6.5 Model assessment and selection

The generalization performance of a classifier model is defined as the model's ability to predict the class label of a new observation in an independent dataset that was not used for training the classifier. Evaluation of such performance is important to get an estimate of the quality of a model. In this section, we describe cross-validation for model assessment. Moreover, the span bound technique for computational efficient model selection for SVMs is explained.

### Cross-Validation

Cross-Validation (CV) is a widely used technique for estimating the prediction performance of a classifier model. This technique divides the given data into two parts: one part for training called training set; another part for validation the model called validation set. Generally, cross-validation has two goals:

- Model selection: several trained models with the same classifier models but different features are compared by their estimated performance in order to select the best one.
- Model assessment: after selecting a model, estimate performance of the model on unseen test data.

In this thesis, cross-validation was mainly used for the model assessment.

**K-fold Cross-Validation** process works as follows: Given a classifier model on the training set $X = \{x_i | x_i \in \mathbb{R}, i = 1, ..., n\}$ with labels $Y = \{y_i | i = 1, ..., n\}$, the loss function to measure the prediction errors is denoted by $L(Y, \hat{f}(X)$. Taken $k : \{1, ..., N\} \mapsto \{1, ..., K\}$ as an indexing function that allocates samples to one of the $k$ randomly partitions, then the cross-validation technique estimates the prediction error as:

$$CV(\hat{f}) = \frac{1}{N} \sum_{i=1}^{n} L(Y, \hat{f}^{-k(i)}(X), \tag{2.39}$$

where $\hat{f}^{-i}(X)$ is a classification function fitted on data from which fold k(i) was eliminated. 5-fold or 10-fold cross-validation are frequently used

in practice (see Figure 7.9 of [HTF08]). If $K = n$, the cross-validation is called leave-one-out (LOO) cross-validation. The LOO-CV usually has a low bias accompanied with high variance as only one observation is take as validation data at each step. Moreover, LOO-CV is computationally intensive compared to 5-fold or 10-fold cross-validation (see Chapter 7.10.1 in [HTF08]).

Generally, the K-fold cross-validation process should be repeated 5 or more times in order to estimate the variance resulting from a random split of the whole dataset into k distinct folds. In this thesis we take 10-fold cross-validation with 10 repeats for each algorithm.

**Prediction error measurement**

Several methods can be used to measure the prediction error of classification and regression models. Here use $\hat{y}_i$ as the predicted class label for the individual $i$ with the true value $y_i$. As described before, a classifier usually outputs a label $+1$ or $-1$. Given two classes, a classifier can create the following assignments:

- True Positive (TP): algorithm predicts a positive instance as positive.
- False Negative (FN): algorithm predicts a positive instance as negative.
- True Negative (TN): algorithm predicts a negative instance as negative.

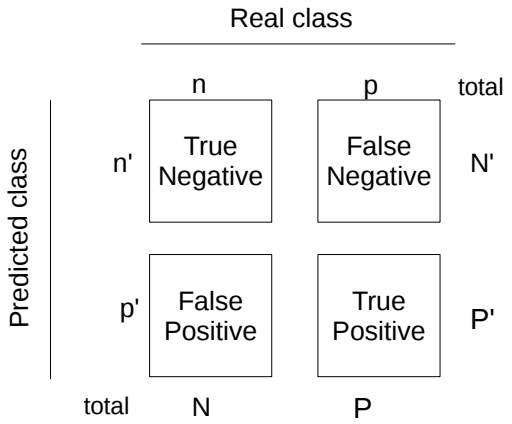



Figure 2.8: A 2 by 2 confusion table.

- False Positive (FP): algorithm predicts a negative instance as positive.

A contingency table shows these class assignments (Figure 2.8). Using this information, a variety of quality measures are used to compute the prediction performance of a classifier algorithm:

- **Accuracy** (ACC) is the ratio of the number of correctly prediction among all predictions:

$$ACC = \frac{\mathrm{TP} + \mathrm{TN}}{(\mathrm{FP} + \mathrm{TN}) + (\mathrm{TP} + \mathrm{FN})}.$$

- **Sensitivity** or **true positive rate** (**TPR**) is the probability / ratio

of the positive sample that are correctly predicted:

$$TPR = \frac{\text{TP}}{(\text{TP} + \text{FN})}.$$

- **Specificity** or **true negative rate** (**TNR**) is the probability / ratio of the negative sample that are incorrectly predicted:

$$TNR = \frac{\text{TN}}{(\text{FP} + \text{TN})} = 1 - \text{FPR}.$$

- **False positive rate** (**FPR**) is the ratio of the negative sample that are incorrectly predicted:

$$FPR = \frac{\text{FP}}{(\text{FP} + \text{TN})}.$$

- **False negative rate** (**FNR**) is the ratio of the positive sample that are correctly predicted:

$$FNR = \frac{\text{FN}}{(\text{FN} + \text{TP})}.$$

- **AUC/AUCROC**: Area Under the **ROC** (Receiver Operating Characteristic) Curve.

An area under the **ROC** (Receiver Operating Characteristic) plot is depicts FPR versus TPR and thus shows the relative balance between true positives and false positives [Bra97]. In the ROC plot, each point corresponds to a defined threshold of a real valued decision function, giving rise to a specific fraction of false positives and false negatives. The area under the ROC curve (AUC) is a common way to summarize whole ROC curves into one number. As the AUC is based on a unit square of the

ROC space, its value is always between 0 and 1, and a bigger AUC value indicates better prediction performance. If a model's $AUC < 0.5$, it is worse than random. In this thesis, R package ROCR ([SSBL05]) is used for calculating AUC values of classification models.

**Model selection via span bound**

As introduced in the previous section, cross-validation is a re-sampling technique to estimate the generalization performance of a classifier. In order to get a well optimized model, most learning algorithms need to tune more than one parameter. For example, a tuning parameter for SVMs is the constant $C$ in Equation (2.19) for penalizing margin and training errors. Hence, the best among a number of candidate models (each defined via a specific value of parameter C) needs to be found. Model selection can then be performed by cross-validating each of these candidate models. However, this nested cross-validation procedure would be a time-consuming method. The span bound technique has been proposed to address this problem. The span bound defines an upper bound for the leave-one-out cross-validation error of a SVM classifier [VC00, CVBM02]. Here I focus on the span bound technique in the hard margin case.

Given any fixed support vector $x_p$ and $\alpha^0 = (\alpha_1^0, \cdots, \alpha_n^0)$ is the vector of Lagrange multipliers for the optimal hyperplane, a set $\Lambda_p$ is defined as a constrained linear combination of the support vectors $\{x_i\}_{i \neq p}$ :

$$\Lambda_p = \left\{ \sum_{i=1, i \neq p}^{n} \lambda_i x_i : \quad \sum_{i=1, i \neq p}^{n} \lambda_i = 1, \; and \; \alpha_i^0 + y_i y_p \alpha_i^0 \lambda_i \geq 0 \right\}, \tag{2.40}$$

where$\lambda_i$ is constrained parameter and can be negative. The span of the support vectors $x_p$ is defined based on the the distance between $x_p$ and

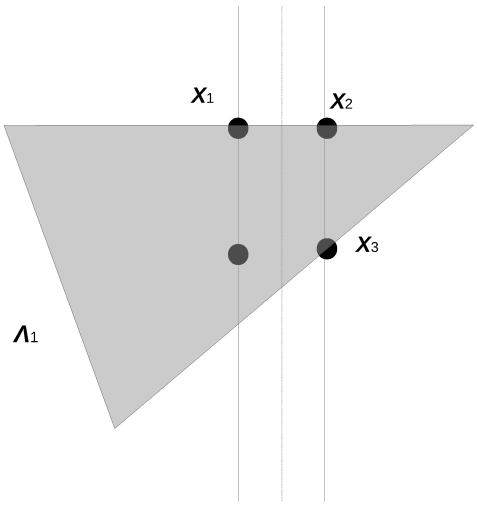



Figure 2.9: Example of the span set $\Lambda_1$ of the support vectors $x_1$. The two real lines are boundary of SVM. As the support vector $x_1$ belong to the span set $\Lambda_1$, the distance form $x_1$ to $\Lambda_1$ is equal to zero. The set $\Lambda_1$ is computed by $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$. Image adopted from [VC00].

the set $\Lambda_p$:

$$S_p^2 = d^2(x_p, \Lambda_p) = \underset{x \in \Lambda_p}{min}(x_p - x)^2. \tag{2.41}$$

As shown in Figure 2.9, $S_p = d(x_p, \Lambda_p) = 0$ when $x_p \in \Lambda_p$ .

The smaller $S_p = d(x_p, \Lambda_p)$, the smaller the LOO cross-validation error on the support vectors $x_p$. The span rule estimates the number of errors via LOO cross-validation via:

$$T = \frac{1}{n} \sum_{p=1}^{n} \Psi(\hat{\alpha}_p S_p^2 - y_p f(x_p)), \tag{2.42}$$

where the value of the span can be computed in closed form as $S_p^2 = \frac{1}{(K_S^{-1}V)_{pp}}$. Here $K_{SV}$ denotes the kernel matrix restricted to the support

vectors. $\Psi$ is the step function:

$$\Psi(x) = \begin{cases} 1, & if\ x > 0 \\ 0, & otherwise \end{cases} .$$

The span rule provides an upper bound of the leave-one-out error. The practical advantage stems from the fact that it can be computed very efficiently, provided that the number of samples is small (which is the typical case for omics data). We use the span bound for choosing multiple parameters for SVM in this thesis.

### 2.6.6 Limitations of purely data driven classification methods

A common approach to obtain a signature for diagnostic or prognostic purposes is to put patients into distinct groups and then construct a classifier that can discriminative patients in the training set and is able to predict well unseen patients. In the past a large number of classification algorithms have been developed or adopted from the machine learning field, like PAM, SVM-RFE, SAM, Lasso and Random Forests [Tib96, Bre01, THNC02, GWBV02b]. Several adaptations of Support Vector Machines(SVM) [Vap00] have been suggested for gene selection in genomic data, like L1-SVMs, SCAD-SVMs and elastic net SVMs [FM04, ZALP06, WZZ08]. Although these methods show reasonably good prediction accuracy, they are often criticized for their lack of gene selection stability and the difficulty to interpret obtained signatures in a biological way [EDKG$^+$05, DD11]. These challenges provide opportunities for the development of new gene selection methods.

To overcome the disadvantages of conventional approaches Chuang et al. [CLL+07] proposed an algorithm that incorporates of protein-protein interaction information into prognostic biomarker discovery. Since then a number of methods going into the same direction have been published [CLL+07, RZD+07, LCK+08, BS09, TLWF+09, ZSP09, JBF+10]. In the next section, I give a brief overview on current network based approaches for biomarker discovery.

## 2.7 Network centric approaches

### 2.7.1 Overview

Nowadays knowledge on protein-protein interactions (PPIs) as described in Section 1.3. Various network based approaches have been proposed to integrate prior knowledge on canonical pathways, Gene Ontology (GO) annotation or protein-protein interactions into feature selection algorithms [GZL+05, CLL+07, RZD+07, LCK+08, TLWF+09, BS09, ZSP09, JBF+10]. A recent review on such approaches can be found in [CF12a]. The general hope of these approaches is that biological knowledge can lead to better interpretable and more stable signatures. Whether network based classification methods automatically also lead to higher prediction accuracies is still a matter of debate [CF12c, SCK+12].

In general one may divide existing methods integrating network knowledge broadly into two main classes:

On one hand there are network centric approaches, which map gene expression data onto a molecular network reconstructed from the literature

and then either try to identify discriminative / differential sub-networks between patient groups, or directly compute summary statistics (pathway activity) for pre-defined sub-networks (e.g. canonical pathways). Afterwards often a conventional classifier (e.g. logistic regression, k-NN) or Cox regressor is applied to make predictions based on the expression profiles of sub-network genes.

On the other hand data centric approaches are closer to traditional machine learning methods. Here the idea is to bias the gene selection process within a machine learning framework in such a way that connected genes are preferably selected. There are two main techniques for this purpose: One is to construct a mathematical embedding of gene expression data into a network graph space via the so-called kernel trick [SS02]. Afterwards existing kernel-based feature selection algorithms, such as SVM-RFE [GWBV02a], can be applied. Another approach is to modify the feature selection process itself, e.g. by imposing specific restrictions on the learnable parameters (so-called regularization) [TA77].

In the following, I give a more detailed overview about these methods.

### 2.7.2 Network features

An approach, which is possibly most focusing on the network structure itself, is to purely select genes based on topological features of the PPI network. An example is the method proposed in [TLWF+09]. Here the idea is to concentrate on hubs in the network, i.e. proteins with an extraordinary high degree of interactions. In their paper Taylor et al. show that the average Pearson correlation of the expression of a hub protein

and its interacting partners can be used to reliably predict survival of breast cancer patients without any further machine learning based variable or feature selection procedure.

### 2.7.3 Pathway activity

Another method to integrate network knowledge is to summarize the expression level of predefined canonical pathways obtained from databases, such as KEGG [KAG+08], into one value, for instance by taking the mean or the median. These newly constructed interpretable features are then correlated with the clinical phenotype to be predicted using conventional machine learning techniques.

Guo *et al.* [GZL+05] report that "functional expression profiles" obtained by taking the average expression of genes annotated to significantly enriched Gene Ontology (GO, [The04]) categories could increase the robustness of a classifier trained to discriminate four cancer types.

Rather than simply looking at mean or median expression [VBS+10] propose a probabilistic approach based on a factor graph model for pathway activity inference from both, gene expression and copy number alterations. In contrast to many others, this method is completely probabilistic and takes the topology of the pathway into account.

Teschendorff *et al.* [TGA+10] further decompose pathways into coherent modules based on the correlation structure in gene expression data. For each module an activation metric is proposed, which specifically takes into account the network architecture.

Another approach following the same direction is proposed by Trey Ideker and co-workers [LCK+08]. In their paper an activity score is derived from the normalized expression of most discriminative genes within each pathway. Logistic regression is applied to discriminate between "good" or "bad" prognosis breast cancer patients based on these scores. In their paper Lee et al. show that their "combined optimal response genes" (CORGs) approach yields better prediction performance than if pathway activity is simply estimated via the mean or median expression level. A further improvement of the method with respect to the selection of discriminative genes within each pathway is proposed in [YDPD12].

Bild *et al.* [BYC+06] estimate pathway activity by so-called "meta-genes", which are obtained by computing the first principal components of the expression of pathway genes. The authors use their method to cluster several tumor entities and identify coordinated patterns of pathway deregulation, which distinguish between specific cancers and tumor subtypes. Bild et al. show that estimated pathway activities are predictive for the respective patient subgroups, and that in cell lines pathway activity also predicts the sensitivity to therapeutic compounds. An extension of the pathway activity classifier to identify oncogene-inducible modules is described in [BWS+08].

Yu *et al.* [YSZ+07] propose to first detect pathways that are significantly associated with the phenotype via a global test strategy [GVDV04]. Afterwards genes annotated to these pathways are selected based on their individual differential expression. Using their approach the authors successfully establish an interpretable signature for predicting metastasis of lymph node negative breast cancer patients.

The paper by [KLH+11] focuses on functional gene groups defined by GO.

Rather than computing an explicit measure of group activity, the authors first identify group representatives via PAM clustering [KR90].

### 2.7.4 Differential sub-networks

Rather than looking at predefined canonical pathways or GO groups another idea, which puts a little bit more emphasis on measured data, is to reconstruct a protein-protein interaction network for all gene products and then use experimental data to identify differentially expressed sub-networks. One of the first approaches in this direction is described in [CLL+07]. The algorithm starts from "seed" proteins in the network, which are highly differentially expressed. Then around each seed protein neighbors are added in a greedy hill climbing fashion until the discriminative power of the corresponding sub-network (measured via the mutual information of the average normalized gene expression together with the clinical outcome variable) reaches a local maximum. In their paper Trey Ideker and co-workers show that their method not only leads to clearly interpretable signatures for discriminating "poor" and "bad" prognosis breast cancer patients, but also improves prediction performance compared to a conventional machine learning setup. Similar greedy algorithms for identification of differential sub-networks have been proposed by other authors, e.g. [CK10, FKJ10, SYD10, AYP+11].

A particular interesting variant has recently been introduced by [DI11]. They modify Random Forests [Bre01], which contain a large ensemble of decision trees, such that individual trees only use neighboring genes in the PPI network. This allows them to draw conclusions about the inherent logic by which stably selected sub-networks are dis-regulated. The

authors show that their method leads to a much better reproducibility of selected markers compared to using a conventional Random Forest.

It has to be mentioned that despite their good performances all so far mentioned approaches are heuristic and thus cannot guarantee to find the *optimal* differential sub-network. Attempts to obtain an optimal sub-network are described in [CNCK11] via branch and bound and in [DCS+10] via exhaustive search. A elegant solution is proposed by [DKR+08]. After calculating a score for differential expression of each node in the protein-protein interaction network, the authors interpret the problem of identifying the optimal differential sub-network as an instance of the prize-collecting Steiner tree problem, which they solve to optimality via integer linear programming (ILP). The authors show that their obtained optimal sub-networks generally correlate well with the clinical phenotype of diffuse large B-cell lymphomas, however no rigorous validation in terms of prediction accuracy is performed.

In general, identification of an optimally discriminative sub-network is an NP-hard problem [DKR+08, DWC+11] and thus algorithms have to face a super-polynomial run time complexity, which can make them intractable for larger datasets. An interesting compromise between computational speed and the goal to obtain a well separating sub-network has thus recently been proposed in [DWC+11]. Their algorithm is based on the color coding paradigm [ADH+08], which allows for identifying optimally discriminative sub-networks up to a certain error rate. Dao et al. use a randomized approximation algorithm to obtain polynomial run time complexity. Afterwards the authors employ a 3-NN classifier on averaged expression levels of each sub-network to discriminate response to chemotherapy in breast cancer.

### 2.7.5 Data centric approaches

#### Mathematical embedding

All previously mentioned approaches deal with a PPI network as the central entity. In contrast, data centric approaches focus on the experimental data. Kernel techniques [SS02] allow for a mathematically elegant way of combining network information with experimental data.

Among many other applications kernel functions have been proposed for nodes in a graph or network based on the notion of random walks. A random walk is a stochastic process that consists of a sequence of moves that are taken along the graph structure according to some defined probability distribution. The diffusion kernels [KL02] is a specific similarity measure for nodes in a graph that considers all random walk paths connecting nodes $x$ and $y$, but weights each path in dependency on the path length (see Chapter 4). This is done in an exponentially decreasing way. Diffusion kernels are mathematically equivalent to the fundamental solution of the heat equation in physics, which describes the evolution of heat in a region under certain boundary conditions. If instead of exponentially decreasing weights for path lengths a linear weighting scheme is preferred, one arrives at the pseudo-inverse of the graph Laplacian [GDCW09]. In the same paper also a random walk kernel is proposed, which simply bounds the number of random walk steps to $p$ (see in Chapter 4).

The afore mentioned graph kernels allow for easily incorporating measurement data, such as gene expression. This is done by weighting each edge $x \rightarrow y$ in the network by the similarity of the gene expression of

$x$ and $y$ (using the dot product). This is equivalent to defining a kernel function between $x$ and $y$ as:

$$k(x, y) = \mathbf{x}^T \mathbf{Q} \mathbf{y}$$

where x and y are the vectors of gene expression values for genes $x$ and $y$, and Q is the graph kernel matrix between nodes in the network. Consequently the expression data is linearly mapped via the graph kernel matrix Q to some different space.

Combining gene expression data with network information in such a way has been described by [RZD+07] and [GDCW09]. In general the intuition of these methods is that genes which are closely connected in the network should also have similar expression levels. Rapaport et al. (2007) in particular emphasize the possibility to conduct unsupervised clustering analysis of gene expression data in this way besides more common supervised classification, which yields to biologically interpretable results. Several other authors have used graph kernels to identify possibly disease causing genes [NTT+09, QZZC10].

Recently, [CXR+11] have introduced a variation of the kernel idea using the pseudo-inverse of the graph Laplacian. In their paper the authors compute an explicit mapping of gene expression data by a matrix square root of Q, which is calculated via singular value decomposition. An ordinary linear Support Vector Machine is then trained on the transformed data. Afterwards the solution is back-transformed to the original space and a permutation test executed for assessing the significance of genes and identifying sub-networks. With their approach the authors successfully predict early vs. late recurrence of ER positive breast cancer patients with comparably high accuracy. Moreover, the obtained sub-

network markers appear to be biologically plausible.

### Biased feature selection

Instead of augmenting the similarity measure of each pair of genes with network information via embedding techniques, another approach is to directly integrate network information into conventional variable/feature selection techniques. [ZSP09] describe a modified Support Vector Machine (SVM) algorithm with embedded feature selection, which strongly prefers to select genes, which are connected to each other. Via their method the authors successfully obtain sub-networks associated to Parkinson's disease and to breast cancer metastasis.

Johannes *et al.* [JBF+10] introduce a modification of the frequently used SVM-RFE algorithm, called SVM-RRFE (Reweighted Recursive Feature Elimination). They use the GeneRank approach [MBHG05], which is based on Google's famous PageRank algorithm [PBMW99] to identify genes that on one hand exhibit a high fold change and on the other hand are central in the PPI network. With this ranking they re-adjust the SVM decision hyperplane, which is learned at each step of the SVM-RFE algorithm. This way they give preference to selecting genes, which have a high GeneRank. It can be shown that this approach is equivalent to run the conventional SVM-RFE algorithm on data that is transformed in a specific way, i.e. embedded into a different space. In their paper the authors demonstrate that SVM-RRFE is not only superior to the conventional SVM-RFE algorithm in predicting an early relapse in breast cancer patients, but can also compete with several other network based

gene selection approaches. Moreover, the stability and interpretability of the obtained gene signatures are significantly improved.

Binder *et al.* [BS09] propose a component-wise likelihood boosting approach (pathBoost) for integrating network information. The idea is to decrease the penalty for selecting variables / genes that are connected in the PPI network. The authors demonstrate on two gene expression datasets, diffuse large B-cell lymphoma and ovarian cancer, that their approach is able to improve survival time predictions via a multivariate penalized Cox regression model compared to conventional likelihood boosting for the same purpose.

In a recent paper [GPF+11] extend the method by Binder and Schumacher by considering a miRNA-mRNA interaction graph rather than a PPI network. Gade et al. show that this way miRNA and mRNA expression data can be combined in a straight forward way for predicting the risk of a relapse in prostate cancer via penalized Cox regression. Moreover, they demonstrate that their approach enhances prediction performance and gene selection stability compared to several other methods.

Lasso regression models [Tib96] have gained a particular attention for high dimensional data analysis during the last years. [LL08] propose a modification of this approach, which down-weights the penalty for selecting genes that are in proximity to each other. They demonstrate that their method can improve over the conventional lasso for predicting survival of glioblastoma patients. Despite the elegance of the approach it has to be mentioned that the authors do not consider the possible censoring of patient survival times in their study. Hence, the application of a conventional regression framework in this context has to be seen critical.

## 2.8 Summary

Integration of biological knowledge, specifically from protein-protein interaction networks and canonical pathways, is widely accepted as an important step to make biomarker signature discovery from high dimensional data more robust, stable and interpretable. Consequently there is an increasing amount of methodologies for this purpose. In this chapter, I gave a general overview about these approaches and grouped them into categories.

In conclusion we see that all approaches that have been proposed so far have specific advantages and disadvantages. Thus there is a strong need for systematic empirical comparisons. In Chapter 3, I conducted a comparison of 14 classification algorithms (8 using network knowledge) for predicting early vs. late relapse of breast cancer patients in 6 microarray datasets. In this context it has to be emphasized that most published methods have been evaluated for one specific clinical questions (e.g. early relapse prediction) in one disease (mostly breast cancer), only. To get a more complete picture, more comprehensive studies including more clinical questions and more disease entities are needed in order to guide practitioners, under which conditions which method would be a good choice. In Chapter 4, a new algorithm is developed to not not only integrate more molecular interaction information, but also more molecular data types. Nonetheless, there will be always a dataset specific dependency of an algorithm's performance, which can never we resolved. Careful checking of assumptions is therefore a prerequisite for the successful application of any algorithm.

# Chapter 3

# Comparison of Current Feature Selection Methods in Terms of Accuracy, Stability and Interpretability

"It is by logic that we prove, but by intuition that we discover. To know how to to criticize is good, to know how to create is better."

– *Jules Henri Poincaré.*

In this chapter, we compare fourteen published gene selection methods (eight using network knowledge) on six public breast cancer datasets with respect to prediction accuracy, biomarker signature stability and biological interpretability in terms of an enrichment of disease related genes, KEGG pathways and known drug targets.

The comparison done here is thus multi-dimensional and goes beyond the typical studies focusing purely on prediction accuracy. The reason is that - as pointed out in section 2.6 - the limitation of current biomarker

signatures is their low reproducibility coupled with the difficulty to interpret them in the context of existing biological knowledge. Hence, the questions, which we address in this chapter, are:

1. Do network based gene selection methods yield a higher prediction accuracy than purely data based ones?
2. Does biological knowledge help to obtain better reproducible and interpretable gene signatures?
3. Which of the tested network based algorithms is most successful with respect to prediction accuracy, reproducibility and interpretability of signatures?

The content of this chapter is based on a previous publication in ***BMC Bioinformatics***[CF12c].

## 3.1 Materials and methods

### 3.1.1 Gene selection methods

We employed fourteen published gene selection methods in this chapter. As already described in section 2.6.4, feature selection methods can be classified into three categories [GE03]: filters, wrappers and embedded methods. Filter methods select a subset of features prior to classifier training according to some measure of relevance for class membership,

e.g. mutual information [Bat94]. Wrapper methods systematically assess the prediction performance of feature subsets, e.g. recursive feature elimination (RFE, [GWBV02b]); and embedded methods perform features selection within the process of classifier training. The methods we employ in this chapter covered all three categories. Furthermore we can classify feature selection methods according to whether or not they incorporate biological network knowledge (conventional vs. network-based approaches).

As one of the most basic approaches, we considered here a combination of significance analysis of microarrays (SAM) [TTC01] as a filter prior to SVM or Naïve Bayes classifier learning [Ris01]. More specifically, only genes with $FDR < 5\%$ (Benjamini-Hochberg method) [BH95] were considered as differentially expressed. As further classical gene selection methods we considered prediction analysis for microarrays (PAM) [THNC02], which is an embedded method, and recursive feature elimination (SVM-RFE) [GWBV02b], an SVM-based wrapper algorithm. Moreover, we included SCAD-SVMs [ZLS$^+$06] and elastic-net penalty SVMs (HHSVM) [WZZ08] as more recently proposed embedded approaches (see section 2.6.3) that particularly take into account correlations in gene expression data. In this chapter, we used SAM+SVM (significant gene SVM), SAM+NB (significant gene Naïve Bayes classifier), PAM, SCAD-SVM, HHSVM and SVM-RFE as "conventional" feature selection methods that do not employ network knowledge.

The following network-based approaches for integrating network or pathway knowledge into gene selection algorithms were investigated: Mean expression profile of member genes within KEGG pathways (aveExpPath) [GZL$^+$05], graph diffusion kernels for SVMs (graphK; diffusion ker-

nel parameter $\delta = 1$) [RZD+07], p-step random walk kernels for SVMs (graphKp; parameters $p = 3, a = 2$, as suggested by Gao et al. [GDCW09]), pathway activity classification (PAC) by CROGs gene sets of each pathways [LCK+08], gradient boosting (PathBoost, [BS09]) and network-based SVMs (parameter $sd.cutoff = 0.8$ for pre-filtering of probesets according to their standard deviation) [ZSP09].

In case of avgExpPath whole KEGG-pathways were selected or not selected based on their average differential expression between patient groups. This was done based on a SAM-test with FDR cutoff 5% (see above). In case of diffusion and p-step random walk kernels the SVM-RFE algorithm was adopted for gene selection using the implementation in the R-package pathClass [JFSB11]. Furthermore, pathClass was used to calculate the diffusion kernel. This implementation is directly based on [RZD+07] and only keeps the 20% smallest eigenvalues and corresponding eigenvectors of the normalized graph Laplacian to compute the kernel matrix.

PAC and PathBoost come with an own mechanism to select relevant genes. PathBoost incorporates network knowledge directly into the gradient boosting procedure to perform gene selection, whereas PAC first selects genes within each KEGG-pathway based on a t-test and then summarizes gene expression in each pathway to a pathway activity score. According to the original paper by Lee et al. [LCK+08] only the top $10\%$ pathways with highest differences in their activity between sample groups were selected.

Recently, Taylor et al. [TLWF+09] found that differentially expressed hub proteins in a protein-protein interaction network could be related to breast cancer disease outcome. We here applied their approach (called

HubClassify) as follows: the random permutation test proposed in Taylor et al. [TLWF$^+$09] was used to select differentially expressed hub genes with FDR cutoff $0.5\%$. Hubs were here defined to be those genes, whose node degree fell into the top 1% percentile of the degree distribution of our protein interaction network.

Afterwards a SVM was trained using only those differential hub genes. Finally, we also include Reweighted Recursive Feature Elimination (RRFE) algorithm [JBF$^+$10], which combines GeneRank [MBHG05] and SVM-RFE as implemented in the pathClass package [JFSB11]. In summary average pathway expression (aveExpPath), graph diffusion kernels for SVMs (graphK), p-step random walk graph kernels for SVMs (graphKp), PAC, PathBoost, networkSVM and HubClassify are considered in our comparison of network-based gene selection methods.

For all linear SVM classifiers used in this study the soft-margin parameter $C$ was tuned in the range $10^{-4}$, $10^{-3}$, $\ldots$, $10^{4}$ on the training data. For that purpose the pathClass package was employed, which uses the span-bound for linear SVMs as a computationally attractive and probably accurate alternative to cross-validation (see section 2.6.5, [CVBM02]). For elastic net SVMs and SCAD-SVMs we used the R-package penalizedSVM [BWT$^+$09], which allows for tuning of hyperparameters (elastic net: $\lambda_1 \in [2^{-8}, 2^{14}]$, $\lambda_2$ set in a fixed ratio to $\lambda_1$ according to [WZZ08]; SCAD-SVM: $\lambda \in [2^{-8}, 2^{14}]$) based on the generalized approximate cross-validation (GACV) error as another computationally attractive alternative to cross-validation. The EPSGO algorithm described in [FZ05] was used for finding optimal hyper-parameter values within the defined ranges. Note that in any case only the training data were used for hyper-parameter tuning.

It should be mentioned that for conventional approaches all probesets on the chip were considered. This is in agreement with a typical purely data driven approach with no extra side information. Please note that an a-priori restriction to probesets, which can be mapped to a pre-defined network, would already include a certain level of extra background knowledge with corresponding assumptions.

### 3.1.2 Classification performance and stability

In order to assess the prediction performance of all tested methods we performed a 10 times repeated 10-fold cross-validation on each dataset. That means the whole data was randomly split into 10 fold, and each fold sequentially left out once for testing, while the rest of the data was used for training and optimizing the classifier (including selection of relevant genes, hyper-parameter tuning, standardization of expression values for each gene to mean 0 and standard deviation 1, etc.). The whole process was repeated 10 times. It should be noted extra that also standardization of gene expression data was only done on each training set separately and the corresponding scaling parameters then applied to the test data.

The area under receiver operator characteristic curve (AUC) was used to measure the prediction accuracy via the R-package ROCR [SSBL05]. To assess the stability of gene selection, we computed the selection frequency of each gene within the 10 times repeated 10-fold cross-validation procedure. That means a particular gene could be selected at most 100 times In order to summarize the selection frequencies for all genes we defined a so-called stability index (SI) as

$$SI = \frac{1}{|P|} \sum_{s \in P} h(s) \tag{3.1}$$

where $P$ is the set of selected genes that had been selected at least once and $h(s)$ is the actual number of times that $s$ was selected. SI represents a weighted histogram count of selection frequencies. Obviously, the larger SI the more stable the algorithm is. In the optimal case $SI = 100$. The $SI$ has to be seen together with the size of gene signature, because trivially a classifier selecting all genes would always achieve $SI = 100$.

### 3.1.3 Functional analysis of signature genes

To interpret a signature gene in terms of biological function, we performed an enrichment analysis in terms of cancer-related disease genes, KEGG pathways and known drug targets for the prognosis biomarkers via Fisher's exact test. We employed FunDO [OFH+09] to look for enrichment of disease related genes. FunDO uses a hyper-geometric test to find relevant diseases. Multiple testing correction was done using Bonferroni's method [BA95]. Furthermore, an analysis of enriched KEGG pathways based on a hypergeometric test was done (multiple testing correction via Benjamini-Yekutieli's method [Ben01]). We also carried out an enrichment analysis for known targets of therapeutic compounds against breast cancer. For that purpose, we retrieved a list of 104 proteins and respective therapeutic compounds in breast cancer, which are either in clinical trials (also withdrawn ones), FDA approved or on the market with the help of the software MetaCore™. Fisher's exact test was then used to assess statistical overrepresentation of drug targets within each signature.

### 3.1.4 Datasets

#### Microarray gene expression data

We collected six public breast cancer Affymetrix HGU133A microarray

(22,283 probesets) datasets [WKZ+05, PBA+05, SBvT+08, SWL+06, IGS+06, DPL+07], which are further described in Table 3.1. The six datasets were obtained via Gene Expression Omnibus [BTW+11], and normalization was carried out using FARMS [HCO06]. As clinical end points we considered metastasis free (datasets by Schmidt et al [SBvT+08]., Ivshina et al [IGS+06].) and relapse free (other datasets) survival time after initial clinical treatment, depending on the availability of the corresponding information in the original data. Time information was dichotomized into two classes according whether or not patients suffered from a reported relapse / metastasis event within 5 years. Patients with a survival time shorter than 5 years without any reported event were not considered and removed from our datasets.

**Protein-Protein interaction (PPI) network**

A comprehensive protein interaction network was compiled from the Pathway Commons database [CGD+11], which was downloaded in tab-delimited format (September 2012). All SIF interactions INTERACTS_WITH and STATE_CHANGE were taken into account[1] and self loops removed, resulting in a large network with 11,361 nodes and 610,185 edges.

The R package, *hgu133a.db* [CFPL09], was employed to map probe sets on the microarray to nodes in the PPI-network. Accordingly, expression values for probesets on the microarray that mapped to the same gene in the network were averaged. In order to consider genes with available probesets on the array but no corresponding network information

[1] http://www.pathwaycommons.org/pc/sif_interaction_rules.do

Table 3.1: Employed breast cancer data sets. The data in () means patients with dmfs/rfs $\geq$ 5 years.

| GEOid | patients | dmfs/rfs $<$ 5 years |
|---|---|---|
| GSE2034 [WKZ+05] | 286 | 93 (183) |
| GSE1456 [PBA+05] | 159 | 34 (119) |
| GSE2990 [SWL+06] | 187 | 42 (116) |
| GSE4922 [IGS+06] | 249 | 69 (159) |
| GSE7390 [DPL+07] | 198 | 56 (135) |
| GSE11121 [SBvT+08] | 200 | 28 (154) |

we added for all these genes unconnected nodes to our initial network, resulting in 9,186 nodes for all breast cancer datasets.

## 3.2 Results and discussion

### 3.2.1 Predictive power and stability

We assessed the prediction performance of prognostic biomarker gene signatures obtained by fourteen gene selection methods on six gene expression datasets in terms of area under ROC curve (AUC) (see Figure 3.1). The gene selection stability of each gene selection method is depicted in Figure 3.3 (fraction of constantly selected probe sets). Here are the abbreviations used for the 14 tested methods: PAM (prediction analysis of microarray data), sigGenNB (SAM + Naïve Bayes), sigGenSVM (SAM + SVM), SCAD-SVM, HHSVM (Huberized Hinge loss SVM), RFE (Recursive Feature Elimination), RRFE (Reweighted Recursive Feature

Elimination), graphK (graph diffusion kernels for SVMs), graphKp (p-step random walk graph kernel for SVMs), networkSVM (Network-based SVM), PAC (Pathway Activity Classification), aveExpPath (average pathway expression ), HubClassify (classification by significant hub genes), pathBoost. In the following graphs of this Chapter, A represents data GSE2034 [WKZ+05]; B represents data GSE11121 [SBvT+08]; C represents data GSE1456 [PBA+05]; D represents data GSE2990 [SWL+06]; E represents data GSE4922 [IGS+06]; F represents data GSE7390 [DPL+07].

In general, we observed a large variability of prediction performances of individual methods between different datasets. This is not necessarily surprising, since it is known that the performance of any machine learning algorithms is dependent on the data at hand. Moreover, each dataset under study here contains different patients with unique characteristics and also clinical end points were slightly different (relapse free versus metastasis free survival after treatment). We are convinced that a comparison on a larger number of datasets reveals more of the true variability of an algorithm than a typical comparison on few selected ones.

In order to get a more objective and comprehensive view we conducted a ranking of all methods in each dataset according to the median cross-validated AUC value. We then calculated a consensus ranking based on the average rank of each method (Table 3.2). Interestingly enough, **aveExpPath** was ranked highest here. Two penalized SVM methods, **SCAD** and **HHSVM**, were ranked second, together with RRFE as a network based approach.

Some network-based methods (specifically network-based SVM, hub-based classification, pathBoost) revealed significantly higher gene selection sta-

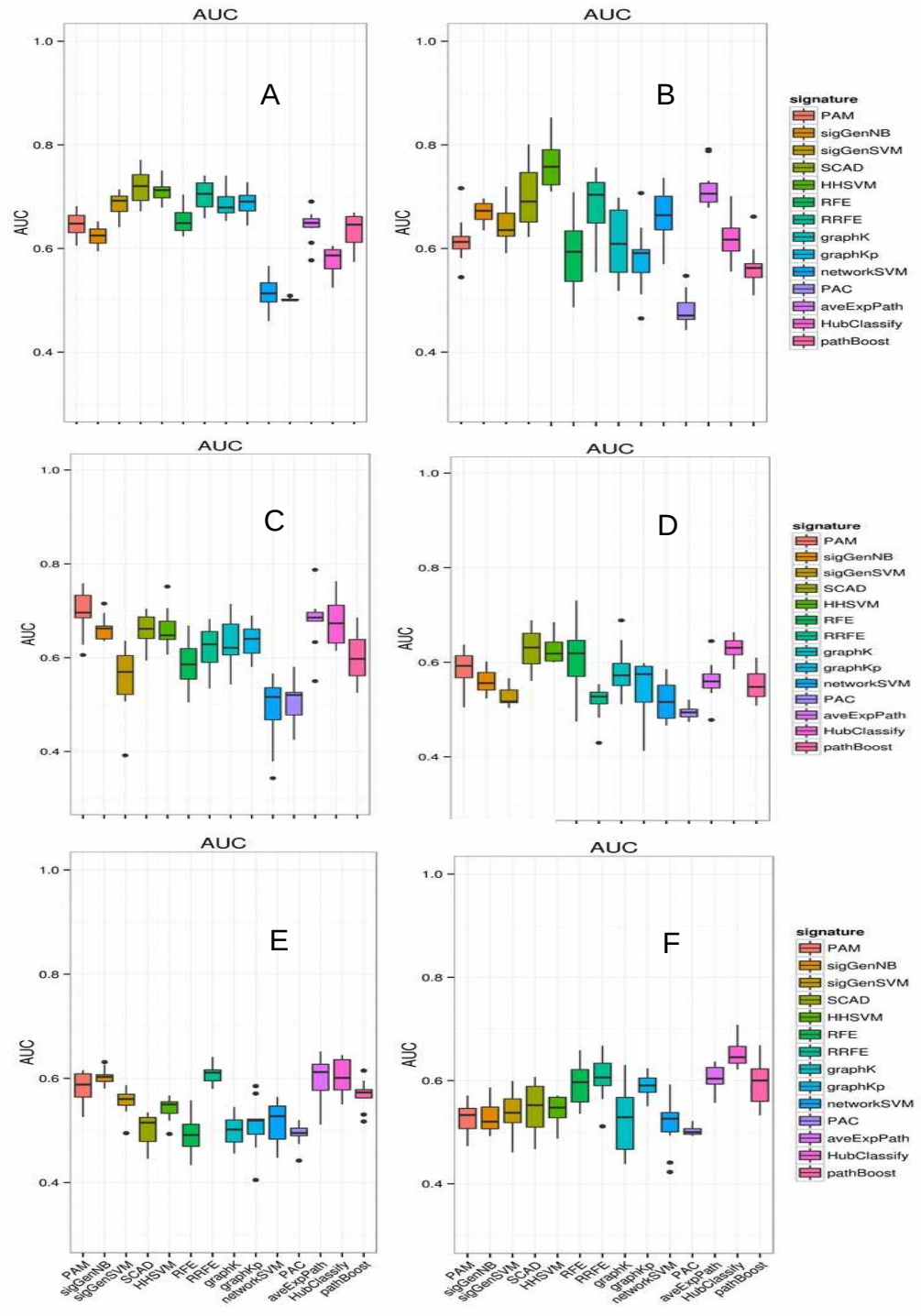



Figure 3.1: Prediction performance in terms of AUC.

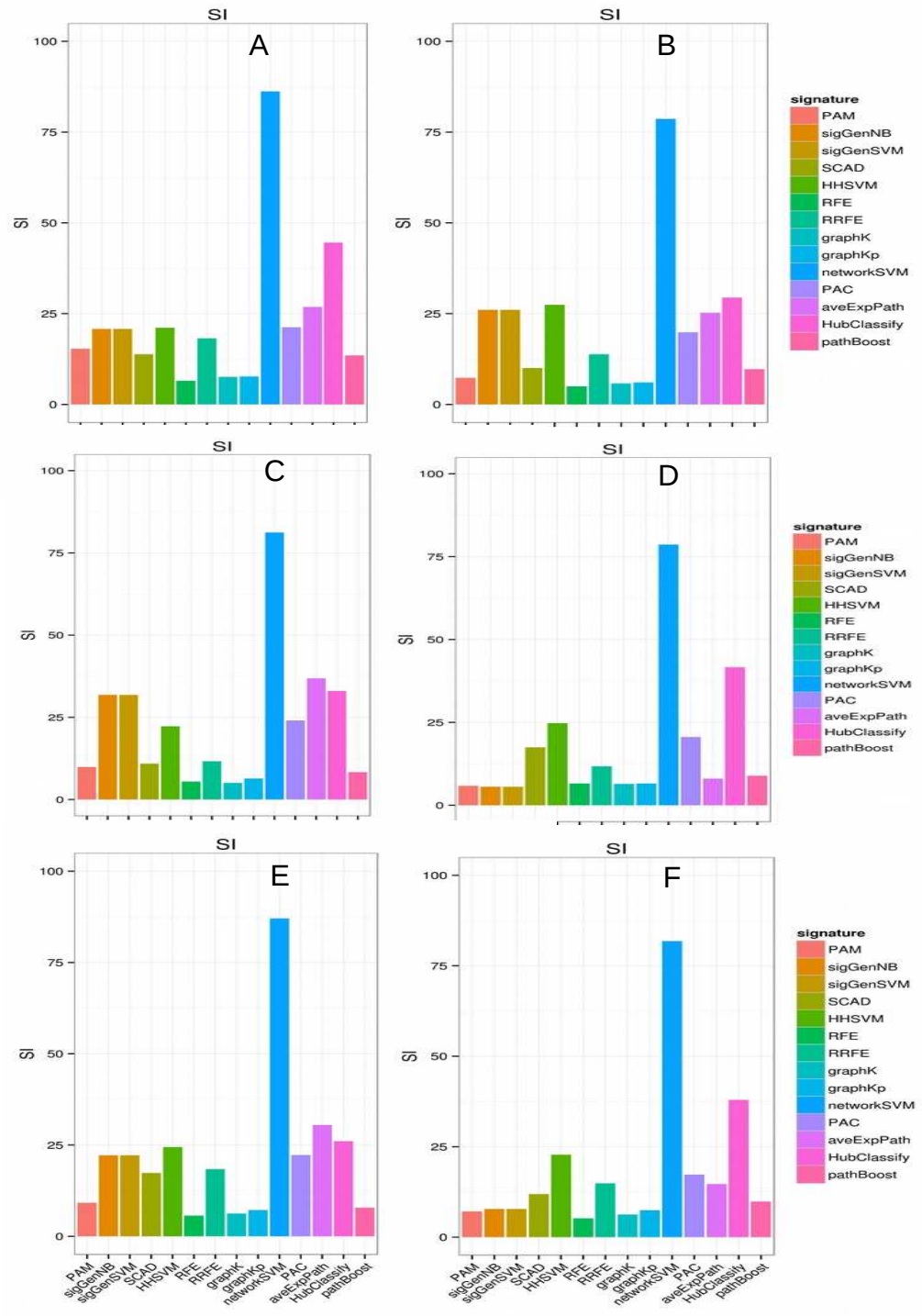



Figure 3.2: Signature stability.

Table 3.2: Ranking of different algorithms with respect to the median AUC in a 10 times repeated 10-fold cross-validation procedure.

| Method | G2034 | G11121 | G1456 | G2990 | G4922 | G7390 | consensus |
|---|---|---|---|---|---|---|---|
| PAM | 9 | 9 | 1 | 5 | 5 | 10 | 4 |
| sigGenNB | 11 | 5 | 4 | 9 | 3 | 13 | 5 |
| sigGenSVM | 4 | 7 | 12 | 12 | 7 | 9 | 7 |
| **SCAD** | **1** | **4** | **5** | **1** | **11** | **7** | **2** |
| **HHSVM** | **2** | **1** | **6** | **4** | **8** | **8** | **2** |
| RFE | 8 | 11 | 11 | 3 | 14 | 5 | 8 |
| **RRFE** | **3** | **3** | **8** | **11** | **2** | **2** | **2** |
| graphK | 6 | 10 | 9 | 7 | 12 | 11 | 10 |
| graphkKp | 5 | 12 | 7 | 6 | 10 | 6 | 6 |
| networkSVM | 13 | 6 | 14 | 13 | 9 | 12 | 11 |
| PAC | 14 | 14 | 13 | 14 | 13 | 14 | 12 |
| **aveExpPath** | **7** | **2** | **2** | **8** | **1** | **3** | **1** |
| HubClassify | 12 | 8 | 3 | 2 | 4 | 9 | 3 |
| pathBoost | 10 | 13 | 10 | 10 | 6 | 4 | 9 |

bility (Figure 3.2). Network-based SVMs performed clearly outstanding here. The reason might be two-fold: On one hand network-based SVMs come with a pre-filtering step of probesets according to their standard deviation, which already drastically reduces the set of considered probesets for the later learning phase and thus naturally enhances stability.

Network-based SVMs have a very effective mechanism for grouped selection of network connected genes via the infinity norm penalty [ZSP09]. Nonetheless, we found network-based SVMs to show a comparably poor prediction performance. This underlines that an improved gene selection stability does not necessarily coincide with better prediction performance. The reason for this behaviour could be that many genes reveal a high correlation in their expression. If such highly correlated genes are itself correlated with the patient group, then picking any of these genes leads to a similar prediction performance.

Picking preferentially one particular gene out of the correlated group (as tried by network-based approaches) increases gene selection stability, but does not necessarily increase prediction performance, either. This is exactly the behaviour we can observe in our datasets: Some network-based approaches (specifically networkSVM) have significantly improved gene selection stability, but do not perform consistently better than "conventional" methods, like PAM. We would like to point out that the high stability of network based SVMs and hub based classification is not at all associated to a higher number of selected genes (Figure 3.2).

As shown in Figure 3.2 and 3.3, which highlighted the much different behavior of **networkSVM** compared to all other approaches, which, given our previously discussed findings, was not very surprising. Most network-

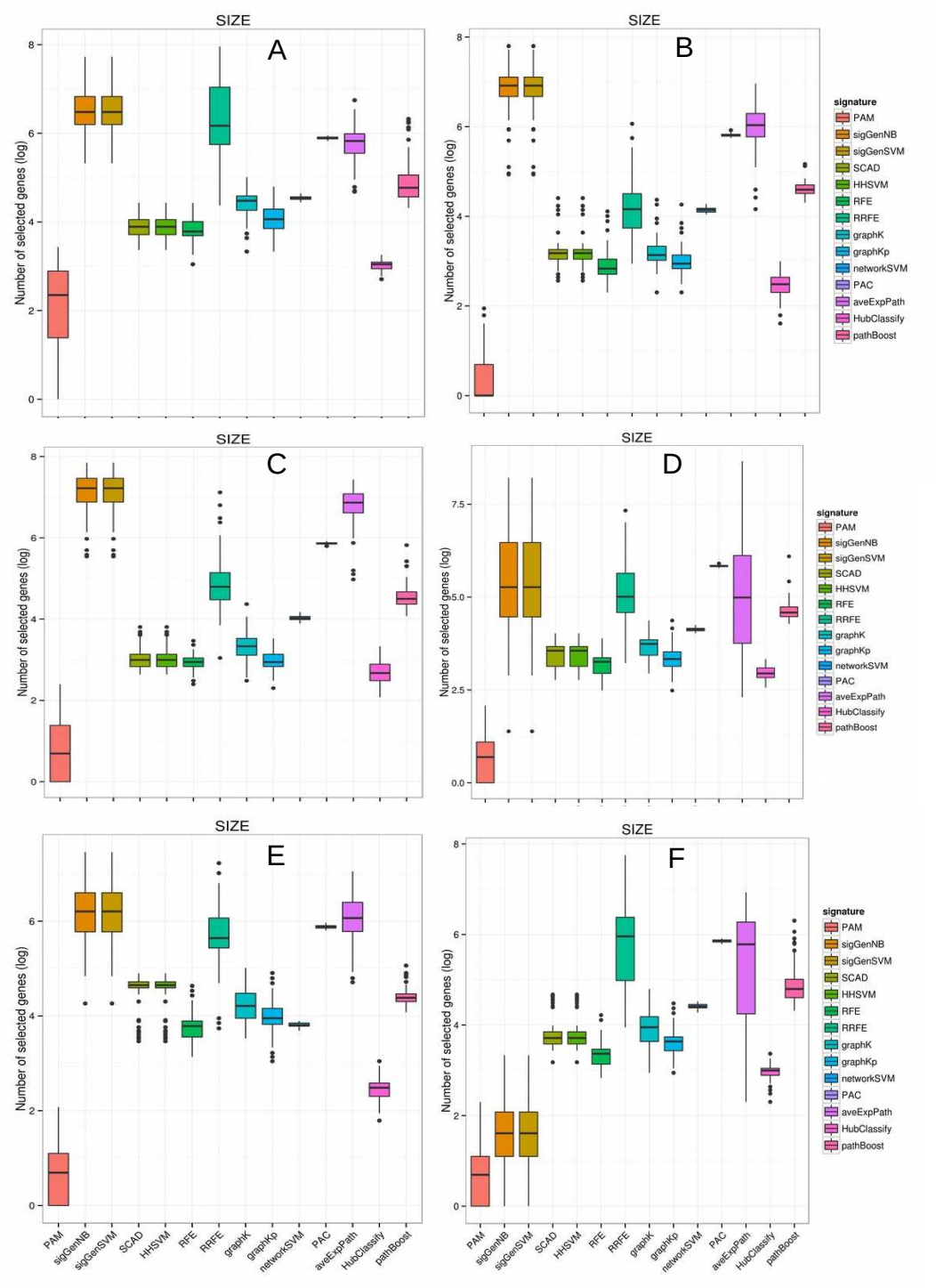



Figure 3.3: Number of selected genes per method. Y-axis is scaled by natural logarithm.

based method with respect to good gene selection stability. The high stability of this approach can be explained by the a-priori restriction on hub genes.

### 3.2.2 Cross datasets comparison

In order to the test the cross prediction performance, we selected the 4 top ranked gene selection algorithms according to Table 3.2 on the six breast cancer datasets. These methods are two network-based methods, namely RRFE and aveExpPath, and two classical approaches are HHSVM and SCAD. For each method, we trained in one datasets and tested on the other one. In consistency with our previous findings we observed RRFE and aveExpPath to show a better prediction performance than the two other methods here. (see Figure 3.4).

A consensus ranking based on the average rank of the prediction accuracy (AUC value) of each comparison study showed that aveExpPath ranked best in the cross dataset comparison, RRFE ranked second, and HHSVM and SCAD ranked as third (Table 3.3). This suggests that prior information might help to find better predictive biomarker signatures.

### 3.2.3 Biological interpretability of signatures

To investigate the biological interpretability of our found signatures, we performed an enrichment analysis with respect to KEGG pathways, Disease Ontology terms and known drug targets. For that purpose we trained

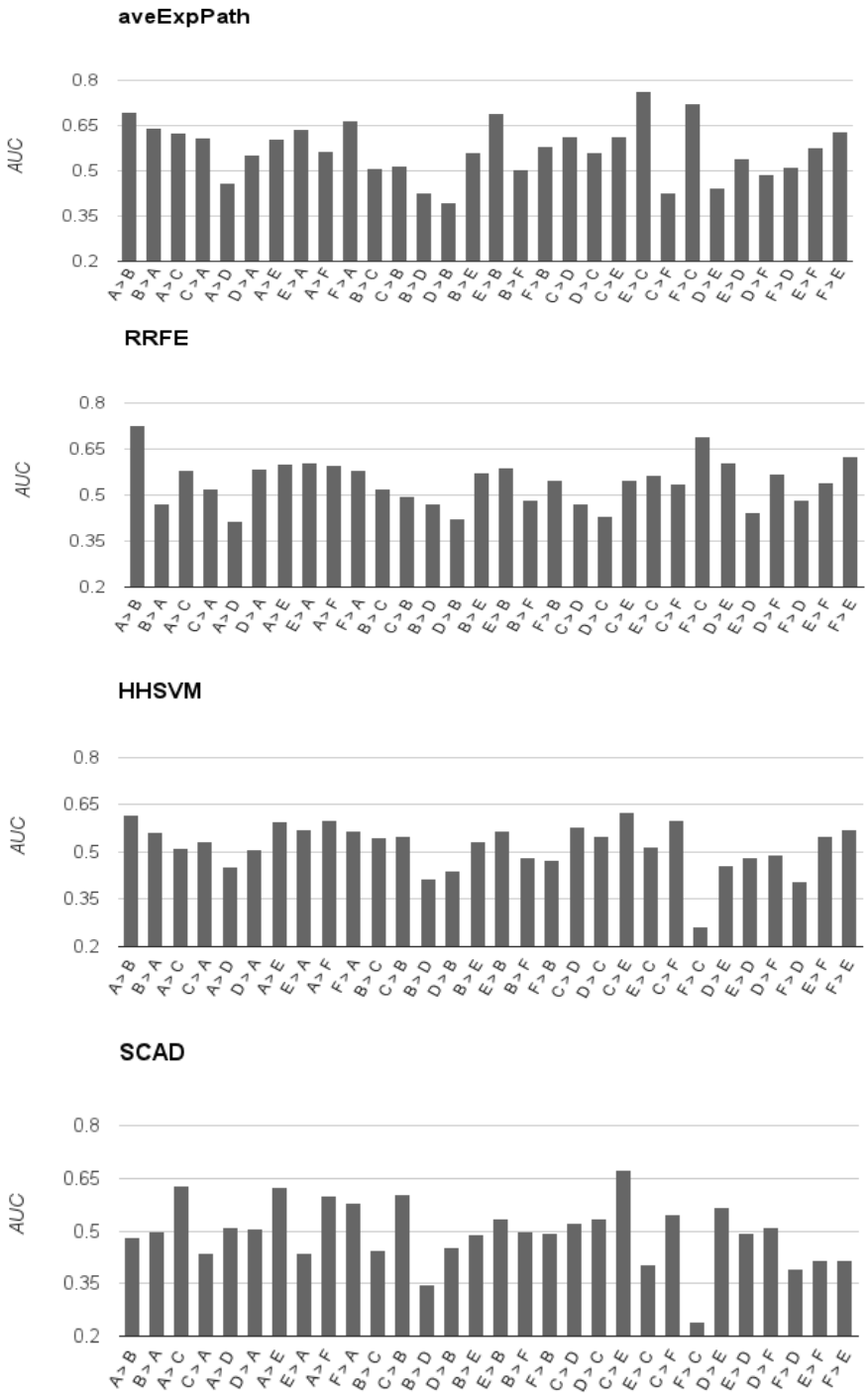



Figure 3.4: Cross comparison of 4 methods on 6 datasets. $A > B$ indicates training on dataset $A$ and predicting on dataset $B$.

Table 3.3: Ranking 4 selected algorithms according to AUC. $A > B$ indicates training on dataset $A$ and predicting on dataset $B$.

| cross comparison | aveExpPath | RRFE | HHSVM | SCAD |
|---|---|---|---|---|
| A > B | 2 | 1 | 3 | 4 |
| B > A | 1 | 4 | 2 | 3 |
| A > C | 2 | 3 | 4 | 1 |
| C > A | 1 | 3 | 2 | 4 |
| A > D | 2 | 4 | 3 | 1 |
| D > A | 2 | 1 | 4 | 4 |
| A > E | 2 | 3 | 4 | 1 |
| E > A | 1 | 2 | 3 | 4 |
| A > F | 4 | 3 | 2 | 1 |
| F > A | 1 | 3 | 4 | 2 |
| B > C | 3 | 2 | 1 | 4 |
| C > B | 3 | 4 | 2 | 1 |
| B > D | 2 | 1 | 3 | 4 |
| D > B | 4 | 3 | 2 | 1 |
| B > E | 2 | 1 | 3 | 4 |
| E > B | 1 | 2 | 3 | 4 |
| B > F | 1 | 3 | 4 | 2 |
| F > B | 1 | 2 | 4 | 3 |
| C > D | 1 | 4 | 2 | 3 |
| D > C | 1 | 4 | 2 | 3 |
| C > E | 3 | 4 | 2 | 1 |
| E > C | 1 | 2 | 3 | 4 |
| C > F | 4 | 3 | 1 | 2 |
| F > C | 1 | 2 | 3 | 4 |
| D > E | 4 | 1 | 3 | 2 |
| E > D | 1 | 4 | 3 | 2 |
| D > F | 4 | 1 | 3 | 2 |
| F > D | 1 | 2 | 3 | 4 |
| E > F | 1 | 3 | 2 | 4 |
| F > E | 1 | 2 | 3 | 4 |
| **consensus rank** | **1** | **2** | **3** | **3** |

each of the above described methods once on a whole dataset to retrieve a final gene signature.

In generally, this analysis revealed a high enrichment of disease related genes, KEGG pathways and known drug targets in signatures selected by network-based approaches (Figure 3.5, Figure 3.6, Figure 3.7). Specifically, RRFE (and partially also AveExpPath with regard to pathways) yielded an extremely high enrichment with respect to all three categories on all datasets. The overrepresentation of known drug targets for genes selected by RRFE was absolutely outstanding on all datasets. Consistently enriched KEGG-pathways for gene signatures selected by RRFE and aveExpPath were “Pathways in cancer”, “MAPK signaling pathway”, “ErbB signaling pathway”, “Adherens junction” and “Focal adhesion”, which have all been related to breast cancer [DYF+03, ONLH00, PBB99, PT00].

The reason for the good interpretability of pathways selected by AvgExpPath is directly clear, since this method focuses on selection of whole pathways. The outstanding interpretability of genes selected by RRFE can be explained as follows: RRFE uses a modification of Google’s PageRank algorithm (GeneRank – [MBHG05]) to compute for each gene a rank according to its own fold change and its connectivity with many other differentially expressed ones (guilt by association principle). This rank is then used to re-scale the hyperplane normal vector of a SVM. This method automatically leads to a preference of genes which are central in the network (c.f. [JBF+10]). These central genes are often well studied and directly known to be disease related [CBKB10].

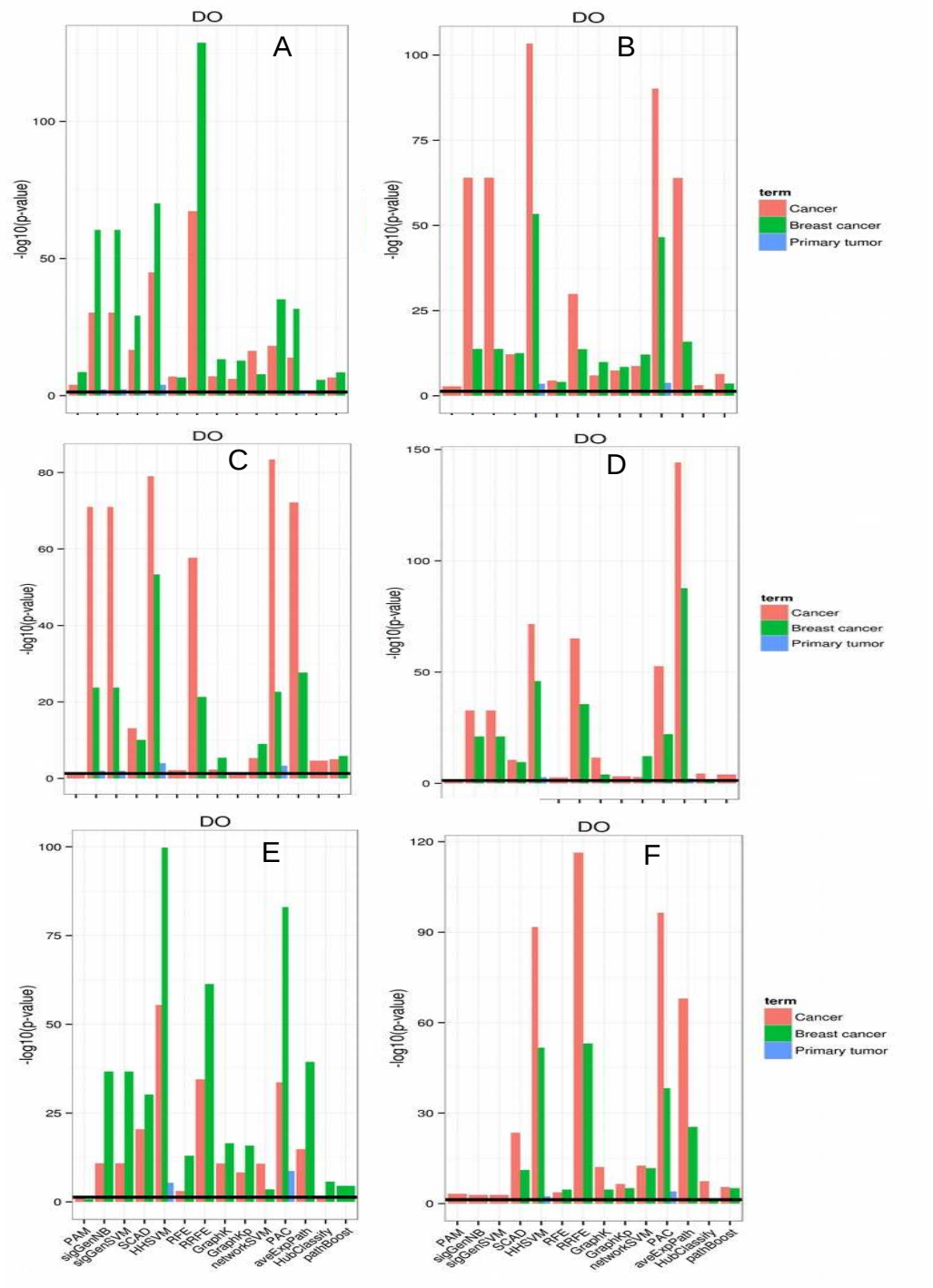



Figure 3.5: Interpretability of signatures (enriched disease genes). For AveExpPath and PAC the enrichment of the particular disease category within selected pathway genes is shown.

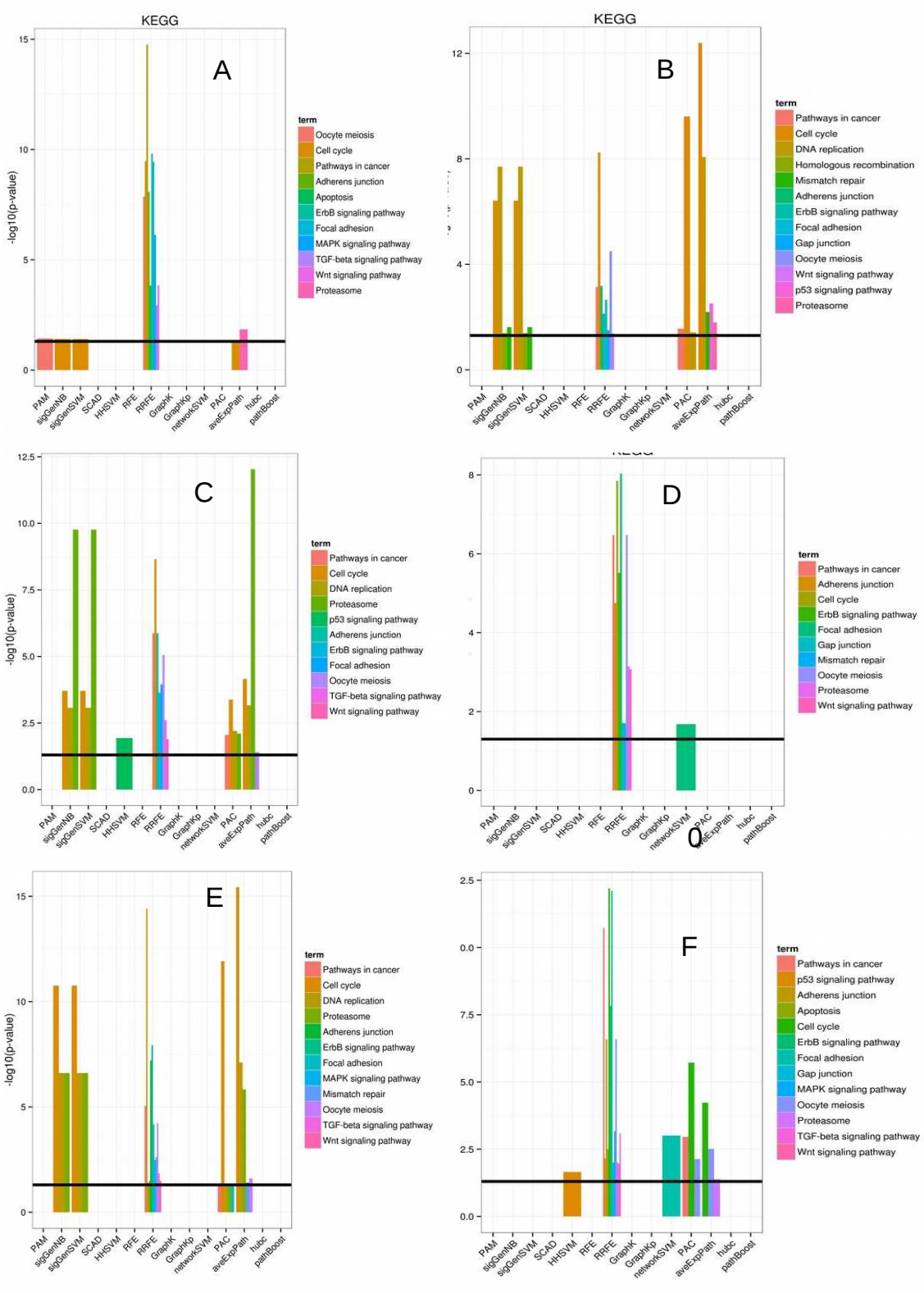



Figure 3.6: Interpretability of signatures (enriched KEGG pathways). For AveExpPath the adjusted p-value for differential expression from the SAM-test is shown. For all other methods we tested pathway enrichment within the set of selected genes.

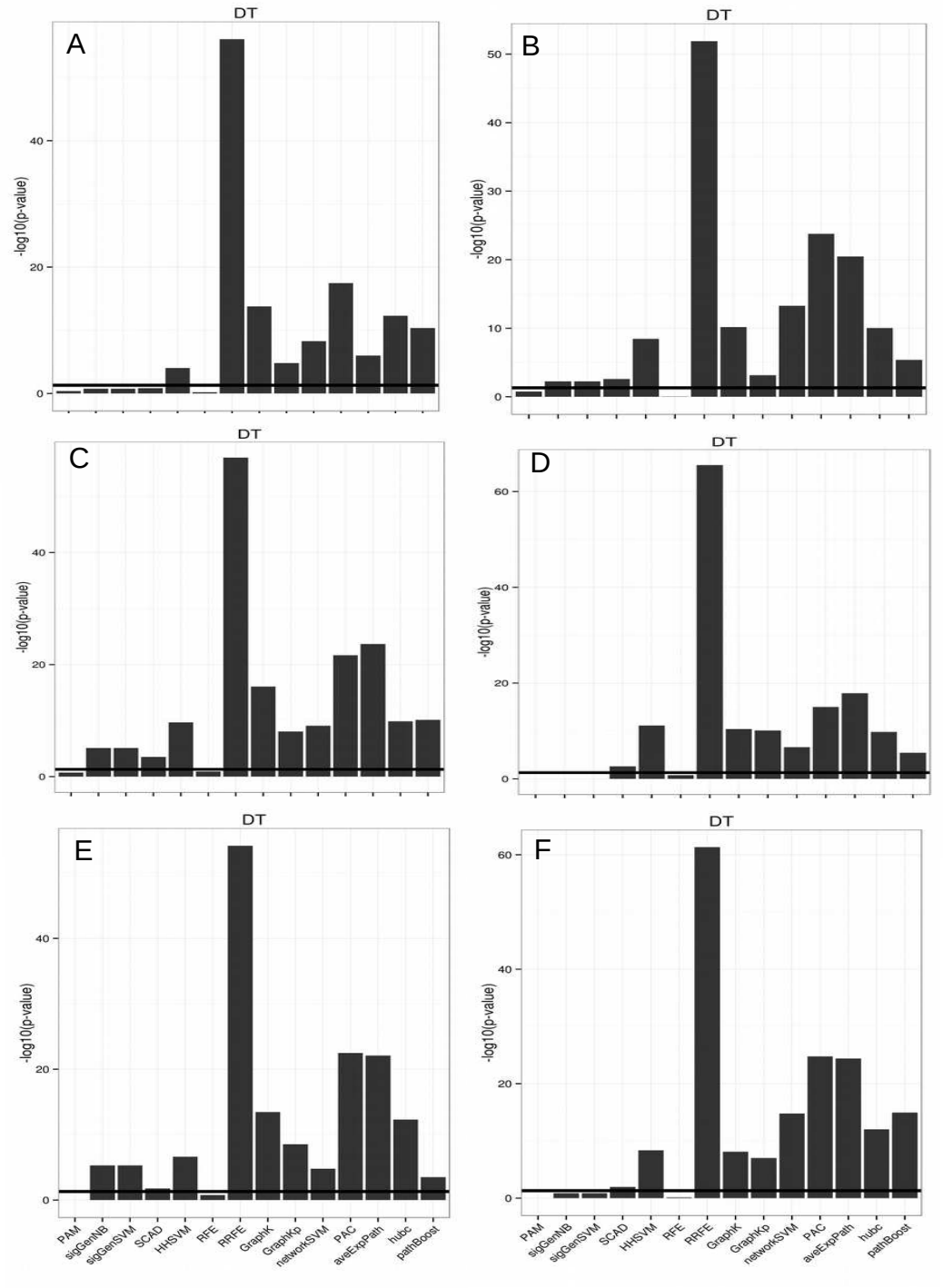



Figure 3.7: Interpretability of signatures (enriched drug targets). For AveExpPath and PAC the enrichment of drug targets within selected pathway genes is shown.

## 3.3 Conclusion

In this chapter, we performed a comprehensive and detailed comparison of fourteen gene selection methods (eight integrating network information) in terms of prediction performance, gene selection stability and interpretability on six public breast cancer datasets.

In general we found identify aveExpPath and RRFE to perform well with respect to all three categories. Moreover, we found that incorporating network or pathway knowledge into gene selection methods in general did not significantly improve classification accuracy compared to classical algorithms. Much more, the choice of the individual algorithm had a significant influence. Most network-based approaches not only drastically enhanced gene selection stability, but also showed a good prediction performance, such as aveExpPath and RRFE. Relatively simple gene selection methods, like average pathway expression, revealed a good prediction accuracy. Similar results have been reported in Haury et al. [HGV11]. Nonetheless, it is worth mentioning that the crucial assumption made by average pathway expression, namely that the mean pathway activity is altered significantly between two patient groups, might not always be fulfilled, for instance, if only few genes in a pathway are differentially expressed. Thus this method should be applied with care.

We found HHSVM and SCAD-SVM in most cases to show a better prediction performance than SVM-RFE. This is, for instance, in agreement with [WZZ08] and [BTLB11], who explained that by the fact that elastic net and SCAD penalties can better deal with correlated features, which are typically observed in gene expression data. In our comparison HHSVM, together with average pathway expression and RRFE, revealed the high-

est prediction performance.

Integrating additional experimental data, such as microRNA measurements, SNP or CNV data in addition to protein-protein interaction information might offer an alternative route to enhance prediction performance as well as stability and interpretability of biomarker signatures in the future.

To our knowledge this work is one of the most detailed and largest comparisons, which has been conducted so far to assess the performance of network-based gene selection methods in a multi-dimensional way. Whereas most previous approaches concentrated only on one aspect of gene selection methods, namely prediction performance, we have here also looked into stability and interpretability of the tested algorithms. Prognostic and diagnostic gene signatures are applied in a biomedical context. Thus, the classical machine learning based perspective of focusing only on prediction performance might be too narrow. Indeed we believe that stability and interpretability of gene signatures will strongly enhance their acceptance and practical practice for personalized medicine. Here we see the largest potential for methods, which incorporate biological background knowledge, for example in form of pathway knowledge, known disease relations or other approaches. This does not, of course, imply that prediction performance should be sacrificed for reproducibility or interpretability, but seen as an additional goal to achieve.

# Chapter 4

# Network and Data Integration for Biomarker Signature Discovery via Network Smoothed T-Statistics

"*Essentially, all models are wrong, but some are useful*."
*– George E. P. Box.*

In this chapter, we propose a new filter feature selection method, which integrates network information by smoothing gene wise t-statistics over the graph structure using a random walk kernel.

Various network based approaches have been proposed to integrate prior knowledge on canonical pathways, Gene Ontology (GO) annotation or protein-protein interactions into feature selection algorithms [GZL+05, CLL+07, RZD+07, LCK+08, TLWF+09, BS09, ZSP09, JBF+10]. A recent review on such approaches can be found in [CF12a]. The general hope

of these approaches is that biological knowledge can lead to better interpretable and more stable signatures. Whether network based classification methods automatically also lead to higher prediction accuracies is still a matter of debate [CF12c, SCK+12].

Another line of research focuses on the integration of different entities of experimental data for the same patient, e.g. mRNA and miRNA expression [VLV+10, GSMK+10, ZYK+11, GPF+11]. The increasing amount of different kinds of molecular data from the same patient, for instance within the TCGA database (www.cancergenome.nih.gov), now opens the door to a broader disease understanding [CHGM11, BBB+11, HAA+10]. Moreover, the integration of data capturing different molecular mechanisms could also lead to improved molecular signatures.

Our approach allows for a straight forward integration of different data entities, like mRNA and miRNA expression. Comparisons of our smoothed t-statistic SVM (stSVM) with several competing approaches on one of previously introduced breast cancer, two prostate cancer and an ovarian cancer dataset demonstrate a favorable prediction performance of early versus late relapse and a high signature stability. Moreover, obtained gene lists are highly enriched with known disease genes and KEGG pathways.

The content of this chapter is based on a previous publication in ***PloS ONE***[CF13].

## 4.1 Materials and methods

### 4.1.1 Datasets

We retrieved one previously described breast cancer [SBvT+08], one ovar-

ian cancer [BBB+11] dataset and two prostate cancer [SG09, TSH+10] from different data repositories. The breast cancer [SBvT+08] and one of the prostate cancer datasets [SG09] were measured on Affymetrix hgu133a microarrays. The purpose for selecting these datasets was on one hand to have mRNA and miRNA expression data available for the same patient and on the other hand to cover different tumor entities. It is expected that different tumor entities exhibit different biological properties, which in turn may have an effect on the performance of the algorithm that we propose here. The breast cancer dataset was picked as an arbitrary representative of the six breast cancer datasets described in the last chapter. The second prostate cancer dataset (MSKCC, [TSH+10]) and the ovarian cancer dataset (TCGA, [BBB+11]) were measured on Affymetrix HuEx 1.0 ST microarrays. The breast and first prostate cancer dataset were normalized via FARMS [HCO06]. The ovarian cancer and MSKCC datasets were downloaded as ready normalized and gene-wise aggregated data from the TCGA and MSKCC homepage, respectively. Both datasets, in contrast to the others, include gene as well as miRNA expression information. They are thus of particular interest here to test our proposed data integration strategy. As clinical end points we considered metastasis free (breast and prostate cancer) and relapse free (ovarian cancer) survival time after initial clinical treatment. For ovarian cancer only tumors with stages IIA - IV and grades G2 and G3 were considered, which after resection revealed at most 10mm residual cancer tissue and responded completely to initial chemotherapy.

Survival time information was dichotomized into two classes according whether or not patients suffered from a reported relapse / metastasis event within 5 years (breast, prostate dataset 1), 3 years (MSKCC prostate cancer dataset) and 1 year (ovarian), respectively. Patients with a sur-

vival time shorter than 5/3/1 year(s) without any reported event were not considered and removed from our datasets. This was done, because these patients can neither reliably be put into the early nor into the late relapse class. A summary of our datasets can be found in Table 4.1.

Table 4.1: Overview about employed datasets. msf: metastasis free survival; rfs: relapse free survival; rec: recurrent.

| ID/source | patients | cancer type | classification | positive class |
|---|---|---|---|---|
| GSE4922 | 228 | breast | mfs >5y | 69 |
| TCGA | 135 | ovarian | rfs >1y | 35 |
| GSE21032 | 79 | prostate | rsf >3y | 29 |
| GSE25136 | 79 | prostate | rec vs. non-rec | 40 |

### 4.1.2 Network information

#### Protein-Protein Interactions (PPI)

A comprehensive protein interaction network was compiled from the Pathway Commons database [CGD+11], which was downloaded in tab-delimited format (September 2012, and already described in Chapter 3). The network includes 11,361 nodes and 610,185 edges. Nodes in this network were identified with Entrez gene IDs. In order to consider genes with available probesets on the array but no corresponding network information we added for all these genes unconnected nodes to our initial network, resulting in 12,611 nodes for breast and the Sun et al. prostate cancer dataset; 11,356 nodes for ovarian cancer and 11,322 nodes for the MSKCC prostate cancer dataset. The reason for these differences is that not all dataset contain the same number of mappable transcripts.

#### KEGG pathways

As an alternative network information we computed a merger of all non-metabolic KEGG pathways [KAG+08]. For retrieval and merger of KEGG pathways, we employed the R-package KEGGgraph[ZCLS09]. Only gene-gene interactions were considered, which resulted in an initial network with 3,087 nodes and 17,518 edges. As before this initial network was extended to contain all genes available on the array, resulting in an overall network with the same number of nodes as described above for the PPI network but a different number of edges.

#### miRNA-target gene network

In addition to PPI or KEGG pathway information we optionally included predicted miRNA-target gene interactions. Target predictions were obtained from the MicroCosm database (version 5) [GJSvDE08] (FDR cut-off 1%). This increased the number of edges in the PPI network to 11,892 nodes for MSKCC's prostate cancer and 11,839 nodes for ovarian cancer.

### 4.1.3 Prediction accuracy, stability and interpretability

In order to assess the prediction performance of all tested methods we performed a 10 times repeated 10-fold cross-validation on each dataset. That means the whole data was randomly split into 10 fold, and each fold sequentially left out once for testing, while the rest of the data was used

for training and optimizing the classifier (including selection of relevant genes, hyper-parameter tuning, standardization of expression values for each gene to mean 0 and standard deviation 1, etc.). The whole process was repeated 10 times. It should be noted extra that also standardization of gene expression data was only done on each training set separately and the corresponding scaling parameters then applied to the test data.

The area under receiver operator characteristic curve (AUC) was used to measure the prediction accuracy via the R-package ROCR [SSBL05]. We use same gene selection stability index ($SI$) of Chapter 3 to investigate gene selection stability in more depth.

In order to check in how far signatures obtained by training the classifier on the whole dataset could be related to existing biological knowledge, we looked for enrichment of disease related genes via the tool FunDO [OFH$^+$09] (hypergeometric test; multiple testing correction: Bonferroni's method). Moreover, we calculated the enrichment with KEGG pathways [KAG$^+$08] via a hyper-geometric test.

### 4.1.4 Network smoothed t-statistic SVMs (stSVMs)

#### Network smoothed t-statistics

Given a simple, undirected graph $G = (V, E)$ with adjacency matrix $A$ the graph Laplacian $L$ is defined as $L := D - A$, where $D = diag(deg(v_1), ..., deg(v_n))$ is a diagonal matrix of node degrees for nodes $v_1, ..., v_n$ [Chu07]. The graph Laplacian can be viewed as a discrete approximation of the negative Laplace operator for functions.

One way of characterizing the degree of relatedness of two nodes (e.g. proteins) $v$ and $w$ in a graph (e.g. a PPI network) can be obtained via the

notion of random walks. The $p$-step random walk kernel is one particular similarity measure, which can be derived from this notion [GDCW09] and is defined as in section 2.6.3:

$$\begin{aligned} K &= (\alpha I - L^{norm})^p \\ &= ((\alpha - 1)I + D^{-1/2}AD^{-1/2})^p \end{aligned} \quad , \tag{4.1}$$

where

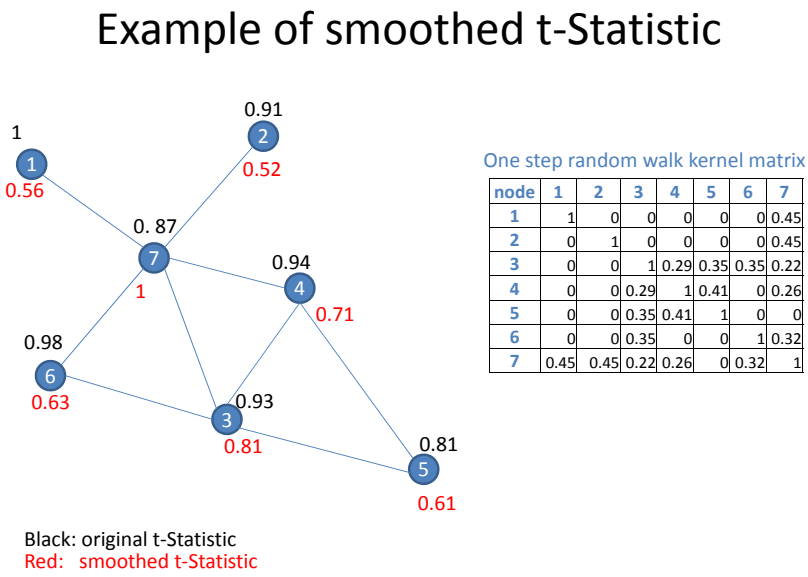
$$\begin{aligned} L^{norm} &:= D^{-1/2}LD^{-1/2} \\ &= I - D^{-1/2}AD^{-1/2} \end{aligned} \quad ,$$

is the normalized graph Laplacian matrix, $\alpha$ is constant, and $p$ is the number of random walk steps (here: $a = 1, p = 2$). The $p$-step random walk kernel gives rise to a symmetric, positive semi-definite similarity matrix between network nodes, capturing their degree of topological relatedness. The advantage compared to shortest path distance based measures is that alternative routes between pairs of nodes are considered. That means, if $v$ and $w$ are connected via many alternative paths of the same length this marks a higher similarity than if there exists only one such path.

Suppose for each network gene we assess its differential expression on the training dataset via a t-test. This results in an absolute t-statistic $|t_i|$ for network node $i$. We summarize the $|t_i|, i = 1, ..., |V|$ into a vector $\mathbf{t}$ and consider the score vector

$$\tilde{\mathbf{t}} = \mathbf{t}^T K. \tag{4.2}$$

Please note that $\tilde{t_i} = \sum_j |t_j| K_{ij}$. Hence, $\tilde{t_i}$ is a network smoothed version of $|t_i|$ (Figure 4.1), but does not follow a t-distribution any more. We thus conduct a permutation test (here: 1000 times) to obtain a p-value for each gene. For reasons of computation time we restrict this to the 10%




| node | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0.45 |
| 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0.45 |
| 3 | 0 | 0 | 1 | 0.29 | 0.35 | 0.35 | 0.22 |
| 4 | 0 | 0 | 0.29 | 1 | 0.41 | 0 | 0.26 |
| 5 | 0 | 0 | 0.35 | 0.41 | 1 | 0 | 0 |
| 6 | 0 | 0 | 0.35 | 0 | 0 | 1 | 0.32 |
| 7 | 0.45 | 0.45 | 0.22 | 0.26 | 0 | 0.32 | 1 |

Figure 4.1: Toy example to demonstrate the network smoothed t-statistic.

genes, which are highest ranked according to the network smoothed t-score (Equation 4.2). Multiple testing correction is then performed using the FDR approach by [BH95].

It is worth mentioning that the smoothing of absolute t-statistics particularly affects nodes with a high number of interaction partners. On one hand our procedure aggregates the scores of neighboring nodes to increase the score for these central proteins. On the other hand there is also a reverse effect, which increases the relevance of proteins in close proximity to hubs.

**SVM training**

We only select genes with FDR <5%. Subsequently a linear Support Vector Machine (SVM) is trained using the optimal parameter $C$ from

{0.0001, 0.001, ..., 10000}. To evaluate each candidate parameter $C$ we here used the span rule, which provides a theoretical upper bound for the leave-one-out cross-validation error, but can be computed much more efficiently for datasets with few samples [CVBM02]. It has been demonstrated theoretical as well as empirically that the span-rule provides an excellent mechanism for parameter selection in SVMs [CVBM02]. An implementation of this procedure can be found in R-packages pathClass [JFSB11] and netClass [CF14](see next Chapter).

**Integration of different experimental data**

Besides network information our approach allows for a straight forward integrating on of different experimental data, e.g. mRNA and miRNA expression, into one classifier. This can be achieved by extending adjacency matrix $A$ to miRNA-mRNA interactions and vector t to absolute t-statistics for miRNAs. Accordingly, network smoothing is now performed over protein-protein as well as miRNA-target gene interactions.

## 4.2 Results

### 4.2.1 stSVM shows overall best prediction performance

We initial considered our proposed stSVM method using only gene expression data and PPI network information. We compared the prediction performance to a number of competing methods, namely PAM [THNC02], a SVM trained with significant differentially expressed genes (FDR cut-off 5%) selected by SAM [TTC01] (sgSVM), average gene expression of KEGG pathways (aepSVM, [GZL+05]), pathway activity classification (PAC, [LCK+08]), reweighted recursive feature elimination (RRFE, [JBF+10]) and the netRank algorithm [WKK+12, CF12b]. NetRank, similar to RRFE,

uses a modification of Google's PageRank method to rank genes according to both, expression and network centrality [MBHG05]. The optimal number of selected genes in both cases was determined via the span-rule inside the cross-validation procedure [CVBM02].

For stSVM, netRank and RRFE, the same large PPI network was used as biological background information. The aepSVM and PAC methods use KEGG pathways. PAC relies on a so-called activity score, which is calculated per individual pathway and then taken as as a feature for classification purposes. For aepSVM we first conducted a global test [GVDV04] to select pathways being significantly associated with the class label (FDR cutoff 1%) on the training data and then calculated the mean expression of each selected pathway as a feature for SVM based classification. The prediction of all methods was assessed via a 10 times repeated 10-fold cross-validation procedure, as described in the Materials and Methods part of this paper.

Generally we observed a large variability of prediction performances of most tested algorithms across different datasets, which is in agreement with our previous observations [CF12c]. However, our proposed stSVM approach showed on all of our four gene expression datasets a consistently high prediction performance with respect to the area under ROC curve (AUC, Figure 4.2) and significantly outperformed several competing methods. Notably on two datasets (breast, prostate dataset 1) the AUC was extremely stable and showed only a very small variance across the cross-validation procedure.

In order to get a more objective and comprehensive view we conducted a ranking of all methods in each dataset according to the median cross-

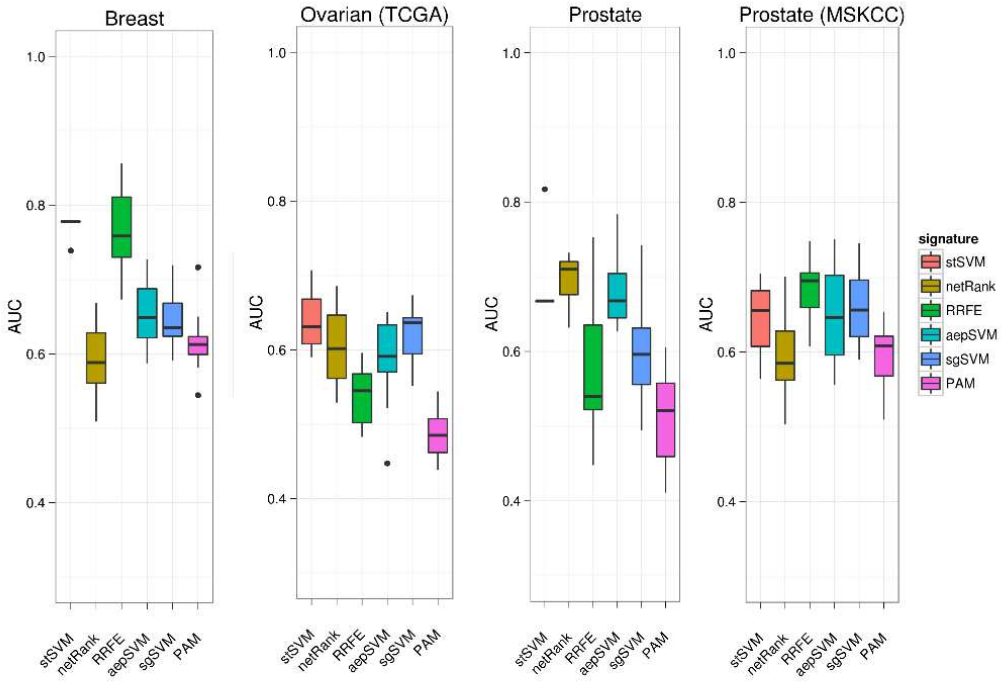



Figure 4.2: Prediction performance of stSVM in comparison to other methods in terms of area under ROC curve (AUC). Breast = GSE11121, Ovarian (TCGA)= GSE25136, Prostate = GSE25136, Prostate (MSKCC) = GSE21032.

Table 4.2: Ranking of different algorithms with respect to the median AUC in a 10 times repeated 10-fold cross-validation procedure.

| | breast | ovarian | prostate | prostate MSKCC | consensus |
|---|---|---|---|---|---|
| **stSVM** | **1** | **2** | **3** | **3** | **1** |
| netRank | 6 | 3 | 1 | 6 | 4 |
| RRFE | 2 | 5 | 5 | 1 | 3 |
| aepSVM | 3 | 4 | 2 | 4 | 5 |
| sgSVM | 4 | 1 | 4 | 2 | 2 |
| PAM | 5 | 6 | 6 | 5 | 6 |

validated AUC value. We then calculated a consensus ranking using Kendall's $\tau$ distance method [PDD09] (Table 4.2). The Kendall's $\tau$ distance measures the distance between two ordered lists. This confirmed our impression that stSVM was the overall best performing method. Interestingly enough, sgSVM was ranked second highest here, which is in agreement with our earlier finding that network based approaches do not consistently outperform classical ones [CF12c].

### 4.2.2 stSVM yields highly stable classification

We investigated the stability of signatures obtained during the 10 times repeated 10-fold cross-validation procedure using the concept of the stability index (Equation 3.1), showing for stSVM an extremely robust behavior (Figure 4.3). Most of the signature probesets were selected consistently during the cross-validation procedure. Interestingly enough, at the same time the number of selected probesets was comparably high for stSVM, which may be attributed to the fact that the network smoothing enforces the selection of correlated genes. As expected these genes typically reveal a high node degree in the PPI network. Many of these hub genes are well known to play a role in the disease pathology, e.g. BRCA1 for all tumors [GSD$^+$99, PCB$^+$96, FJP$^+$10] and AR for prostate cancer

[CCWH+99]. Other disease related and consistently selected genes include p53 (all datasets), EGFR (breast and prostate cancer [CSH99, BSF+04]), RB1 (breast and ovarian tumors [MV98, CSC+98, TST+99]) and EP300 (prostate cancer [BBL+12]).

### 4.2.3 stSVM signatures can be related to existing biological knowledge

In order to test the association with existing biological knowledge more systematically we trained each of our tested methods on complete datasets and subsequently tested the resulting signatures for enrichment of disease related genes and KEGG pathways and known drug targets (see Section 3.1.3 for detail description, Figures 4.4, 4.5, 4.6). For testing the association with disease related genes we used the FunDO tool [OFH+09], which is based on a hyper-geometric test.

Our analysis revealed a high enrichment of signatures obtained via stSVM to known disease genes and drug targets on all datasets. The enrichment was always higher than for non-network based methods (sgSVM, PAM) as well as for signatures obtained via the netRank algorithm. The latter might be attributed to the fact that netRank typically selects only very few genes, which thus could cause a loss of statistical power for enrichment analysis.

Besides disease related genes we also found a high enrichment of stSVM derived signatures for several KEGG pathways in all datasets (Figure 4.5). Examples were *Pathways in cancer* (prostate, breast cancer), *Prostate Cancer* (both prostate cancer datasets), *Wnt signaling, MAPK signaling* and *ERBB signaling*. The latter three were significant in breast and

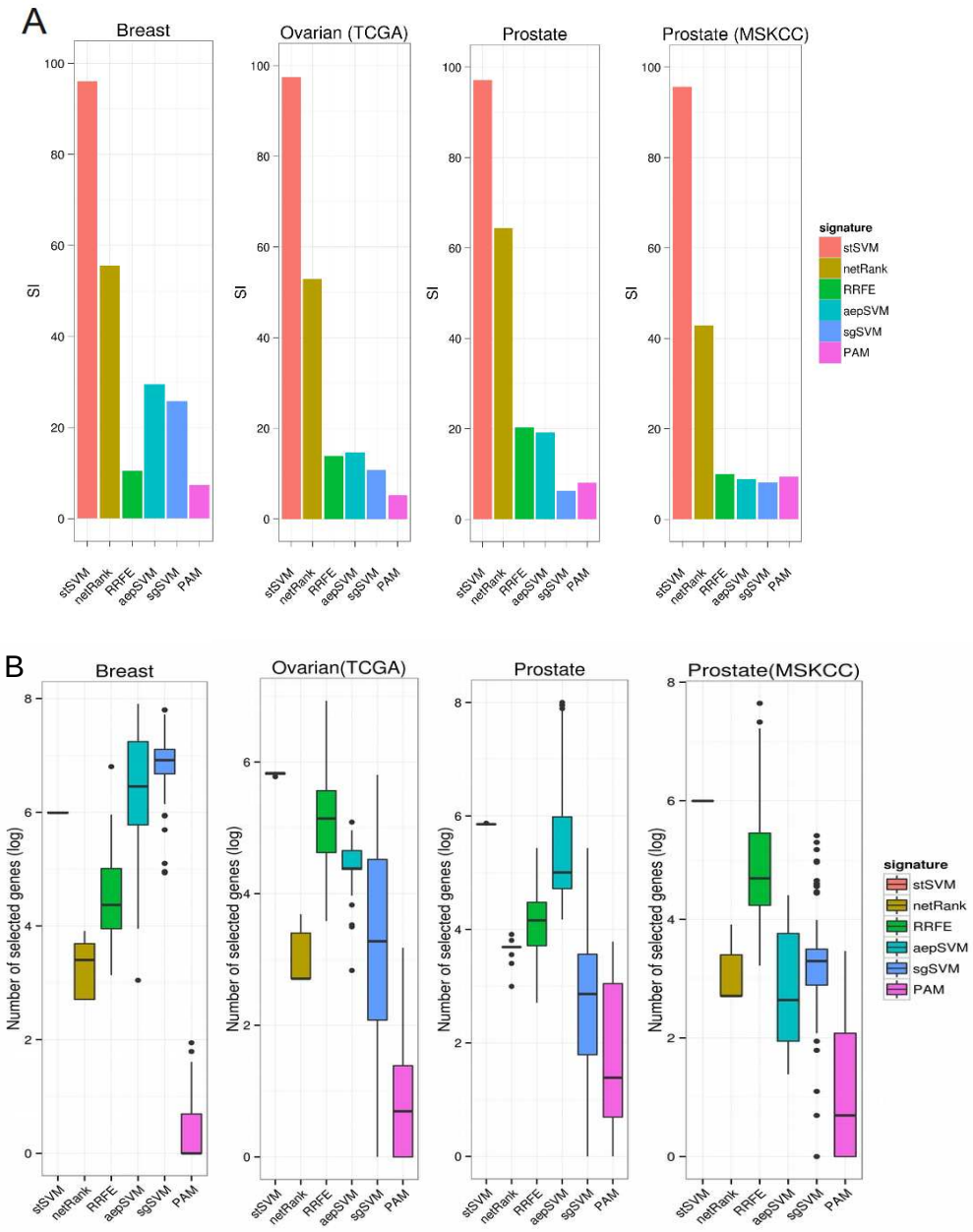



Figure 4.3: Stability index and signature sizes within the 10 times repeated 10-fold CV procedure. A) stability index according to Equation (3.1); B) Number of selected probesets. Y-axis is scaled by natural logarithms scale.

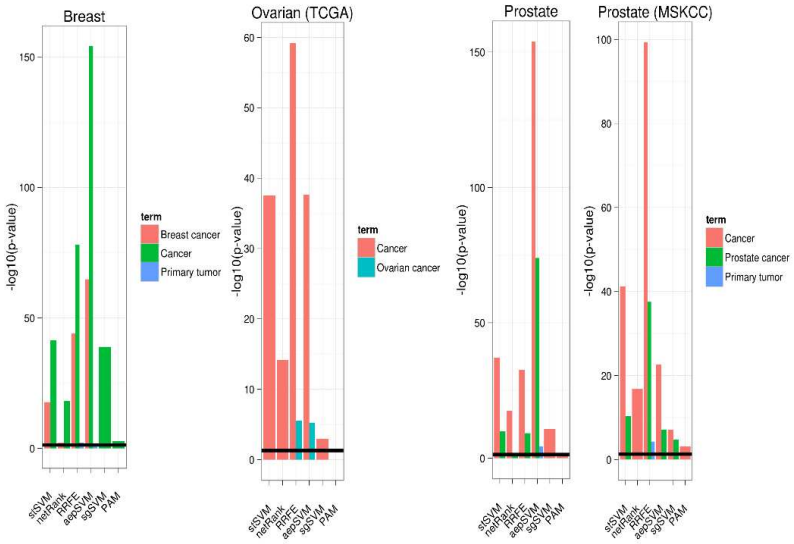



Figure 4.4: Enrichment of signatures with disease related genes. The y-axis shows -$\log_{10}$ p-values computed via a hypergeometric test (Bonferroni correction for multiple testing). Black horizontal line = 5% significance cutoff.

prostate cancer and are known to play a role in the respective disease pathologies [HB04, YB05, YB06, SOC+11, KCMPK+08, SP08, HBH+10]. In ovarian cancer we particularly detected a high enrichment of several metabolic pathways, such as *Fatty acid metabolism*. This fits to the fact that adipocytes were recently found to promote rapid tumor growth in ovarian tumors [NKP+11]. The significance of enrichment for KEGG pathways was generally higher for stSVM than for all other methods.

We also tested the enrichment with known drug targets (compare Chapter 3). This revealed for stSVM in all but one dataset (ovarian cancer) a highly significant result.

Taken together stSVM derived signatures showed a clear association to existing biological knowledge, which eases their biological understanding.

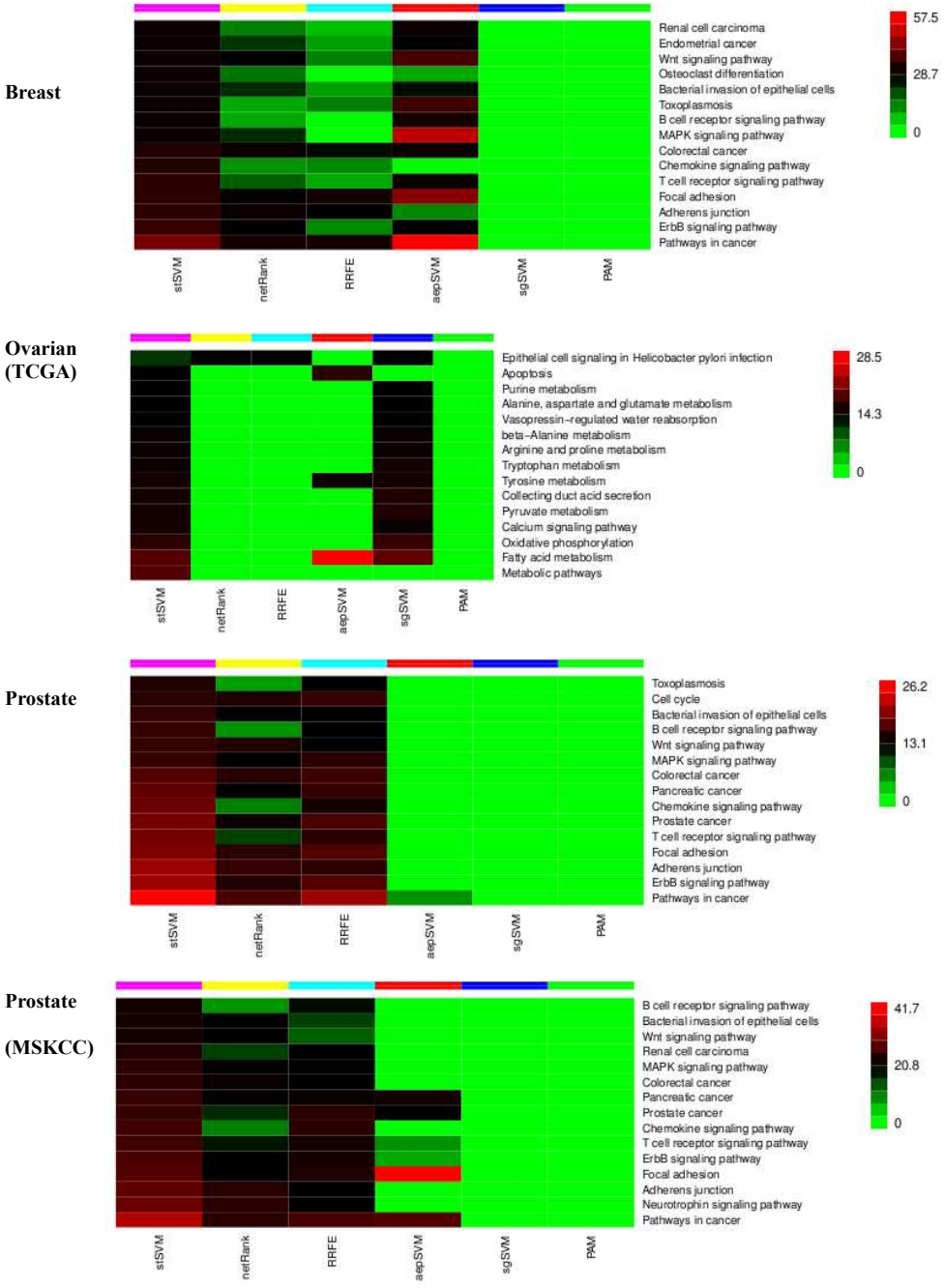



Figure 4.5: Enrichment of signatures (KEGG pathways). Only the 10 most significant pathways are shown for clearer visibility.

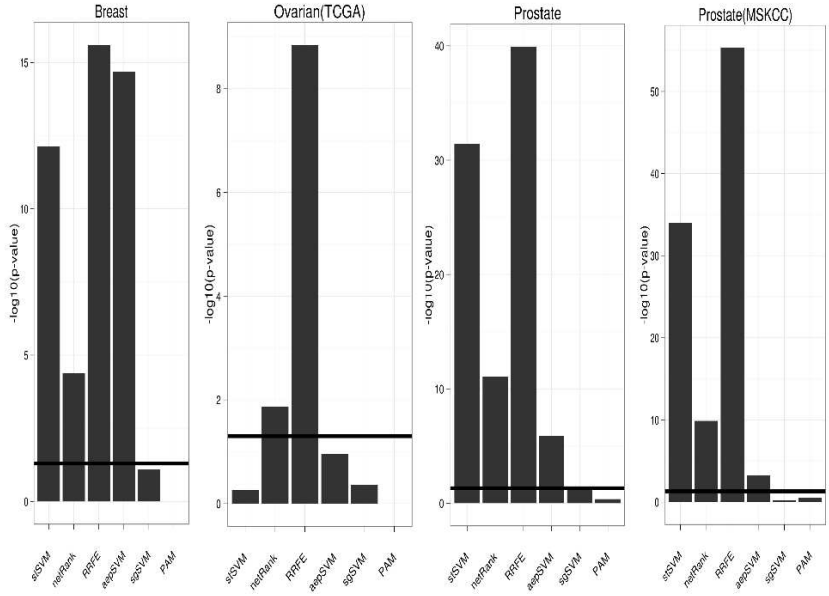



Figure 4.6: Enrichment of signatures with known drug targets.

### 4.2.4 Influence of network structure

We asked the question, in how far the observed good prediction performance of stSVM was dependent on the incorporated network structure. We hence re-ran our cross-validation procedure with a different network structure, which was compiled from a merger of all non-metabolic KEGG pathways (see Materials and Methods). It is worthwhile to mention that both networks contained the same number of nodes, but different number of edges. The KEGG derived network was much sparser then the previously used PPI network.

We observed that our original PPI network in all but one case (ovarian cancer dataset) yielded significantly higher AUCs, which highlights the principle influence of the network structure (Figure 4.7). We can only speculate why on the ovarian cancer dataset the KEGG based network

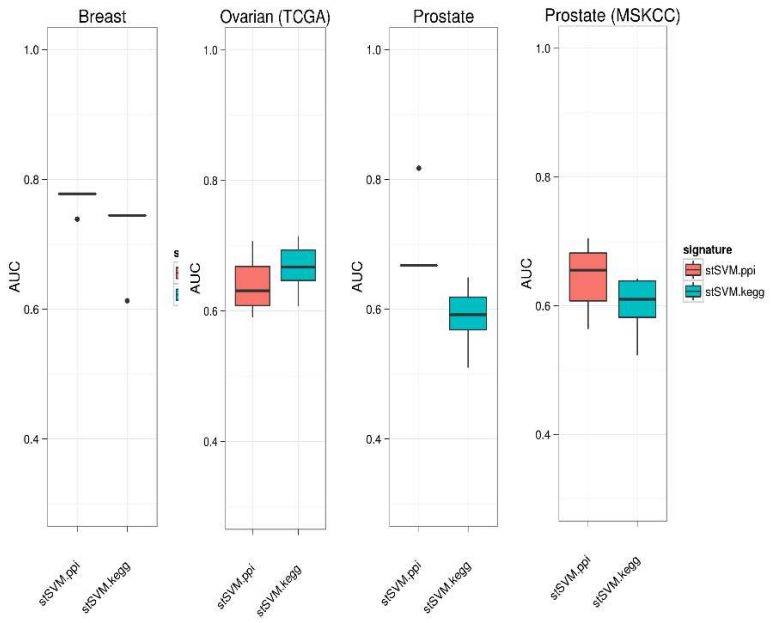



Figure 4.7: Classification performance of stSVM on two different network information.

appeared to work at least as good as the PPI network. Principally KEGG pathways capture different biological aspects (canonical pathways) than large scale protein-protein interaction networks. It may be due to the nature of the disease that KEGG pathways reflect better the relevant biology for ovarian cancer than for breast and prostate tumors.

### 4.2.5 Cross comparison in prostate cancer

In order to the test the prediction performance if our tested methods across different datasets, we focused on the two prostate cancer datasets. For each method, we trained in one datasets and tested on the other one. We observed that our stSVM and netRank revealed a similar good prediction performance across datasets (Figure 4.8).

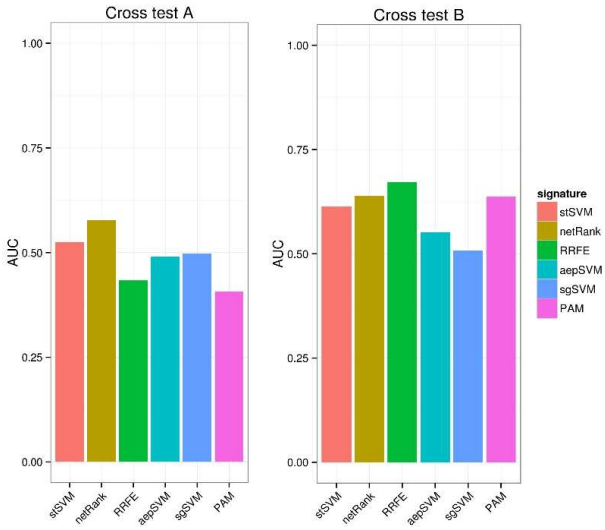



Figure 4.8: Cross comparison of 6 methods on two prostate cancer datasets. Cross test A: training on Prostate (MSKCC) cancer , test on Prostate. Cross test B: training on Prostate cancer , test on Prostate (MSKCC).

### 4.2.6 stSVM for mRNA and miRNA data integration

Our stSVM method allows for a straight forward integration of different types of experimental data on network level (see Materials and Methods). We here exemplify this property by using gene expression together with miRNA expression data for the TCGA ovarian cancer and for the MSKCC prostate cancer datasets. Correspondingly network information now consisted of a combined PPI and miRNA-target gene network. We call the corresponding variant of our method stSVM(mi-mRNA). We compared stSVM(mi-mRNA) to the graph fusion approach by Gade et al. [GPF+11] (GraphFusion). In their original paper Gade et al. used CoxBoost [BS09] to make survival risk prediction. In our classification based framework we replaced CoxBoost by the related PathBoost algorithm [BS09].

Moreover, we compared stSVM(mi-mRNA) to sgSVM trained on mRNA

data only, on miRNA data only and to a meta-classifier, which combines classification outputs from the mRNA / miRNA sgSVM classifiers into one consensus classifier (sgSVM(meta)). This was done as follows: The sgSVM method was separately trained on both datasets to yield a linear SVM classifier using significant differentially expressed genes and miRNAs, respectively. Each of these SVM classifiers yields a ranking (not classification) function of the form

$$f(\mathbf{w}) = \sum_{i=1}^{n} \alpha_i y_i \mathbf{w}_i + b,$$

where $\alpha_i$ are the fitted Lagrangian multipliers, $y_i \in \{-1, 1\}$ the class labels and $b$ the intercept (see section 2.6.3). Note that the corresponding classification function can be obtained by taking the sign of $f(\mathbf{w})$. Let $f_1(\mathbf{x})$, $f_2(\mathbf{z})$ be the SVM ranking functions for mRNA profile x and miRNA profile z, respectively. Then both rankings can be combined into a meta-classifier by fitting a logistic regression function

$$\Pr(y_i = 1 \mid f_1(\mathbf{x}), f_2(\mathbf{z})) = \frac{1}{1 + \exp(-\theta_0 - \theta_1 f_1(\mathbf{x}) - \theta_2 f_2(\mathbf{z}))},$$

where $\theta_0, \theta_1, \theta_2$ are parameters, which can be fitted to the data.

The comparison of our stSVM(mi-mRNA) approach to the graph fusion algorithm same to the above described meta-classifier approach (sgSVM(meta)) revealed a superior performance of our method. GraphFusion was outperformed with large margin (Figure 4.9), while the gain compared to sgSVM(meta) was still weakly / moderately significant ($p = 0.065$ for ovarian and $p = 0.041$ for prostate cancer; Wilcoxon signed rank test). In that context it was interesting that only on the prostate cancer dataset a significant improvement by integration of mRNA and miRNA data could be observed at all: The comparison of stSVM(meta) versus stSVM yielded

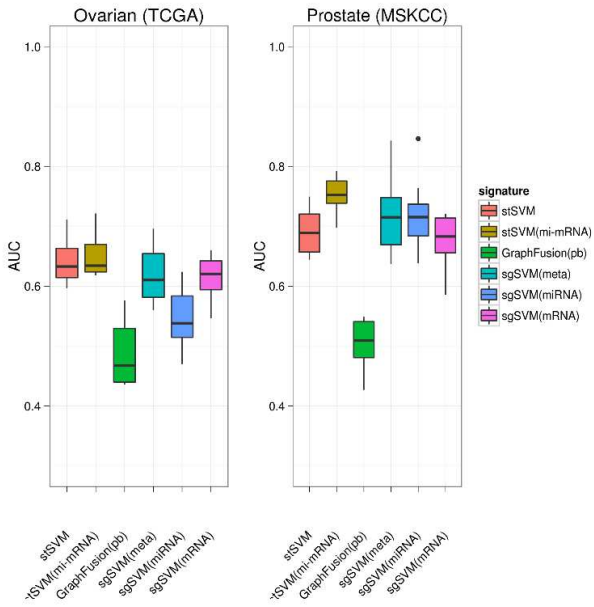



Figure 4.9: Prediction performance of stSVM on integrated gene and miRNA expression data compared to other approaches.

a p-value of $0.008$ (Wilcoxon signed rank test). On the ovarian cancer dataset miRNA expression data did not appear to contribute any useful classification information. This is also highlighted by the weak performance of the sgSVM classifier trained only on miRNA expression data (sgSVM(miRNA)).

### 4.2.7 Consistently signatures form disease related modules

Taking the set of genes and miRNAs, which were consistently selected by stSVM in the above investigated ovarian and MSKCC prostate cancer datasets, we asked the question, whether these features were connected to each other on network level, indicating that stSVM preferentially selected network connected genes and miRNAs.

To answer this question we looked for the largest sub-network that was purely formed by consistently selected features. In case of the ovarian cancer dataset we found 368 genes and 50 miRNAs out of 377 genes and 235 miRNAs to form such a network module. In case of the MSKCC prostate cancer dataset 384 genes and 96 miRNAs out of 386 genes and 254 miRNAs were inside one network module. This demonstrates that stSVM preferentially selected features, which were connected to each other on network level. The fraction of consistently selected genes that were inside one network module was, however, higher than the corresponding fraction of miRNAs. The reason could be that differential expression of a miRNA does not automatically imply that its target genes are also differentially expressed. Consequently miRNA markers do not always (but still in a significant proportion – see prostate cancer dataset) cluster together with gene markers on network level.

For both, ovarian and prostate cancer, network modules were highly enriched for known disease genes ($p = 4.39e-11$ for prostate cancer in MSKCC prostate cancer case, $p = 1.18e-3$ for ovarian cancer in ovarian cancer case) according to FunDO. Figure 4.10 and Figure 4.11 visualize sub-networks of these modules centered at the AR (MSKCC prostate cancer) and BRCA1 (ovarian cancer), respectively.

## 4.3 Discussion and conclusion

In this chapter we proposed network smoothed t-statistics as a method to integrate network information as well as different types of experimental data into one classifiers for biomarker signature discovery. Our method smoothed a widely used marginal statistic (the t-statistic) for differential

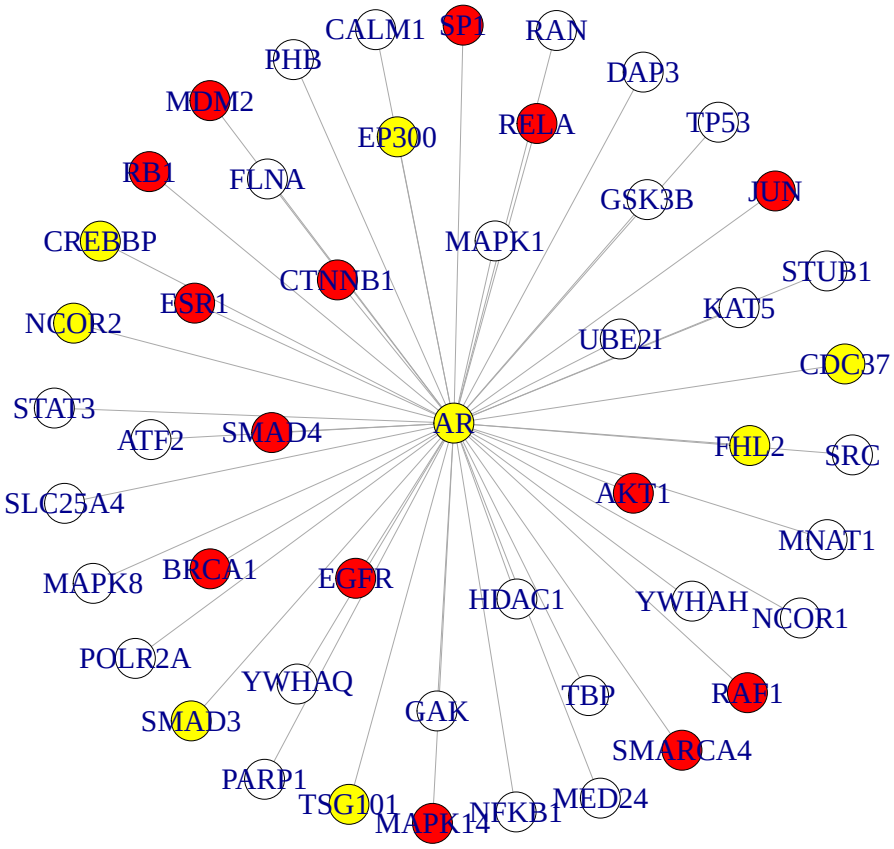



Figure 4.10: Sub-graph of disease related module of MSKC (prostate cancer), which identified by stSVM. The shown sub-graph consists of consistently selected genes in the interactome of the AR. For better visualization edges between neighbors of the AR are omitted. Red: cancer related genes; yellow: prostate cancer related genes.

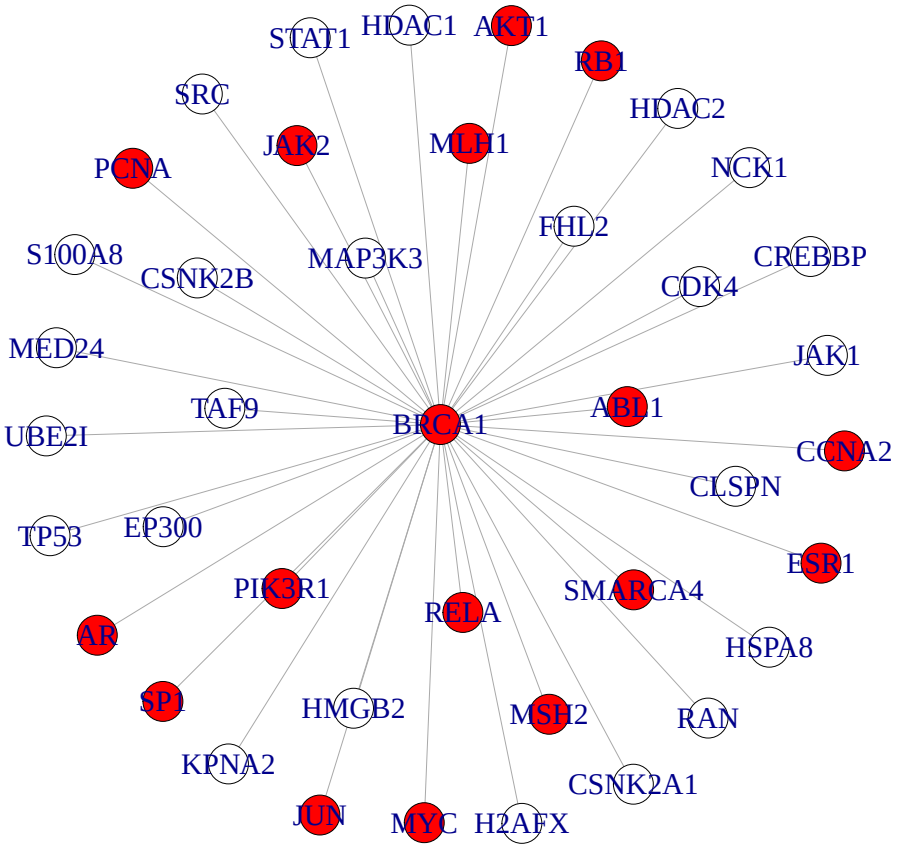



Figure 4.11: Sub-network of disease related module of (ovarian cancer) , which identified by stSVM. The shown sub-graph consists of consistently selected genes in the interactome of the BRCA1. For better visualization edges between neighbors of the BRCA1 are omitted. Red: cancer related genes.

expression over the graph structure of a biological network using random walk kernels. Our approach has on the technical level certain similarities with kernel based ranking methods for gene prioritization, which have been proposed e.g. by Moreau and co-workers to predict putative disease causing genes in genetic disorders [DTvOM07, GFMM12, MT12]. Note, that this is a rather different problem than finding prognostic biomarker signatures.

We showed that our approach overall leads to a highly predictive, stable and biologically interpretable classifier. We exemplified the straight forward integration of different types of experimental data here by building joint classifiers of gene and miRNA expression data. Other kinds of data (e.g. methylation, copy number variations) could principally be integrated in a similar manner. This is, however, not necessarily straight forward and thus subject to future research.

Taken together we think that our method is a step towards the challenging goal to build integrative classification models, which not only make use of biological background information, but also allow to combine various kinds of molecular data in order to make accurate predictions for an individual patient. In the light of the TCGA project and other large scale efforts the time is now ripe to move into this direction.

# Chapter 5

# netClass: An R-package for network based, integrative biomarker signature discovery

"If the only tool you have is a hammer, you tend to see every problem as a nail."

*– Abraham Maslow.*

In this chapter, we present our R-package *netClass*, which implements five network-based gene selection methods [CF14]. In addition, *netClass* is to our knowledge the first software that allows for integrating miRNA and mRNA expression data together with protein-protein interactions and predicted miRNA-target gene information [CF13] into one biomarker signature. *netClass* thus complements the functionality of *pathClass* [JFSB11]. It is worth emphasizing that *netClass* focuses on classification algorithms only. A software package that is more tailored to Cox regression is e.g. *CoxBoost [BS09]*.

## 5.1 Packages overview

*netClass* currently implements five network-based gene selection methods:

1. Average expression profile of pathways [GZL+05].
2. Pathway activity classification [LCK+08].
3. Classification based on differential expression of hub genes and correlated partners [TLWF+09].
4. Filtering of genes according to a modified Google PageRank algorithm [WKK+12, CF12b].
5. Random walk kernel based smoothing of t-statistics over a network structure [CF13].

Specifically, the latter approach also allows for integrating miRNA and mRNA expression data. Neither of the five above mentioned methods have been implemented in *pathClass*, which mainly focuses on the SVM-RFE algorithm and variants thereof [JFSB11]. Hence, *netClass* and *pathClass* complement each other.

Pathway activity classification is the only non-SVM based classification approach in *netClass*, since it uses logistic regression [LCK+08]. All the other algorithms internally use (linear) SVM classification. *netClass* enables to tune the soft margin parameter automatically in a computationally efficient manner using the span rule, which provides a theoretical

upper bound on the leave-one-out cross-validation error and can be calculated from training data only[CV99]. Furthermore, to evaluate the prediction performance of classification algorithms, in *netClass* feature selection and soft margin parameter tuning are embedded into a repeated $k$-fold cross-validation scheme. Cross-validation can be started via user friendly interface functions and allows for parallel computing.

### 5.1.1 Data and network integration via kernel based smoothing of t-statistics

A specific feature of *netClass* is the implementation of our recently proposed *stSVM* algorithm, which allows for joint integration of network information together with miRNA and mRNA expression data [CF13]. The basic idea behind *stSVM* is to smooth a feature-wise marginal statistic (like the commonly used t-statistic) over the structure of a joint protein-protein and miRNA-target gene interaction graph. For this purpose a random walk kernel is employed [GDCW09]. A permutation test is used to select features in a highly consistent manner, and then these features are employed for subsequent SVM training. In our paper we demonstrated the utility of this approach on four datasets from different tumor entities and specifically showed that integration of miRNA and mRNA expression could enhance the prediction power for prostate cancer prognosis [CF13].

### 5.1.2 Integration of *igraph*

*netClass* facilities the post-hoc analysis of obtained feature sets by integrating the R-package *igraph* [CN06]. Algorithms incorporating network

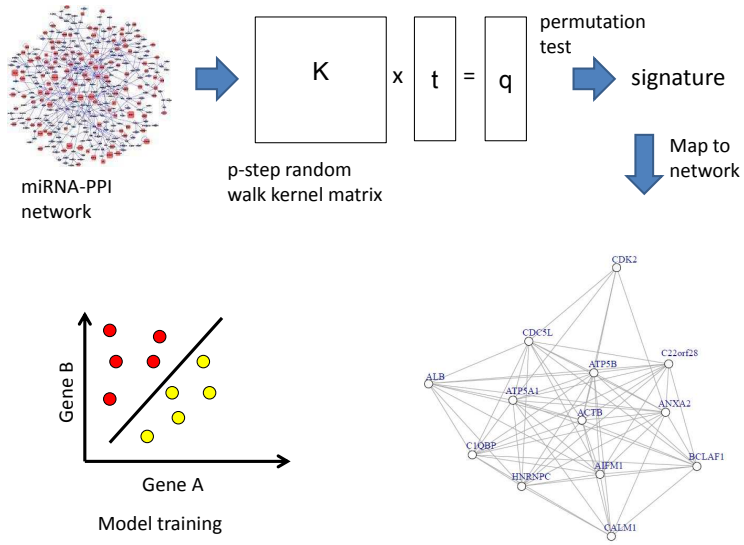



Figure 5.1: Workflow of stSVM: Marginal statistics for features in each -omics dataset are computed and smoothed over the structure of a joined miRNAPPI network. After re-ranking a permutation test selects the most relevant features and trains a SVM model. The obtained signature can be visualized as a network.

structures return the connected sub-graph(s) between selected features. This enables the full functionality of graph algorithms and plotting routines (Figure 5.1). In this context specifically Steiner tree methods as e.g. implemented in our package *SteinerNet* may provide a useful tool [SF13].

### 5.1.3 Example usage

To illustrate the use of *netClass* we show an example for running **stSVM** on a small sample dataset. First we get the sample data $expr$ with gene expression matrix $genes$, miRNA expression matrix and $miRNA$ class labels $y$. The adjacency matrix for the network is given in $ad.matrix$. We then train **stSVM** on the whole dataset and plot the sub-graph induced by selected features as following codes.

```
> library(netClass)
> data(expr)
> data(ad.matrix))
> data(EN2SY)
> dk <- calc.diffusionKernelp(L=ad.matrix, p=2,a=1)
>t=train.stsvm(x=cbind(expr$genes,expr$miRNA),y=expr$y,
Gsub=ad.matrix,dk=dk,EN2SY=EN2SY)
> plot(st$trained$graph)
```

## 5.2 Conclusion

*netClass* is an R-package that allows for network and data integration for biomarker signature discovery. It includes several published approaches for incorporating network information into gene selection. Moreover, *netClass* contains our recently published **stSVM** algorithm, which allows for additional integration of miRNA and mRNA expression data. All implemented methods can perform repeated cross-validation to estimate the prediction performance. Moreover, integration of *igraph* facilitates the follow-up analysis of selected features via graph algorithms and plotting functions. In summary we believe that *netClass* provides a useful tool for biomarker signature discovery in personalized medicine.

# Chapter 6

# Summary and Future Plans

"*W*e must know. We will know."
– *David Hilbert.*

THE work contained in this thesis can be summarized as development of algorithms for prognostic / diagnostic biomarker discovery, which integrate prior knowledge on protein-protein interactions as well as different types of omics data (here: mRNA and miRNA expression data).

## 6.1 Summary

In order to get a landscape overview on current feature selection algorithms for biomarker discovery, we first compared fourteen recognized published methods, which include network-based and classical approaches. The goal was to get a comprehensive overview of current methods in terms of prediction performance, gene selection stability and interpretability on six public breast cancer datasets. In our comparison study, we

found no methods always performed best on 6 breast cancer datasets, and incorporating protein-protein interaction network or pathway knowledge into gene selection methods in general did not significantly improve classification accuracy compared to classical algorithms. However, signatures obtained by network based methods could often be better interpreted in terms of existing biological knowledge. In addition to RRFE, an interesting results was that relatively simple gene selection methods, like average pathway expression, revealed a good prediction accuracy. Our comparison study was one of the most detailed and largest investigations that evaluate the performance between network-based and classical gene selection methods in a multi-dimensional manner.

In the next step, we developed network smoothed t-statistics as a method to integrate network information as well as different types of experimental data, such as microRNA and mRNA expression profiles into one classifiers for gene selection. Integrating different types of omic data might generally provide one possible way to improve prediction performance as well as stability and interpretability of molecular signatures. The method was named stSVM, it smoothes a widely used marginal statistic (the t-statistic) for differential expression values over the network structure of a biological network using random walk kernels. The technical principle of stSVM has similarities to kernel based ranking methods for disease gene prioritization [DTvOM07]. We showed that stSVM generally leads to a highly predictive, stable and biologically interpretable genes signatures. The stSVM allows for integrating gene and microRNA expression profiles into a joint classifier in a straight forward manner. This is a step forward towards the challenging goal to integrate multiple omics data to classification models, which not only employ prior biological information, but also make use of a combination of various kinds of molecular data of

individual patient.

In this thesis, we showed that the classical machine learning based approaches that only focus on prediction performance might be too narrow. In clinical diagnosis and prognosis, a small set of stable and interpretable biomarkers will strongly raise their acceptance and practical application value for personalized medicine. Most biomarkers, which are detected by current gene selection methods, are hard to functional validate. Integrating prior knowledge and multiple omics data types may thus be a way to overcome current problems [BGL11, KL12].

## 6.2 Personal future plans

Cancer is a complex disease and one of the leading lethal diseases worldwide, it is a genetic disease and caused by somatic mutations [VK04, Wei07]. Recent progress in genomics technologies enable to detect genetic or epigenetic alterations in the genomes of tumor cells in a high resolution resolution in 'real-time'. This allows to understand genetic variations in cancer at unprecedented detail. Powerful next-generation sequencing instruments and the ability of bioinformatics enable to accurately find somatic mutations in clinically characterized cancer samples [GW08, GW$^+$09, BEC$^+$12, Gin13]. Genomic alterations (mutations, copy number changes, structure variations, indels, genomic rearrangements, etc.), which dysregulate key intracellular signal transduction pathways influence the growth and survival of cells. Characterizing these genomic alteration events as well as their impact on cellular signal transduction pathways is thus a crucial step for the development of novel drugs for

cancer therapy.

The perspective for my future research will focus on the following two topics in cancer genomics. The first perspective project is developing computational and statistical tools for characterizing genetic alteration profiles of individual tumor samples, and identification of driver alterations, which cause oncogenesis or tumor survival. The integrative cancer genome analysis of individual tumor samples permit to identify critical driver or key abnormalities. Such abnormalities converge on a single molecular target that can be used as therapeutic target [PFCS+12]. The second step is to employ these discovered alterations in cancer genome to develop sensitive statistical models that ensure detection of accurate biomarker(s) for diagnosis and therapeutic application. Such biomarker(s) will help clinical doctors to tailor individual treatment, which is the aim of personalized medicine.

# Bibliography


[ABB+00] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al., *Gene ontology: tool for the unification of biology*, Nature genetics **25** (2000), no. 1, 25–29.

[Abe10] Shigeo Abe, *Support vector machines for pattern classification*, Springer, 2010.

[ADH+08] Noga Alon, Phuong Dao, Iman Hajirasouliha, Fereydoun Hormozdiari, and S. Cenk Sahinalp, *Biomolecular network motif counting and discovery by color coding.*, Bioinformatics **24** (2008), no. 13, i241–i249.

[AHVdP+10] Thomas Abeel, Thibault Helleputte, Yves Van de Peer, Pierre Dupont, and Yvan Saeys, *Robust biomarker identification for cancer diagnosis with ensemble feature selection methods*, Bioinformatics **26** (2010), no. 3, 392–398.

[AYP+11] Jaegyoon Ahn, Youngmi Yoon, Chihyun Park, Eunji Shin, and Sanghyun Park, *Integrative gene network construction for predicting a set of complementary prostate cancer genes.*, Bioinformatics **27** (2011), no. 13, 1846–1853.

[BA95] J. M. Bland and D. G. Altman, *Multiple significance tests: the bonferroni method.*, BMJ **310** (1995), no. 6973, 170 (eng).

[BAAS03] B. M. Bolstad, Irizarry R. A., M. Astrand, and T. P. Speed, *A comparison of normalization methods for high density oligonucleotide array data based on bias and variance*, Bioinformatics **19** (2003), 185–193.

[Bat94] R. Battiti, *Using mutual information for selecting features in supervised neural net learning.*, IEEE Trans Neural Netw **5** (1994), no. 4, 537–550 (eng).

[BBB+11] D Bell, A Berchuck, M Birrer, J Chien, DW Cramer, F Dao, R Dhir, P DiSaia, H Gabra, P Glenn, et al., *Integrated genomic analyses of ovarian carcinoma*.

[BBL+12] Christopher E Barbieri, Sylvan C Baca, Michael S Lawrence, Francesca Demichelis, Mirjam Blattner, Jean-Philippe Theurillat, Thomas A White, Petar Stojanov, Eliezer Van Allen, Nicolas Stransky, Elizabeth Nickerson, Sung-Suk Chae, Gunther Boysen, Daniel Auclair, Robert C Onofrio, Kyung Park, Naoki Kitabayashi, Theresa Y MacDonald, Karen Sheikh, Terry Vuong, Candace Guiducci, Kristian Cibulskis, Andrey

Sivachenko, Scott L Carter, Gordon Saksena, Douglas Voet, Wasay M Hussain, Alex H Ramos, Wendy Winckler, Michelle C Redman, Kristin Ardlie, Ashutosh K Tewari, Juan Miguel Mosquera, Niels Rupp, Peter J Wild, Holger Moch, Colm Morrissey, Peter S Nelson, Philip W Kantoff, Stacey B Gabriel, Todd R Golub, Matthew Meyerson, Eric S Lander, Gad Getz, Mark A Rubin, and Levi A Garraway, *Exome sequencing identifies recurrent spop, foxa1 and med12 mutations in prostate cancer.*, Nat Genet **44** (2012), no. 6, 685–689 (eng).

[BCR+12] Carsten Bokemeyer, Eric Van Cutsem, Philippe Rougier, Fortunato Ciardiello, Steffen Heeger, Michael Schlichting, Ilhan Celik, and Claus-Henning Köhne, *Addition of cetuximab to chemotherapy as first-line treatment for< i> kras< /i> wild-type metastatic colorectal cancer: Pooled analysis of the crystal and opus randomised clinical trials*, European Journal of Cancer (2012).

[BEC+12] Barillot, Emmanuel, Calzone, Laurence, Hupe, Philippe, Vert, Jean-Philippe, and Andrei Yu Zinovyev, *Computational systems biology of cancer*, vol. 47, CRC Press, 2012.

[Ben01] Benjamini, Y. and Yekutieli, D., *The control of the false discovery rate in multiple testing under dependency*, Annals of Statistics **29** (2001), 1165 – 1188.

[BGL11] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo, *Network medicine: a network-based approach to human disease*, Nature Reviews Genetics **12** (2011), no. 1, 56–68.

[BH95] Yoav Benjamini and Yosef Hochberg, *Controlling the false discovery rate: A practical and powerful approach to multiple testing*, Journal of the Royal Statistical Society. Series B (Methodological) **57** (1995), no. 1, pp. 289–300 (English).

[BH02] Pierre Baldi and G Wesley Hatfield, *Dna microarrays and gene expression: from experiments to data analysis and modeling*, Cambridge University Press, 2002.

[BHOS+08] Asa Ben-Hur, Cheng Soon Ong, Sören Sonnenburg, Bernhard Schölkopf, and Gunnar Rätsch, *Support vector machines and kernels for computational biology*, PLoS computational biology **4** (2008), no. 10, e1000173.

[BL97] Avrim L Blum and Pat Langley, *Selection of relevant features and examples in machine learning*, Artificial intelligence **97** (1997), no. 1, 245–271.

[BM98] P. Bradley and O. Mangasarian, *Feature selection via concave minimization and support vector machines*, Proc. 13th Int. Conf. Machine Learning, 1998, pp. 82 – 90.

[Bra97] Andrew P Bradley, *The use of the area under the roc curve in the evaluation of machine learning algorithms*, Pattern recognition **30** (1997), no. 7, 1145–1159.

[Bre01] Leo Breiman, *Random forests*, Machine Learning **45** (2001), 5–32, 10.1023/A:1010933404324.

[BS09] Harald Binder and Martin Schumacher, *Incorporating pathway information into boosting estimation of high-dimensional risk prediction models.*, BMC Bioinformatics **10** (2009), 18 (eng).

[BSF+04] Magdalena Brys, Magdalena Stawinska, Marek Foksinski, Andrzej Barecki, Cezary Zydek, Eugeniusz Miekos, and Wanda M Krajewska, *Androgen receptor versus erbb-1 and erbb-2 expression in human prostate neoplasms.*, Oncol Rep **11** (2004), no. 1, 219–224 (eng).

[BTLB11] Natalia Becker, Grischa Toedt, Peter Lichter, and Axel Benner, *Elastic scad as a novel penalization method for svm classification tasks in high-dimensional data.*, BMC Bioinformatics **12** (2011), 138.

[BTW+11] Tanya Barrett, Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Rolf N Muertter, Michelle Holko, Oluwabukunmi Ayanbule, Andrey Yefanov, and Alexandra Soboleva, *Ncbi geo: archive for functional genomics data sets–10 years on.*, Nucleic Acids Res **39** (2011), no. Database issue, D1005–D1010 (eng).

[BWS+08] S. Bentink, S. Wessendorf, C. Schwaenen, M. Rosolowski, W. Klapper, A. Rosenwald, G. Ott, A. H. Banham, H. Berger, A. C. Feller, M-L. Hansmann, D. Hasenclever, M. Hummel, D. Lenze, P. MÃ¶ller, B. Stuerzenhofecker, M. Loeffler, L. Truemper, H. Stein, R. Siebert, R. Spang, and Molecular Mechanisms in Malignant Lymphomas Network Project of the, *Pathway activation patterns in diffuse large b-cell lymphomas.*, Leukemia **22** (2008), no. 9, 1746–1754.

[BWT+09] Natalia Becker, Wiebke Werft, Grischa Toedt, Peter Lichter, and Axel Benner, *penalizedsvm: a r-package for feature selection svm classification.*, Bioinformatics **25** (2009), no. 13, 1711–1712 (eng).

[BYC+06] Andrea H Bild, Guang Yao, Jeffrey T Chang, Quanli Wang, Anil Potti, Dawn Chasse, Mary-Beth Joshi, David Harpole, Johnathan M Lancaster, Andrew Berchuck, John A Olson, Jeffrey R Marks, Holly K Dressman, Mike West, and Joseph R Nevins, *Oncogenic pathway signatures in human cancers as a guide to targeted therapies.*, Nature **439** (2006), no. 7074, 353–357.

[BZK11] Michalis E Blazadonakis, Michalis E Zervakis, and Dimitris Kafetzopoulos, *Complementary gene signature integration in multiplatform microarray experiments*, Information Technology in Biomedicine, IEEE Transactions on **15** (2011), no. 1, 155–163.

[CBKB10] Sreenivas Chavali, Fredrik Barrenas, Kartiek Kanduri, and Mikael Benson, *Network properties of human disease genes with pleiotropic effects.*, BMC Syst Biol **4** (2010), 78.

[CC+10] Brian Charlesworth, Deborah Charlesworth, et al., *Elements of evolutionary genetics*.

[CCWH+99] L. Correa-Cerro, G. Wöhr, J. Häussler, P. Berthon, E. Drelon, P. Mangin, G. Fournier, O. Cussenot, P. Kraus, W. Just, T. Paiss, J. M. Cantú, and W. Vogel, *(cag)ncaa and ggn repeats in the human androgen receptor gene are not associated with prostate cancer in a french-german population.*, Eur J Hum Genet **7** (1999), no. 3, 357–362 (eng).

[CF12a] Yupeng Cun and Holger Fröhlich, *Biomarker gene signature discovery integrating network knowledge*, Biology **1** (2012), no. 1, 5–17.

[CF12b] ———, *Integrating prior knowledge into prognostic biomarker discovery based on network strucutre*, (preprint) arXiv:1212.3214 (2012).

[CF12c] ———, *Prognostic gene signatures for patient stratification in breast cancer - accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions*, BMC Bioinformatics **13:69** (2012).

[CF13] ———, *Network and data integration for biomarker signature discovery via network smoothed t-statistics*, PloS one **8** (2013), no. e73074, 9.

[CF14] ———, *netclass: An r-package for network based, integrative biomarker signature discovery*, Bioinformatics (2014), btu025.

[CFPL09] Marc Carlson, Seth Falcon, Herve Pages, and Nianhua Li, *Affymetrix human genome u133 set annotation data (chip hgu133a) assembled using data from public repositories*, 2009.

[CGD+11] Ethan G Cerami, Benjamin E Gross, Emek Demir, Igor Rodchenkov, Ozgün Babur, Nadia Anwar, Nikolaus Schultz, Gary D Bader, and Chris Sander, *Pathway commons, a web resource for biological pathway data.*, Nucleic Acids Res **39** (2011), no. Database issue, D685–D690.

[CHGM11] Lynda Chin, William C Hahn, Gad Getz, and Matthew Meyerson, *Making sense of cancer genomic data*, Genes & development **25** (2011), no. 6, 534–555.

[Chu07] Fan Chung, *The heat kernel as the pagerank of a graph*, Proceedings of the National Academy of Sciences **104** (2007), no. 50, 19735–19740.

[CK10] Salim A Chowdhury and Mehmet Koyutürk, *Identification of coordinately dysregulated subnetworks in complex phenotypes.*, Pac Symp Biocomput (2010), 133–144.

[CKZ+07] Sean R Collins, Patrick Kemmeren, Xue-Chu Zhao, Jack F Greenblatt, Forrest Spencer, Frank C P Holstege, Jonathan S Weissman, and Nevan J Krogan, *Toward a comprehensive atlas of the physical interactome of saccharomyces cerevisiae.*, Mol Cell Proteomics **6** (2007), no. 3, 439–450.

[CLL+07] Han-Yu Chuang, Eunjung Lee, Yu-Tsueng Liu, Doheon Lee, and Trey Ideker, *Network-based classification of breast cancer metastasis.*, Mol Syst Biol **3** (2007), 140 (eng).

[CN06] Gabor Csardi and Tamas Nepusz, *The igraph software package for complex network research*, InterJournal **Complex Systems** (2006), 1695.

[CNCK11] Salim A Chowdhury, Rod K Nibbe, Mark R Chance, and Mehmet Koyutürk, *Subnetwork state functions define dysregulated subnetworks in cancer.*, J Comput Biol **18** (2011), no. 3, 263–281.

[CR07] Kevin Chen and Nikolaus Rajewsky, *The evolution of gene regulation by transcription factors and microrna*s, Nature Reviews Genetics **8** (2007), no. 2, 93–103.

[CSC+98] C. Ceccarelli, D. Santini, P. Chieco, M. Taffurelli, M. Gamberini, S. A. Pileri, and D. Marrano, *Retinoblastoma (rb1) gene product expression in breast carcinoma. correlation with ki-67 growth fraction and biopathological profile.*, J Clin Pathol **51** (1998), no. 11, 818–824 (eng).

[CSH99] J. H. Clement, J. Sänger, and K. Höffken, *Expression of bone morphogenetic protein 6 in normal mammary tissue and breast cancer cell lines and its regulation by epidermal growth factor.*, Int J Cancer **80** (1999), no. 2, 250–256 (eng).

[CV95] Corinna Cortes and Vladimir Vapnik, *Support-vector networks*, Machine learning **20** (1995), no. 3, 273–297.

[CV99] Olivier Chapelle and Vladimir Vapnik, *Model selection for support vector machines.*, NIPS, 1999, pp. 230–236.

[CVBM02] O Chapelle, V Vapnik, O Bousquet, and S Mukherjee, *Choosing multiple parameters for support vector machines*, Machine Learning **46** (2002), no. 1-3, 131–159.

[CXR+11] Li Chen, Jianhua Xuan, Rebecca Riggins, Robert Clarke, and Yue Wang, *Identifying cancer biomarkers by network-constrained support vector machines*, BMC Systems Biology **5** (2011), no. 1, 161.

[DCS+10] Phuong Dao, Recep Colak, Raheleh Salari, Flavia Moser, Elai Davicioni, Alexander Schönhuth, and Martin Ester, *Inferring cancer subnetwork markers using density-constrained biclustering.*, Bioinformatics **26** (2010), no. 18, i625–i631.

[DD11] Yotam Drier and Eytan Domany, *Do two machine-learning based prognostic signatures for breast cancer capture the same biological processes?*, PLoS One **6** (2011), no. 3, e17795 (eng).

[DHS01] R. Duda, P. Hart, and D. Stork, *Pattern classification*, Wiley-Interscience, New York, 2001.

[DI11] Janusz Dutkowski and Trey Ideker, *Protein networks as logic functions in development and cancer.*, PLoS Comput Biol **7** (2011), no. 9, e1002180.

[DKR+08] Marcus T Dittrich, Gunnar W Klau, Andreas Rosenwald, Thomas Dandekar, and Tobias Müller, *Identifying functional modules in protein-protein interaction networks: an integrated exact approach.*, Bioinformatics (Oxford, England) **24** (2008), no. 13, i223–31.

[DPL+07] Christine Desmedt, Fanny Piette, Sherene Loi, Yixin Wang, Françoise Lallemand, Benjamin Haibe-Kains, Giuseppe Viale, Mauro Delorenzi, Yi Zhang, Mahasti Saghatchian d'Assignies d'Assignies d'Assignies d'Assignies d'Assignies d'Assignies, Jonas Bergh, Rosette Lidereau, Paul Ellis, Adrian L Harris, Jan G M Klijn, John A Foekens, Fatima Cardoso, Martine J Piccart, Marc Buyse, Christos Sotiriou, and T. R. A. N. S. B. I. G. Consortium, *Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series.*, Clin Cancer Res **13** (2007), no. 11, 3207–3214 (eng).

[DTvOM07] Tijl De Bie, Léon-Charles Tranchevent, Liesbeth MM van Oeffelen, and Yves Moreau, *Kernel-based data fusion for gene prioritization*, Bioinformatics **23** (2007), no. 13, i125–i132.

[DUdA06a] Ramon Diaz-Uriarte and Sara Alvarez de Andres, *Gene selection and classification of microarray data using random forest.*, BMC Bioinformatics **7** (2006), 3.

[DUDA06b] Ramón Díaz-Uriarte and Sara Alvarez De Andres, *Gene selection and classification of microarray data using random forest*, BMC bioinformatics **7** (2006), no. 1, 3.

[DWC+11] Phuong Dao, Kendric Wang, Colin Collins, Martin Ester, Anna Lapuk, and S. Cenk Sahinalp, *Optimally discriminative subnetwork markers predict response to chemotherapy.*, Bioinformatics **27** (2011), no. 13, i205–i213.

[DWWTM06] Dennise D Dalma-Weiszhausz, Janet Warrington, Eugene Y Tanimoto, and C Garrett Miyada, *The affymetrix genechip® platform: An overview*, Methods in enzymology **410** (2006), 3–28.

[DYF+03] Paul Dent, Adly Yacoub, Paul B Fisher, Michael P Hagan, and Steven Grant, *Mapk pathways in radiation responses.*, Oncogene **22** (2003), no. 37, 5885–5896 (eng).

[EDKG+05] Liat Ein-Dor, Itai Kela, Gad Getz, David Givol, and Eytan Domany, *Outcome signature genes in breast cancer: is there a unique set?*, Bioinformatics **21** (2005), no. 2, 171–178 (eng).

[EKS06] Aurora Esquela-Kerscher and Frank J Slack, *Oncomirs – microrna with a role in cancer*, Nature Reviews Cancer **6** (2006), no. 4, 259–269.

[FBB+11] Simon A Forbes, Nidhi Bindal, Sally Bamford, Charlotte Cole, Chai Yin Kok, David Beare, Mingming Jia, Rebecca Shepherd, Kenric Leung, Andrew Menzies, et al., *Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer*, Nucleic acids research **39** (2011), no. suppl 1, D945–D950.

[FCD+11] Guy Haskin Fernald, Emidio Capriotti, Roxana Daneshjou, Konrad J Karczewski, and Russ B Altman, *Bioinformatics challenges for personalized medicine*, Bioinformatics **27** (2011), no. 13, 1741–1748.

[FJP+10] Michelangelo Fiorentino, Gregory Judson, Kathryn Penney, Richard Flavin, Jennifer Stark, Christopher Fiore, Katja Fall, Neil Martin, Jing Ma, Jennifer Sinnott, Edward Giovannucci, Meir Stampfer, Howard D Sesso, Philip W Kantoff, Stephen Finn, Massimo Loda, and Lorelei Mucci, *Immunohistochemical expression of brca1 and lethal prostate cancer.*, Cancer Res **70** (2010), no. 8, 3136–3139 (eng).

[FKJ10] Kristen Fortney, Max Kotlyar, and Igor Jurisica, *Inferring the functions of longevity genes with modular subnetwork biomarkers of caenorhabditis elegans aging.*, Genome Biol **11** (2010), no. 2, R13.

[FM04] Glenn Fung and O.L. Mangasarian, *A feature selection newton method for support vector machine classification*, Computational Optimization and Applications **28** (2004), 185–202, 10.1023/B:COAP.0000026884.66338.df.

[FZ05] H. Fröhlich and A. Zell, *Efficient Parameter Selection for Support Vector Machines in Classification and Regression via Model-Based Global Optimization*, Proc. Int. Joint Conf. Neural Networks, 2005, pp. 1431 – 1438.

[GBP+08] Philip A Gregory, Andrew G Bert, Emily L Paterson, Simon C Barry, Anna Tsykin, Gelareh Farshid, Mathew A Vadas, Yeesim Khew-Goodall, and Gregory J Goodall, *The mir-200 family and mir-205 regulate epithelial to mesenchymal transition by targeting zeb1 and sip1*, Nature cell biology **10** (2008), no. 5, 593–601.

[GDCW09] Cuilan Gao, Xin Dang, Yixin Chen, and Dawn Wilkins, *Graph ranking for exploratory gene data analysis.*, BMC Bioinformatics **10 Suppl 11** (2009), S19.

[GE03] Isabelle Guyon and André Elisseeff, *An introduction to variable and feature selection*, J. Mach. Learn. Res. **3** (2003), 1157–1182.

[GFMM12] Joana P Gonçalves, Alexandre P Francisco, Yves Moreau, and Sara C Madeira, *Interactogeneous: Disease gene prioritization using heterogeneous networks and full topology scores*, PloS one **7** (2012), no. 11, e49634.

[Gin13] Geoffrey S Ginsburg, *Realizing the opportunities of genomics in health carethe opportunities of genomics in health care*, JAMA **309** (2013), no. 14, 1463–1464.

[GJSvDE08] Sam Griffiths-Jones, Harpreet Kaur Saini, Stijn van Dongen, and Anton J. Enright, *mirbase: tools for microrna genomics*, Nucleic Acids Research **36** (2008), no. suppl 1, D154–D158.

[GL13] Levi A Garraway and Eric S Lander, *Lessons from the cancer genome*, Cell **153** (2013), no. 1, 17–37.

[GM12] Ramiro Garzon and Guido Marcucci, *Potential of microrna s for cancer diagnostics, prognostication and therapy*, Current opinion in oncology **24** (2012), no. 6, 655–659.

[Goe10] J. Goeman, *L-1 penalized estimation in the cox proportional hazards model*, Biometrical Journal **52** (2010), no. 1, 70 – 84.

[Gön09] Mithat Gönen, *Statistical aspects of gene signatures and molecular targets*, Gastrointestinal cancer research: GCR **3** (2009), no. 2 Supplement 1, S19.

[GPF+11] Stephan Gade, Christine Porzelius, Maria Faelth, Jan Brase, Daniela Wuttig, Ruprecht Kuner, Harald Binder, Holger Sueltmann, and Tim Beissbarth, *Graph based fusion of mirna and mrna expression data improves clinical outcome prediction in prostate cancer*, BMC Bioinformatics **12** (2011), no. 1, 488.

[GSD+99] R. Gonzalez, J. M. Silva, G. Dominguez, J. M. Garcia, G. Martinez, J. Vargas, M. Provencio, P. España, and F. Bonilla, *Detection of loss of heterozygosity at rad51, rad52, rad54 and brca1 and brca2 loci in breast cancer: pathological correlations.*, Br J Cancer **81** (1999), no. 3, 503–509 (eng).

[GSMK+10] Norma Carmen Gutiérrez, María Eugenia Sarasquete, I Misiewicz-Krzeminska, M Delgado, J De Las Rivas, FV Ticona, E Ferminan, P Martin-Jimenez, C Chillon, A Risueno, et al., *Deregulation of microrna expression in the different genetic subtypes of multiple myeloma and correlation with gene expression profiling*, Leukemia **24** (2010), no. 3, 629–637.

[GVDV04] Jelle J Goeman, Sara A Van De Geer, Floor De Kort, and Hans C Van Houwelingen, *A global test for groups of genes: testing association with a clinical outcome*, Bioinformatics **20** (2004), no. 1, 93–99.

[GW08] Geoffrey S Ginsburg and Huntington F Willard, *Genomic and personalized medicine*, vol. 1, Academic Press, 2008.

[GW+09] Geoffrey S Ginsburg, Huntington F Willard, et al., *Genomic and personalized medicine: foundations and applications*, Translational Research-the Journal of Laboratory and Clinical Medicine **154** (2009), no. 6, 277.

[GWBV02a] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, *Gene Selection for Cancer Classification using Support Vector Machines*, Machine Learning **46** (2002), 389 – 422.

[GWBV02b] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik, *Gene selection for cancer classification using support vector machines*, Mach. Learn. **46** (2002), 389–422.

[GZL+05] Zheng Guo, Tianwen Zhang, Xia Li, Qi Wang, Jianzhen Xu, Hui Yu, Jing Zhu, Haiyun Wang, Chenguang Wang, Eric J Topol, Qing Wang, and Shaoqi Rao, *Towards precise classification of cancers based on robust gene functional expression profiles.*, BMC Bioinformatics **6** (2005), 58.

[HAA+10] Thomas J Hudson, Warwick Anderson, Axel Aretz, Anna D Barker, Cindy Bell, Rosa R Bernabé, MK Bhan, Fabien Calvo, Iiro Eerola, Daniela S Gerhard, et al., *International network of cancer genome projects*, Nature **464** (2010), no. 7291, 993–998.

[Hal99] Mark A Hall, *Correlation-based feature selection for machine learning*, Ph.D. thesis, The University of Waikato, 1999.

[HB04] Louise R Howe and Anthony M C Brown, *Wnt signaling and breast cancer.*, Cancer Biol Ther **3** (2004), no. 1, 36–41 (eng).

[HBH+10] Katharine M Hardy, Brian W Booth, Mary J C Hendrix, David S Salomon, and Luigi Strizzi, *Erbb/egf signaling and emt in mammary development and breast cancer.*, J Mammary Gland Biol Neoplasia **15** (2010), no. 2, 191–199 (eng).

[HCO06] Sepp Hochreiter, Djork-Arné Clevert, and Klaus Obermayer, *A new summarization method for Affymetrix probe level data.*, Bioinformatics (Oxford, England) **22** (2006), no. 8, 943–9.

[Hel02] Michael J Heller, *Dna microarray technology: devices, systems, and applications*, Annual review of biomedical engineering **4** (2002), no. 1, 129–153.

[HGV11] Anne-Claire Haury, Pierre Gestraud, and Jean-Philippe Vert, *The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures*, PLoS ONE **6** (2011), no. 12, e28210.

[HL03] Paul Matsudaira Harvey Lodish, Arnold Berk, *Molecular cell biology (5th revised edition edition)*, W.H. Freeman & Company, 2003.

[HTF08] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*, Springer, New York, NY, USA, 2008.

[Hud07] Clifford A Hudis, *Trastuzumab–mechanism of action and use in clinical practice*, New England Journal of Medicine **357** (2007), no. 1, 39–51.

[Hui05] Trevor Hastie Hui Zou, *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **Volume 67, Issue 2** (April 2005), 301–320.

[IGS+06] Anna V Ivshina, Joshy George, Oleg Senko, Benjamin Mow, Thomas C Putti, Johanna Smeds, Thomas Lindahl, Yudi Pawitan, Per Hall, Hans Nordgren, John E L Wong, Edison T Liu, Jonas Bergh, Vladimir A Kuznetsov, and Lance D Miller, *Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer.*, Cancer Res **66** (2006), no. 21, 10292–10301 (eng).

[JBF+10] Marc Johannes, Jan C Brase, Holger Fröhlich, Stephan Gade, Mathias Gehrmann, Maria Fälth, Holger Sültmann, and Tim Beissbarth, *Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients.*, Bioinformatics **26** (2010), no. 17, 2136–2144 (eng).

[JFSB11] Marc Johannes, Holger Fröhlich, Holger Sültmann, and Tim Beissbarth, *pathclass: an r-package for integration of pathway knowledge into support vector machines for biomarker discovery.*, Bioinformatics **27** (2011), no. 10, 1442–1443 (eng).

[KAG+08] Minoru Kanehisa, Michihiro Araki, Susumu Goto, Masahiro Hattori, Mika Hirakawa, Masumi Itoh, Toshiaki Katayama, Shuichi Kawashima, Shujiro Okuda, Toshiaki Tokimatsu, and Yoshihiro Yamanishi, *Kegg for linking genomes to life and the environment.*, Nucleic Acids Res **36** (2008), no. Database issue, D480–D484 (eng).

[KCMPK+08] Carolyn Waugh Kinkade, Mireia Castillo-Martin, Anna Puzio-Kuter, Jun Yan, Thomas H Foster, Hui Gao, Yvonne Sun, Xuesong Ouyang, William L Gerald, Carlos Cordon-Cardo, and Cory Abate-Shen, *Targeting akt/mtor and erk mapk signaling inhibits hormone-refractory prostate cancer in a preclinical mouse model.*, J Clin Invest **118** (2008), no. 9, 3051–3064 (eng).

[KL02] Risi Imre Kondor and John Lafferty, *Diffusion kernels on graphs and other discrete input spaces*, Proc. of ICML 2002, 2002.

[KL12] Maricel Kann and Fran Lewitter (eds.), *Translational bioinformatics*, PLOS Computational Biology, 2012.

[KLH+11] Kai Kammers, Michel Lang, Jan G Hengstler, Marcus Schmidt, and Jorg Rahnenfuhrer, *Survival models with preclustered gene groups as covariates.*, BMC Bioinformatics **12** (2011), no. 1, 478.

[KR90] L. Kaufman and P. Rousseeuw, *Finding groups in data: An introduction to cluster analysis*, Wiley, New York, 1990.

[KS96] Daphne Koller and Mehran Sahami, *Toward optimal feature selection*.

[LCK+08] Eunjung Lee, Han-Yu Chuang, Jong-Won Kim, Trey Ideker, and Doheon Lee, *Inferring pathway activity toward precise disease classification.*, PLoS Comput Biol **4** (2008), no. 11, e1000217 (eng).

[LFA93] Rosalind C Lee, Rhonda L Feinbaum, and Victor Ambros, *The c. elegans heterochronic gene< i> lin-4< /i> encodes small rnas with antisense complementarity to< i> lin-14< /i>*, Cell **75** (1993), no. 5, 843–854.

[LGM+05] Jun Lu, Gad Getz, Eric A Miska, Ezequiel Alvarez-Saavedra, Justin Lamb, David Peck, Alejandro Sweet-Cordero, Benjamin L Ebert, Raymond H Mak, Adolfo A Ferrando, et al., *Microrna expression profiles classify human cancers*, nature **435** (2005), no. 7043, 834–838.

[LL08] Caiyan Li and Hongzhe Li, *Network-constrained regularization and variable selection for analysis of genomic data.*, Bioinformatics **24** (2008), no. 9, 1175–1182.

[LLB+01] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, et al., *Initial sequencing and analysis of the human genome*, Nature **409** (2001), no. 6822, 860–921.

[Mar11] Elaine R Mardis, *A decade/'s perspective on dna sequencing technology*, Nature **470** (2011), no. 7333, 198–203.

[MBHG05] Julie L Morrison, Rainer Breitling, Desmond J Higham, and David R Gilbert, *Generank: using search engine technology for the analysis of microarray experiments.*, BMC Bioinformatics **6** (2005), 233 (eng).

[MH05] Shuangge Ma and Jian Huang, *Regularized roc method for disease classification and biomarker selection with microarray data*, Bioinformatics **21** (2005), no. 24, 4356–4362.

[MH08] ———, *Penalized feature selection and classification in bioinformatics*, Briefings in bioinformatics **9** (2008), no. 5, 392–403.

[Mit97] Tom M Mitchell, *Machine learning. wcb*, 1997.

[MMG13] Jeanette J. McCarthy, Howard L. McLeod, and Geoffrey S. Ginsburg, *Genomic medicine: A decade of successes, challenges, and opportunities*, Science Translational Medicine **5** (2013), no. 189, 189sr4.

[MRT12] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar, *Foundations of machine learning*, The MIT Press, 2012.

[MT12] Yves Moreau and Léon-Charles Tranchevent, *Computational tools for prioritizing candidate genes: boosting disease gene discovery*, Nature Reviews Genetics (2012).

[MV98] K. L. Marsh and J. M. Varley, *Frequent alterations of cell cycle regulators in early-stage breast lesions as detected by immunohistochemistry.*, Br J Cancer **77** (1998), no. 9, 1460–1468 (eng).

[NCI13] USNIH National Cancer Institute, *Personalized medicine, http://www.cancer.gov/ dictionary/?cdrid=561717, accessed 12 june, 2013*.

[NKP+11] Kristin M Nieman, Hilary A Kenny, Carla V Penicka, Andras Ladanyi, Rebecca Buell-Gutbrod, Marion R Zillhardt, Iris L Romero, Mark S Carey, Gordon B Mills, Gökhan S Hotamisligil, S. Diane Yamada, Marcus E Peter, Katja Gwin, and Ernst Lengyel, *Adipocytes promote ovarian cancer metastasis and provide energy for rapid tumor growth.*, Nat Med **17** (2011), no. 11, 1498–1503 (eng).

[NTT+09] Daniela Nitsch, LÃ©on-Charles Tranchevent, Bernard Thienpont, Lieven Thorrez, Hilde Van Esch, Koenraad Devriendt, and Yves Moreau, *Network analysis of differential expression for the identification of disease-causing genes.*, PLoS One **4** (2009), no. 5, e5526.

[OFH+09] John D Osborne, Jared Flatow, Michelle Holko, Simon M Lin, Warren A Kibbe, Lihua Julie Zhu, Maria I Danila, Gang Feng, and Rex L Chisholm, *Annotating the human genome with disease ontology.*, BMC Genomics **10 Suppl 1** (2009), S6 (eng).

[ONLH00] M. A. Olayioye, R. M. Neve, H. A. Lane, and N. E. Hynes, *The erbb signaling network: receptor heterodimerization in development and cancer.*, EMBO J **19** (2000), no. 13, 3159–3167 (eng).

[PBA+05] Yudi Pawitan, Judith Bjöhle, Lukas Amler, Anna-Lena Borg, Suzanne Egyhazi, Per Hall, Xia Han, Lars Holmberg, Fei Huang, Sigrid Klaar, Edison T Liu, Lance Miller, Hans Nordgren, Alexander Ploner, Kerstin Sandelin, Peter M Shaw, Johanna Smeds, Lambert Skoog, Sara Wedrén, and Jonas Bergh, *Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts.*, Breast Cancer Res **7** (2005), no. 6, R953–R964 (eng).

[PBB99] E. Pötter, C. Bergwitz, and G. Brabant, *The cadherin-catenin system: implications for growth and differentiation of endocrine tissues.*, Endocr Rev **20** (1999), no. 2, 207–239 (eng).

[PBMW99] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd, *The pagerank citation ranking: Bringing order to the web.*, Technical Report 1999-66, Stanford InfoLab, November 1999, Previous number = SIDL-WP-1999-0120.

[PCB+96] J. Papp, B. Csokay, P. Bosze, Z. Zalay, J. Toth, B. Ponder, and E. Olah, *Allele loss from large regions of chromosome 17 is common only in certain histological subtypes of ovarian carcinomas.*, Br J Cancer **74** (1996), no. 10, 1592–1597 (eng).

[PDD09] Vasyl Pihur, Susmita Datta, and Somnath Datta, *Rankaggreg, an r package for weighted rank aggregation*, BMC bioinformatics **10** (2009), no. 1, 62.

[PFCS+12] Martin Peifer, Lynnette Fernández-Cuesta, Martin L Sos, Julie George, Danila Seidel, Lawryn H Kasper, Dennis Plenker, Frauke Leenders, Ruping Sun, Thomas Zander, et al., *Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer*, Nature genetics **44** (2012), no. 10, 1104–1110.

[PKP09] T. S Keshava Prasad, Kumaran Kandasamy, and Akhilesh Pandey, *Human protein reference database and human proteinpedia as discovery tools for systems biology.*, Methods Mol Biol **577** (2009), 67–79.

[PSE+00] Charles M Perou, Therese Sørlie, Michael B Eisen, Matt van de Rijn, Stefanie S Jeffrey, Christian A Rees, Jonathan R Pollack, Douglas T Ross, Hilde Johnsen, Lars A Akslen, et al., *Molecular portraits of human breast tumours*, Nature **406** (2000), no. 6797, 747–752.

[PT00] V. Petit and J. P. Thiery, *Focal adhesions: structure and dynamics.*, Biol Cell **92** (2000), no. 7, 477–494 (eng).

[QZZC10] Yu-Qing Qiu, Shihua Zhang, Xiang-Sun Zhang, and Luonan Chen, *Detecting disease associated modules and prioritizing active genes based on high throughput data.*, BMC Bioinformatics **11** (2010), 26.

[RG02] Sridhar Ramaswamy and Todd R Golub, *Dna microarrays in clinical oncology*, Journal of Clinical Oncology **20** (2002), no. 7, 1932–1941.

[Ris01] Irina Rish, *An empirical study of the naive bayes classifier*, IJCAI-01 workshop on "Empirical Methods in AI", 2001.

[RZD+07] Franck Rapaport, Andrei Zinovyev, Marie Dutreix, Emmanuel Barillot, and Jean-Philippe Vert, *Classification of microarray data using gene networks.*, BMC Bioinformatics **8** (2007), 35 (eng).

[SBvT+08] Marcus Schmidt, Daniel Böhm, Christian von Törne, Eric Steiner, Alexander Puhl, Henryk Pilch, Hans-Anton Lehr, Jan G Hengstler, Heinz Kölbl, and Mathias Gehrmann, *The humoral immune system has a key prognostic impact in node-negative breast cancer.*, Cancer Res **68** (2008), no. 13, 5405–5413 (eng).

[SCF09] Michael R Stratton, Peter J Campbell, and P Andrew Futreal, *The cancer genome*, Nature **458** (2009), no. 7239, 719–724.

[SCK+12] Christine Staiger, Sidney Cadot, Raul Kooter, Marcus Dittrich, Tobias Müller, Gunnar W Klau, and Lodewyk FA Wessels, *A critical evaluation of network and pathway-based classifiers for outcome prediction in breast cancer*, PloS one **7** (2012), no. 4, e34796.

[SF13] Afshin Sadeghi and Holger Fröhlich, *Steiner tree methods for optimal sub-network identification: an empirical study*, BMC bioinformatics **14** (2013), no. 1, 144.

[SG09] Yijun Sun and Steve Goodison, *Optimizing molecular signatures for predicting prostate cancer recurrence*, Prostate. Jul 1; **69(10)** (2009), 1119–27.

[SIL07] Yvan Saeys, Iñaki Inza, and Pedro Larrañaga, *A review of feature selection techniques in bioinformatics*, Bioinformatics **23** (2007), no. 19, 2507–2517.

[SMS99] Edwin Southern, Kalim Mir, and Mikhail Shchepinov, *Molecular interactions on microarrays*, Nature genetics **21** (1999), 5–9.

[SOC+11] Devki Sukhtankar, Alec Okun, Anupama Chandramouli, Mark A Nelson, Todd W Vanderah, Anne E Cress, Frank Porreca, and Tamara King, *Inhibition of p38-mapk signaling pathway attenuates breast cancer induced bone pain and disease progression in a murine model of cancer-induced bone pain.*, Mol Pain **7** (2011), 81 (eng).

[SP08] Greg Shaw and David M Prowse, *Inhibition of androgen-independent prostate cancer cell growth is enhanced by combination therapy targeting hedgehog and erbb signalling.*, Cancer Cell Int **8** (2008), 3 (eng).

[SPT+01] T. Sorlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. Eystein Lonning, and A. L. Borresen-Dale, *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.*, Proc Natl Acad Sci U S A **98** (2001), no. 19, 10869–10874.

[SS02] B Schölkopf and A Smola, *Learning with kernels*, Cambridge: MIT Press. Schölkopf, B., Mika, S., Burges, C. J., P. Knirsch, K.-R. M., Rätsch, G., & Smola, A. J (2002), –2000–81.

[SSBL05] Tobias Sing, Oliver Sander, Niko Beerenwinkel, and Thomas Lengauer, *Rocr: visualizing classifier performance in r.*, Bioinformatics **21** (2005), no. 20, 3940–3941 (eng).

[STV04] Bernhard Schèolkopf, Koji Tsuda, and Jean-Philippe Vert, *Kernel methods in computational biology*, The MIT press, 2004.

[SWL+06] Christos Sotiriou, Pratyaksha Wirapati, Sherene Loi, Adrian Harris, Steve Fox, Johanna Smeds, Hans Nordgren, Pierre Farmer, Viviane Praz, Benjamin Haibe-Kains, Christine Desmedt, Denis Larsimont, Fatima Cardoso, Hans Peterse, Dimitry Nuyten, Marc Buyse, Marc J. Van de Vijver, Jonas Bergh, Martine Piccart, and Mauro Delorenzi, *Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis*, Journal of the National Cancer Institute **98** (2006), no. 4, 262–272.

[SYD10] Junjie Su, Byung-Jun Yoon, and Edward R Dougherty, *Identification of diagnostic subnetwork markers for cancer in human protein-protein interaction network.*, BMC Bioinformatics **11 Suppl 6** (2010), S8.

[TA77] A. Tikhonov and V. Arsenin, *Solutions of ill-posed problems*, W.H. Winston & Sons, Washington, 1977.

[TGA+10] Andrew E Teschendorff, Sergio Gomez, Alex Arenas, Dorraya El-Ashry, Marcus Schmidt, Mathias Gehrmann, and Carlos Caldas, *Improved prognostic classification of breast cancer defined by antagonistic activation patterns of immune response pathway modules.*, BMC Cancer **10** (2010), 604.

[The04] The Gene Ontology Consortium, *The gene ontology (GO) database and informatics resource*, Nucleic Acids Research **32** (2004), D258–D261.

[THNC02] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu, *Diagnosis of multiple cancer types by shrunken centroids of gene expression.*, Proc Natl Acad Sci U S A **99** (2002), no. 10, 6567–6572 (eng).

[Tib96] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Royal. Statist. Soc B. **58** (1996), no. 1, 267–288.

[TLWF+09] Ian W Taylor, Rune Linding, David Warde-Farley, Yongmei Liu, Catia Pesquita, Daniel Faria, Shelley Bull, Tony Pawson, Quaid Morris, and Jeffrey L Wrana, *Dynamic modularity in protein interaction networks predicts breast cancer outcome.*, Nat Biotechnol **27** (2009), no. 2, 199–204.

[TSH+10] Barry S Taylor, Nikolaus Schultz, Haley Hieronymus, Anuradha Gopalan, Yonghong Xiao, Brett S Carver, Vivek K Arora, Poorvi Kaushik, Ethan Cerami, Boris Reva, et al., *Integrative genomic profiling of human prostate cancer*, Cancer cell **18** (2010), no. 1, 11–22.

[TST+99] K. Terasawa, S. Sagae, T. Takeda, S. Ishioka, K. Kobayashi, and R. Kudo, *Telomerase activity in malignant ovarian tumors with deregulation of cell cycle regulatory proteins.*, Cancer Lett **142** (1999), no. 2, 207–217 (eng).

[TTC01] V. G. Tusher, R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response.*, Proc Natl Acad Sci U S A **98** (2001), no. 9, 5116–5121 (eng).

[VAM+01] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al., *The sequence of the human genome*, Science Signaling **291** (2001), no. 5507, 1304.

[Vap00] Vladimir Vapnik, *The nature of statistical learning theory*, 2ed ed., Springer, 2000.

[VBS+10] C. J. Vaske, S. C. Benz, J. Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler, and J. M. Stuart, *Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM*, Bioinformatics **26** (2010), no. 12, i237–i245.

[VC00] Vladimir Vapnik and Olivier Chapelle, *Bounds on error expectation for support vector machines*, Neural computation **12** (2000), no. 9, 2013–2036.

[VCKH+09] Eric Van Cutsem, Claus-Henning Köhne, Erika Hitre, Jerzy Zaluski, Chung-Rong Chang Chien, Anatoly Makhson, Geert D'Haens, Tamás Pintér, Robert Lim, György Bodoky, et al., *Cetuximab and chemotherapy as initial treatment for metastatic colorectal cancer*, New England Journal of Medicine **360** (2009), no. 14, 1408–1417.

[VDVHvV+02] Marc J Van De Vijver, Yudong D He, Laura J van't Veer, Hongyue Dai, Augustinus AM Hart, Dorien W Voskuil, George J Schreiber, Johannes L Peterse, Chris Roberts, Matthew J Marton, et al., *A gene-expression signature as a predictor of survival in breast cancer*, New England Journal of Medicine **347** (2002), no. 25, 1999–2009.

[VK04] Bert Vogelstein and Kenneth W Kinzler, *Cancer genes and the pathways they control*, Nature medicine **10** (2004), no. 8, 789–799.

[VLV+10] Ilse Van der Auwera, R Limame, P Van Dam, PB Vermeulen, LY Dirix, and SJ Van Laere, *Integrated mirna and mrna expression profiling of the inflammatory breast cancer subtype*, British journal of cancer **103** (2010), no. 4, 532–541.

[VMR+08] Jan B Vermorken, Ricard Mesia, Fernando Rivera, Eva Remenar, Andrzej Kawecki, Sylvie Rottey, Jozsef Erfan, Dmytro Zabolotnyy, Heinz-Roland Kienzer, Didier Cupissol, et al., *Platinum-based chemotherapy plus cetuximab in head and neck cancer*, New England Journal of Medicine **359** (2008), no. 11, 1116–1127.

[VPV+13] Bert Vogelstein, Nickolas Papadopoulos, Victor E Velculescu, Shibin Zhou, Luis A Diaz, and Kenneth W Kinzler, *Cancer genome landscapes*, science **339** (2013), no. 6127, 1546–1558.

[vtVDvdV+02] Laura J van 't Veer, Hongyue Dai, Marc J van de Vijver, Yudong D He, Augustinus A M Hart, Mao Mao, Hans L Peterse, Karin van der Kooy, Matthew J Marton, Anke T Witteveen, George J Schreiber, Ron M Kerkhoven, Chris Roberts, Peter S Linsley, RenÃ© Bernards, and Stephen H Friend, *Gene expression profiling predicts clinical outcome of breast cancer.*, Nature **415** (2002), no. 6871, 530–536.

[VVK+11] Urmo Võsa, Tõnu Vooder, Raivo Kolde, Krista Fischer, Kristjan Välk, Neeme Tõnisson, Retlav Roosipuu, Jaak Vilo, Andres Metspalu, and Tarmo Annilo, *Identification of mir-374a as a prognostic marker for survival in patients with early-stage nonsmall cell lung cancer*, Genes, Chromosomes and Cancer **50** (2011), no. 10, 812–822.

[Wei07] Robert Allan Weinberg, *The biology of cancer*, vol. 255, Garland Science New York, 2007.

[WKK+12] Christof Winter, Glen Kristiansen, Stephan Kersting, Janine Roy, Daniela Aust, Thomas Knösel, Petra Rümmele, Beatrix Jahnke, Vera Hentrich, Felix Rückert, et al., *Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes*, PLoS Computational Biology **8** (2012), no. 5, e1002511.

[WKZ+05] Yixin Wang, Jan G. Klijn, Yi Zhang, Anieta M. Sieuwerts, Maxime P. Look, Fei Yang, Dmitri Talantov, Mieke Timmermans, Marion E. Meijer-van Gelder, Jack Yu, Tim Jatkoe, Els M. Berns, David Atkins, and John A. Foekens, *Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer.*, Lancet **365** (2005), no. 9460, 671–679.

[WZZ08] Li Wang, Ji Zhu, and Hui Zou, *Hybrid huberized support vector machines for microarray classification and gene selection.*, Bioinformatics **24** (2008), no. 3, 412–419 (eng).

[YB05] G. W. Yardy and S. F. Brewster, *Wnt signalling and prostate cancer.*, Prostate Cancer Prostatic Dis **8** (2005), no. 2, 119–126 (eng).

[YB06] George W Yardy and Simon F Brewster, *The wnt signalling pathway is a potential therapeutic target in prostate cancer.*, BJU Int **98** (2006), no. 4, 719–721 (eng).

[YDPD12] Ruoting Yang, Bernie J Daigle, Linda R Petzold, and Francis J Doyle, *Core module biomarker identification with network exploration for breast cancer metastasis.*, BMC Bioinformatics **13** (2012), no. 1, 12.

[YL03] Lei Yu and Huan Liu, *Feature selection for high-dimensional data: A fast correlation-based filter solution*, ICML, vol. 3, 2003, pp. 856–863.

[YSZ+07] Jack X Yu, Anieta M Sieuwerts, Yi Zhang, John W M Martens, Marcel Smid, Jan G M Klijn, Yixin Wang, and John A Foekens, *Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer.*, BMC Cancer **7** (2007), 182.

[ZALP06] Hao Helen Zhang, Jeongyoun Ahn, Xiaodong Lin, and Cheolwoo Park, *Gene selection using support vector machines with non-convex penalty.*, Bioinformatics **22** (2006), no. 1, 88–95.

[ZCLS09] Song Zhang, Hu Chen, Ke Liu, and Zhirong Sun, *Inferring protein function by domain context similarities in protein-protein interaction networks.*, BMC bioinformatics **10** (2009), 395.

[ZLS+06] Xuegong Zhang, Xin Lu, Qian Shi, Xiu-Qin Xu, Hon-Chiu E Leung, Lyndsay N Harris, James D Iglehart, Alexander Miron, Jun S Liu, and Wing H Wong, *Recursive svm feature selection and sample classification for mass-spectrometry and microarray data.*, BMC Bioinformatics **7** (2006), 197.

[ZRHT04] Ji Zhu, Saharon Rosset, Trevor Hastie, and Rob Tibshirani, *1-norm support vector machines*, Advances in neural information processing systems **16** (2004), no. 1, 49–56.

[ZSP09] Yanni Zhu, Xiaotong Shen, and Wei Pan, *Network-based support vector machine for classification of microarray samples.*, BMC Bioinformatics **10 Suppl 1** (2009), S21 (eng).

[ZYK+11] Min Zhu, Ming Yi, Chang Hee Kim, Chuxia Deng, Yi Li, Daniel Medina, Robert Stephens, and Jeffrey Green, *Integrated mirna and mrna expression profiling of mouse mammary tumor models identifies mirna signatures associated with mammary tumor lineage*, Genome biology **12** (2011), no. 8, R77.

# Curriculum Vita

Yupeng Cun

## Personal Information


| | |
|---|---|
| Date of birth: | 24 September 1981 |
| Place of birth: | Heqing, P.R.China |
| Nationality: | P.R.China |
| Address: | Jonas-Cahn-Straße 14, 53115 Bonn, Germany |
| E-Mail: | yupeng.cun@gmail.com |


## Education:

| | |
|---|---|
| 9.2010-9.2013, | PhD student, B-IT Research School of University of Bonn. |
| 9.2004-7.2007, | Master degree in Physics, Yunnan University, Kunming, China. |
| 9.2000-7.2004, | Bachelor degree of Informatics, Yunnan University, Kunming, China. |

## Experiences:

| | |
|---|---|
| From 9.2013, | Postdoc at Department of Translational Genomics, University of Cologne. |
| 8.2009-7.2010, | Research assistant, Beijing Institute of Genomics, Beijing, China |
| 8.2008-7.2009, | Staff scientist, Partner Institute for Computational Biology, Shanghai, China |
| 7.2007-7.2008, | Research assistant, Kunming Institute of Zoology, Kunming, China |

# Publications

## Peer-reviewed journal papers

- **Yupeng Cun**, Holger Fröhlich (2014) netClass: An R-package for network based, integrative biomarker signature discovery. **Bioinformatics,** doi: 10.1093/bioinformatics/btu025
- **Yupeng Cun**, Holger Fröhlich (2013) Network and Data Integration for Biomarker Signature Discovery via Network Smoothed T-Statistics, **PLoS One**, dio: 10.1371/journal.pone.0073074
- **Yupeng Cun**, Holger Fröhlich (2012) Prognostic Gene Signatures for Patient Stratification in Breast Cancer - Accuracy, Stability and Interpretability of Gene Selection Approaches Using Prior Knowledge on Protein-Protein Interactions. **BMC Bioinformatics** 13:69 doi:10.1186/1471-2105-13-69
- **Yupeng Cun**, Holger Fröhlich (2012) Biomarker Gene Signature Discovery Integrating Network Knowledge, **Biology** 1, no. 1: 5-17. doi:10.3390/biology1010005

## Reseach manuscripts

- **Yupeng Cun**, Holger Fröhlich (2012) Integrating Prior Knowledge Into Prognostic Biomarker Discovery Based on Network Structure. arXiv:1212.3214

# Acknowledgements

I would like to express my thanks to:

- My advisor, Prof. Dr. Holger Fröhlich, who gave me patiently advise, encouragement and helps for me to completed my PhD projects and my dissertation. My second advisor, Prof. Dr. Armin B. Cremers, who gave me his help and funding support during my PhD study. Another two professors of my defense committee, Prof. Dr. Achim Tresch and Prof. Dr. Björn Schefflerfor, for their valuable comments and review on my thesis.
- the B-IT research school for three year's financial supporting during my PhD study and the stuff of B-IT main office: Dr. Alexandra Reitelmann, Susan Tietz, Sabine Burch, Katharina Rösel and Thomas Thiel for their helps during my studies at Bonn.
- Prof. Dr. Andreas Dress for his long time powerful supports, kind encouragements and promotation during my study.
- My colleges: Dr. Jörg Zimmermann, Paurush Praveen, Khalid Abnaof, Gloria Isabel Valderrama Bahamóndez for their helps.
- My friends: Weiwei Zhai, Yan Leng, Qi Zuo, Jiamin Zhuang, Wenming Peng, Chunfeng Yu, Wuyan Zhou, Fann Gao, Kang Yu, Ruicheng Yan and Shidong Wang for their helps to me and my family, and Beifei Zhou for proof reading on my dissertation draft and comments.
- My parents and my wife for their understanding and supporting on me. Grateful thanks to my parents for their endless encouragement and long term supporting to my studies. I would like to express my very special thanks my wife, Lixiang Duan, for her love and supporting in the past years; and my daughter, Yunxi, for her coming bring a lots happiness to my family.