Czech Technical University in Prague
Faculty of Electrical Engineering

Doctoral Thesis

July 2014

Matěj Holec

Czech Technical University in Prague
Faculty of Electrical Engineering
Department of Computer Science and Engineering

Set-level gene expression data analysis with machine learning

Doctoral Thesis

Matěj Holec

Prague, July 2014

Ph.D. Programme: Electrical Engineering and Information Technology
Branch of Study: Artificial Intelligence and Biocybernetics

Supervisor: Doc. Ing. Filip Železný, Ph.D.

Abstract

Gene expression data analysis methods that exploit formalized prior knowledge have expanded significantly in the last 10 years. This thesis explores strategies in which prior knowledge is used to support the transformation of the original gene expression data into a so called *set-level* representation, on which machine-learning algorithms are applied usually for the sake of predictive classification. The transformation generalizes the input data towards "more abstract" data dimensions that correspond to various biological functional or structural properties rather than genes. For example, the original gene expressions may be replaced by activity levels of a specific set of regulatory pathways, which are estimated from the gene expressions during the transformation.

We hypothesize that the outlined transformation has positive influence on the outcome of the subsequent machine-learning experiments due to the following intuition. Typically, analysis of gene expression data suffers from the well-known "large p, small n" problem (large data dimension and few data samples) leading to an increased risk of overfitting. The transformed set-level representation typically has a smaller dimension, thereby mitigating the "large p" part of the problem. Interestingly, the set-level approach enables us to address also the "small n" part of the problem, since it allows to merge several gene expression data sets originally using different feature sets (genes) but unified to the same more abstract units during the transformation. These two reasons for the suspected boost of the machine-learning performance represent two hypotheses which are tested as the main contribution of this dissertation work.

We successfully prove that generalization caused by the set-level techniques exploiting functional relationships among genes of prior defined gene sets allows the integration of additional data obtained by different platforms or even species, i.e., we confirm the latter hypothesis. For the former hypothesis, the situation is more subtle. We show that using standard gene sets proposed in state-of-the-art research, the performance of predictive analysis will not be significantly improved in terms of classification accuracy. However, we propose some more sophisticated definitions of gene sets which indeed lead to the improvement of classification. These new definitions based on a careful analysis of gene-regulatory principles represent another significant contribution of this thesis, albeit limited to prokaryotes.

A lateral contribution of this work is the designed evaluation framework in which numerous techniques from state-of-the-art set-level gene expression analysis can be compared in an unbiased and objective manner, from the point of view of predictive accuracy. Lastly, this thesis contributed to the development of the public web-based software tools XGENE.ORG and its successor miXGENE.

Abstrakt

V posledních deseti letech došlo k významnému rozšíření metod pro analýzu dat genové exprese využívajících apriorní znalosti. Tato práce zkoumá strategie ve kterých jsou tyto znalosti využity k podpoře transformace původních dat genové exprese do nové tzv. *množinové reprezentace*, na kterou jsou aplikovány metody strojového učení za účelem prediktivní klasifikace. Tato transformace zobecňuje vstupní data tak, že v nové “abstraktnější” reprezentaci odpovídají dimenze různým biologickým funkcím a strukturám spíše než jednotlivým genům. Například, původní genová exprese může být nahrazena aktivitami specifické množiny regulačních stezek, které jsou vypočteny na základě genové exprese během výše uvedené transformace.

Předpokládáme, že uvedená transformace má pozitivní vliv na výstup experimentů založených na aplikaci algoritmů strojového učení na transformovaných datech, a to na základě následující zkušenosti. Analýza genové exprese typicky doplácí na velmi známý problém, který spočívá nadbytku měřených atributů a nedostatku pozorování, a vede k zvýšenému riziku přeučení. Množinová reprezentace získaná pomocí zmíněné transformace má typicky menší dimenzi, tudíž umožňuje zmírnění té části problému, která spočívá v nadbytku měřených atributů. Je rovněž pozoruhodné, že tento množinový přístup umožňuje řešit i druhou část problému spočívající v nedostatku pozorování, protože umožňuje slučovat datasety reprezentované pomocí různých množin atributů (genů), které je ale možno sloučit na úrovni aktivit abstraktnějších jednotek zavedených během transformace. Tyto dva důvody pro předpokládané zvýšení výkonu metod strojového učení odpovídají dvěma hypotézám jejichž testování předkládáme jako hlavní přínos této disertační práce.

V této práci úspěšně ukazujeme, že generalizace způsobená použitím množinových transformačních technik využívajících funkcionálních vztahů mezi elementy předem stanovených množin genů umožňuje integraci dodatečných dat získaných pomocí různých platforem či dokonce živočišných druhů, čímž tedy potvrzujeme posledně uvedené hypotézu. U dříve jmenované hypotézy je situace trochu složitější. Ukazujeme, že při použití standardních genových množin, předkládaných v nejnovějších výzkumných pracích, není výkon při prediktivní analýze významně pozitivně ovlivněn, co se prediktivní přesnosti týká. Nicméně v této práci navrhujeme důmyslnější definice množin genů, které vskutku vedou k zlepšení prediktivní přesnosti. Tyto nové definice založené na důkladné analýze genově regulačních principů reprezentují další významný přínos této práce, který je nicméně omezen pouze na prokaryotické organismy.

Stranou od výše uvedených přínosů stojí navrhovaný vyhodnocovací framework pro objektivní vyhodnocení množinově orientovaných technik pro analýzu genové exprese, jenž posuzuje techniky z pohledu prediktivní přesnosti. Poslední přínos této práce je příspěvek k vývoji veřejného webového nástroje XGENE.ORG a jeho nástupce miXGENE.

Acknowledgements

I am grateful to all my colleagues, especially to Filip Železný, Jiří Kléma, Ondřej Kuželka who helped me a lot with the experiments presented in this thesis. I would also like to thank to former MSc student Valentin Gologuzov and my former colleague Jiří Bělohradský who helped with implementation of miXGENE and XGENE.ORG functionalities, respectively. Last but not least I would like to thank my family, Lenka, and Luis for their support, love, and patience.

Keywords

artificial intelligence, machine learning, supervised learning, classification, prior knowledge, feature extraction, bioinformatics, molecular genomics, gene expression, Gene ontology, metabolic and signaling pathways.

Klíčová slova

umělá inteligence, strojové učení, učení s učitelem, klasifikace, přeučení, apriorní znalosti, extrakce příznaků, bioinformatika, molekulární genomika, genová exprese, genová ontologie, metabolické a signální dráhy.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Analysis of whole-genome gene expression data has received attention in the last 15 years within some prominent publications (Golub et al., 1999; Subramanian et al., 2005; Tusher et al., 2001). The limited amount and immense biological complexity of the underlying gene expression data led to the idea of *set-level analysis*, which has also received attention in recent years (Abraham et al., 2010; Goeman and Mansmann, 2008; Hwang, 2012; Michaud et al., 2008; Mramor et al., 2010; Staiger et al., 2012; Tarca et al., 2012). This approach typically yields more compact and interpretable results than those produced by traditional methods relying on individual genes. The set-level strategy can also be adopted with similar benefits in predictive classification tasks accomplished with machine learning algorithms; however, some studies into the predictive performance of set-level classifiers have yielded rather controversial results (Abraham et al., 2010; Lee et al., 2008; Mramor et al., 2010; Staiger et al., 2012).

This thesis addresses the topic of the set-level gene expression data classification using machine-learning methods and targets the controversy over whether the available set-level approaches can improve accuracy of machine-learning models. The key idea behind the set-level techniques consists in exploiting prior knowledge in the form of predefined gene sets corresponding to various biological processes to enrich the analysis. Basically, these techniques allow the reduction of data dimensionality by transforming into an alternative data representation where the data samples are represented in "activities of the biological processes" instead of the gene activities (gene expressions). The reduction follows from the fact that the number of the predefined gene sets is typically much smaller than the number of the measured genes in the data. At the same time, the set-level methods introduce *abstraction* that substantially improves results interpretability, mainly because the transformed data samples are represented by activities of the particular cellular process (in a simplified representation using the gene sets) instead of the genes. Furthermore, the abstraction provides an interesting tool for data integration obtained from different technologies or even species. We hypothesise that the predictive accuracy of learned models will increase when the set-level approach is used due to a reduction of *overfitting*, the ability to fit random noise instead of the desired phenomenon, caused by reducing the number of *features*[1] by the data transformation. For given data and machine learning methods, predictive model performances depend on a trade off between the number of features, p, and samples, n, where reduction of the ratio $^{p}/_{n}$, by decreasing the nominator or increasing the denominator, can lead to the reduced risk of overfitting.

The main goal of this thesis is to test two hypotheses about the ratio between p and n which have a severe impact on learned model performance and interpretability, we address both hypotheses experimentally.

- The first hypothesis is related to the increase of n. We analyze the effect for various n with constant p not only when data originate from a single homogeneous data set but also when the data were integrated from different studies using transformation to the same abstract representation.

- The second hypothesis is connected to the decrease of p. Two approaches are used (including their combination): firstly it is the feature transformation from the "concrete" gene-level into the "abstract" set-level, and secondly we use traditional machine-learning methods for feature selection.

The contributions related to the first hypothesis are the following. For variable number of samples, we demonstrate the integration of gene expression data taken from different platforms and species based on the abstracted data representation. We learn the classification models on gene set representation corresponding to different abstraction levels and study whether this abstraction increases classification robustness in cross-species tasks. We also explore different ways of defining gene sets and empirically test if gene set features outperform the features based on individual genes.

The contributions related to the second hypothesis are the following. Using the state-of-the-art gene sets, we compare the performance of predictive models based on the gene-level or the abstracted representation which use both the genuine gene sets and gene sets assembled without biological relevance, we assess the state-of-the-art methods for set-level analysis in machine learning settings, and compare various aggregation methods transforming gene expression data into a new set-level features. We also assess the performance of novel gene sets based on transcriptional regulatory network topology.

The final contribution of this thesis is based on an implementation of the used methodology into online tools *xgene.org*[2] and *miXGENE*[3]. The first tool is designed primarily for the integrated analysis of gene expression data obtained by different platforms or species. The latter tool, miXGENE, provides a general framework for analysis of the genetic (gene expression) and epigenetic (DNA methylation) data .

[1] A general term for a measurable property of an object or phenomenon (e.g, level of gene transcription activity for a given gene).

[2] `http://xgene.org`

[3] `http://mixgene.felk.cvut.cz`

1.1 Thesis outline

The rest of this thesis is organized as follows. Chapter 2 provides the background of this work: (i) we briefly introduce relevant aspects of molecular biology and machine learning, and (ii) summarize the state-of-the-art methods. In Chapter 3 we state our working hypotheses. Chapter 4 contains an experimental evaluation of the first hypothesis, and Chapters 5 and 6 contain an experimental evaluation of the second hypothesis. Chapter 7 describes the tools, XGENE.ORG and miXGENE. The final Chapter, 8, summarizes the results of our experiments, concludes the whole work, and indicates possible ways to extend this thesis.

Chapter 2

Background

The aim of this chapter is to describe the biological and machine-learning background of this thesis. Firstly, we briefly introduce relevant aspects of molecular biology (Section 2.1). In the next sections, we review basic methods for gene expression data analysis (Section 2.2) and the state-of-the-art set-level analysis methods (Section 2.3). The last three sections are dedicated to the machine-learning viewpoint of gene expression analysis; Section 2.4 describes the fundamentals of machine learning, Section 2.5 introduces machine learning for gene expression analysis, and Section 2.6 presents usage of machine learning methods for the set-level analysis.

2.1 Gene expression in the cell

The complete hereditary information of any known living organism is stored in the cell as a deoxyribonucleic acid (DNA) molecule. The DNA molecule is composed of a sequence of simpler units, nucleotides, where each nucleotide contains one of four specific biological compounds[1] and can be seen as a sequence of symbols drawn from a 4-symbol alphabet. Typically, the DNA molecule is composed of two such sequences (strands) and organized as a double helix. The sequences are complementary; a symbol on one strand determines the symbol on the opposite strand. This arrangement allows an effective replication of the information during the gene multiplication or copying of the information into another information-storing molecule, ribonucleic acid (RNA), which differs from the DNA (i) by one nucleobase (using uracil instead of thymine) and (ii) typically remains single-stranded (Fig. 2.1).

The basic elements of the hereditary information are continuous sequences of DNA, *genes.* Each gene can either encode an RNA sequence which by itself plays a role in the cell, or this sequence serves as a template for synthesis of a polypeptide; an aminoacid sequence, which is done in ribosomes (Fig. 2.1). Long polypeptide sequences, *proteins*, are the essential substance of life; they constitute a wide range of different functionalities: proteins take part in different molecular functions, comprise cellular components, or participate in biological processes.

[1] The compounds (nucleobases) are guanine, adenine, thymine, and cytosine.

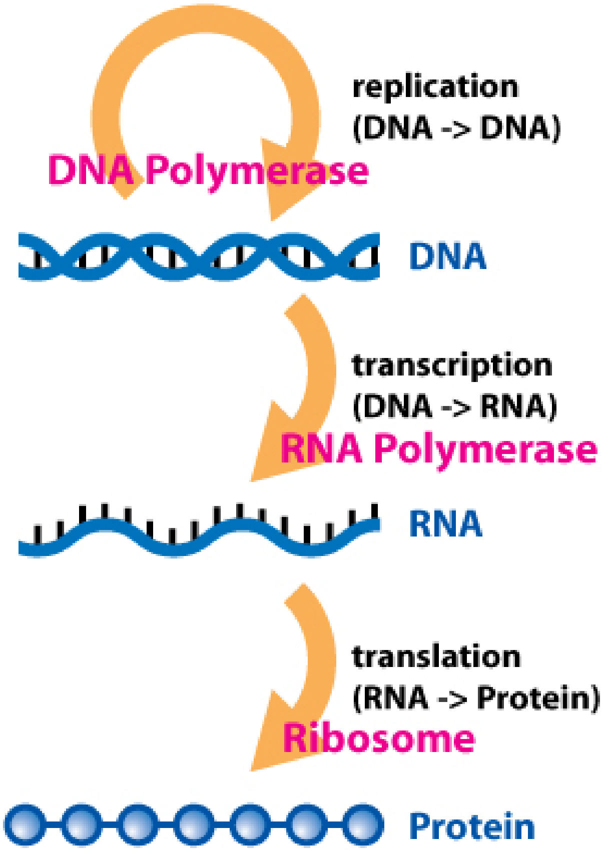


Figure 2.1: Flow of genetic information (also known as "the central dogma of molecular biology"). *Replication*, provided by the DNA polymerase, occurs during the transmission of hereditary information to the progeny of any cell or organism. *Transcription*, provided by the RNA polymerase, is a process where information stored in a sequence of DNA, gene, is transferred into an mRNA sequence. *Translation* is performed by ribosomes and consists in a protein production on the basis of the mature mRNA. (From Wikimedia Commons)

Gene expression is a general name for a process during which genes are processed into functional gene products, proteins, in the case of protein coding genes, or RNA sequences, in cases where the non-coding genes (e.g., small RNA sequences, like tRNA or microRNA, which participate in protein production or gene regulation, respectively). This process is comprised of two phases: (i) *transcription*, in this phase a copy of a particular gene (mRNA) is made according a specific part of the DNA sequence by a special enzyme, RNA polymerase, and (ii) *translation*, which consists of building a protein according to the mRNA sequence in special protein complexes, ribosomes. Generally, the genetic information in the cell always flows from the DNA to the RNA to the protein (Fig. 2.1).[2]

[2] In special cases (e.g., induced by viruses such as HIV) the flow can be from the RNA to the DNA or from the RNA to the RNA.

Gene expression regulation in prokaryotes

Simple prokaryotic organisms (e.g., bacteria) have a relatively simple cellular structure in comparison to higher organisms. The main difference consists in a missing membrane separating the place which contains most of the cellular genetic material, *nucleus*, from other cellular parts. In the prokaryotic DNA, genes are organized in *operons*, clusters with associated regulatory elements and transcribed as a single unit. Each operon contains one or more DNA regions where the RNA polymerase begins with the transcription (*promoter regions*) and a *terminator region* where the RNA polymerase releases the DNA and, therefore, finishes the transcription. Transcription rate is controlled by proteins which act as activators or suppressors of transcription (*transcription factors*). These proteins bind near the promoter region (as a response to external or internal cell stimuli) and positively or negatively affect the transcription process. Thanks to the simple architecture of the prokaryotic cell, translation starts as soon as the mRNA is free, so the translation rate is proportional to the transcription rate unless it is decreased by some functional proteins (enzymes) which can affect the translation rate of the ribosomes.

Gene expression regulation in eukaryotes

The eukaryotes typically constitute multi-cell organisms with a complicated cellular structure where the cell nucleus is clearly separated from the other cellular compartments. Regulation of the gene expression occurs primarily at three distinct levels. (i) At the *transcriptional level*—in contrast to prokaryotes—structures like the operons are missing and an additional regulation is provided by changes in the DNA by mechanisms changing spatial structure of the DNA called histone acetylation and DNA methylation. (ii) *Post-transcriptional modifications* lead to production of different mRNA sequences from a single transcribed gene which subsequently lead to the production of different proteins. It is either caused by trimming the non-coding regions of the gene and alternative assignment of the remaining sequences (*alternative splicing*) or rarely by direct modifications of certain transcript bases (*RNA editing*). (iii) At the *translational level*, the regulatory mechanisms operate with several aspects of mRNA sequences as a response to changing cellular requirements (e.g., *mRNA stability*). It is worth noting that mRNA stability of prokaryotes is only a few minutes, but in the eukaryotic cell it can stretch from dozens of minutes to a day. There are also post-translation control mechanisms which ensure persistence of proteins which can be from minutes to weeks for eukaryotes.

Organization of gene interactions in pathways

From the genomic point of view, genes constitute networks or *pathways* of complex dynamic interactions providing various cellular functions (e.g., cell signaling or metabolism) in order to keep a cell in homeostasis, represented as stable molecular attractor states towards which individual cells are drawn over time (MacArthur et al., 2009). The interactions are basically indirect; particularly, they are provided by gene-encoded mRNAs and proteins and by other substances in the cell (e.g., intermediate products of the cellular metabolism). Various databases (e.g, Croft et al., 2014; Kanehisa et al., 2004) provide such pathways which cover processes like metabolism, information processing, and diseases (Kanehisa et al., 2004).

Functional annotation of genes and gene products

Controlled vocabularies with gene annotations are typical instances of knowledge of gene and protein roles in the cell. One of the biggest examples of these vocabularies is represented by the Gene ontology (GO) project (Ashburner et al., 2000) which provide three ontologies describing genes and the products, places, or processes where they affect: (i) *cellular component*, which refers to places in a cell where a product is active, (ii) *molecular function*, that defines the biochemical activity of gene products and (iii) *biological process* referring to a biological objective to which a gene product contributes.

2.2 Gene expression data analysis

The advent of technologies capable of measuring gene expression on an entire genome level have brought new challenges to the analysis of gene expression data (Zhang, 2006).

2.2.1 High-throughput technologies

High-throughput technologies, like *microarrays* (Lipshutz et al., 1999) and *next-generation sequencing* (Wang et al., 2009), bring a relatively cheap way to analyze the gene expression. These technologies provide a complete gene expression snapshot of cells where the number of interrogated genes may vary according the analyzed organism from a few thousand for bacteria to tens of thousands for higher organisms[3]. The area of applicability of these technologies ranges across molecular-level-based disease diagnostic, drug response analysis, and other areas of molecular biology.

DNA microarrays are a technology measuring the presence of mRNA transcripts by a binding[4] of the transcripts on probes corresponding to interrogated genes (Fig. 2.2). Not in all cases does one probe measure the presence of one mRNA transcript; for some genes several probes measure the same gene mRNA transcript, and—on the contrary—some probes can bind mRNAs from two or more genes. Physically, microarrays are small solid chips with probes attached to their surface. Each probe occurs in many copies (probe set) in order to increase stability of the results. As the first step of a microarray experiment, examined sample RNA sequences are labeled by a florescent tag, then the sequences are hybridized on the chip, washed away, and scanned. The result is an image with intensities for each spot (probe) where the

[3] Human genome contains approximately 20,000 protein-coding genes.

[4] The process of binding, *hybridization*, establishes sequence specific interaction between a mRNA transcript and a probe.

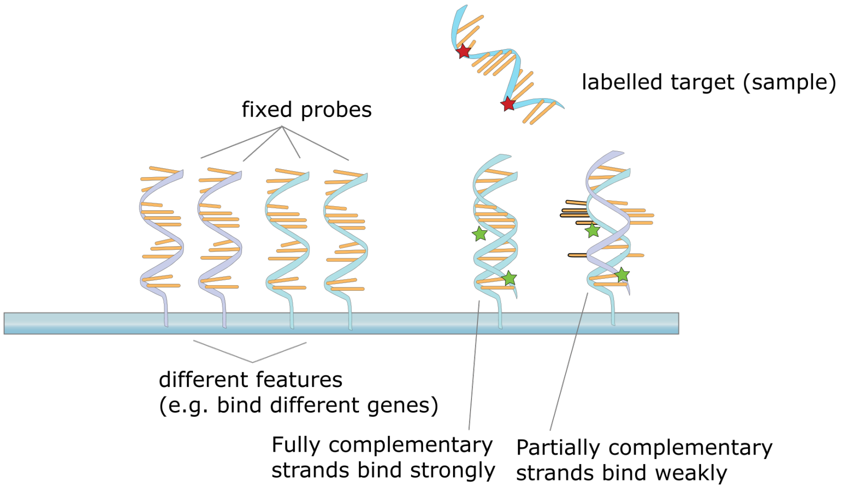


Figure 2.2: Hybridization of target samples and probes in a microarray. The whole process is the following: (i) the available samples are purified, (ii) labeled by a fluorescent label, (iii) then follows the hybridization and washing to remove weakly bound samples, and (iv) scanning. (From Wikimedia Commons)

strength of the signal from a probe depends upon the quantity of the target sample (the transcription rate). After normalization (e.g., Bolstad, 2004), the intensities are ready for analysis.

The next-generation sequencing methods measure the available mRNA transcripts directly in a quantitative manner; therefore, they are able to determine the RNA expression rate more accurately than microarrays (Wang et al., 2009).

In the rest of the thesis only single-channel *Affymetrix GeneChip* microarray data are considered. This is solely due to their availablity and for simplicity of the experimental pipeline. Performing the same experiments on data obtained by comparable platforms or the next-generation sequencing should bring similar or even more significant result when more accurate methods are used. We note here that we use the terms “genechip” and “microarray” interchangeably from now on.

2.2.2 Gene level analysis of high-throughput data

Analysis of the high-throughput gene expression data is a challenging task. The main challenge resides in the size of the processed data, inherent noise, and complicated interactions among genes. Typically, we face data which contain a relatively huge number of measured genes and a small number of interrogated samples. The typical datasets contain measurements of thousands (or even tens of thousands) of genes, g, while the number of samples, n, does not exceed hundreds and is very often only a few dozen; thus, the ratio $^p/_n$ is about 100–1000.

The microarray experiment provides a matrix of gene expressions. Here we assume that the matrix columns correspond to individual observations and each observation is a vector of real values representing expressions of the interrogated genes. In this thesis, we presume that the samples pertain into one of two sample classes (e.g., healthy and diseased samples).

A basic approach for the analysis of the data is to find a list of differentially expressed genes which exhibit a strong relationship between the gene expression and the response variable. One can use fold change, statistical tests like the t-test or its non-parametric equivalents (Wilcoxon's signed-rank test and rank-sum test). As results, the methods provide a list of differentially expressed genes and test scores. The genes with extreme scores (small-enough p-values) are declared as significant. Unfortunately, an inherent problem of this approach is caused by noise in the data and the simultaneous statistical inference on the complete set of genes which lead to incorrectly rejecting the null hypotheses and an overoptimistic conclusion about achieved significance (the multiple testing problem). Therefore, some methods use more appropriate methods to assess the significance, namely they implement multiple testing criterion (FWER) or the false discovery rate (FDR) (Benjamini and Hochberg, 1995) control procedures. (For a detailed overview of these methods see, e.g., Zhang, 2006, Ch. 4).

Another major challenge is to interpret the results obtained from the lists provided by the above mentioned methods. The challenge consists in difficult interpretability of the list in a biological context due to its length and the fact that genes can play a role in many different processes in the cell. Interpretation of such a list can be easier if genes from a gene list exhibit similarities in their functional or chromosomal location (Goeman and Bühlmann, 2007). The *overrepresentation analysis* (ORA) provides such a tool which allows biological terms significantly covered by the differentially expressed genes to be found. The general idea of the ORA is to look for an abnormal representation of significantly expressed genes in a particular biological term using, e.g., the χ^2 test or the hypergeometric test (Goeman and Bühlmann, 2007). A typical choice for the biological terms are the pathways or the GO terms represented as sets of genes.

The main limitations of the ORA approach, according to Khatri et al. (2012), are the following: (i) ORA ignores information about the strength of how the genes are differentially expressed. (ii) Genes marked as not significant are removed from the interpretation; which means that informative (but non-significant) genes are removed from the subsequent analysis. (iii) This approach (e.g., the hypergeometric or the χ^2 tests) assumes independence among the genes. (iv) ORA also assumes independence between the gene sets, which is certainly not true.

The main advantage brought by the set-level analysis is compactness and improved interpretability of the analysis results due to the smaller number of the set-level units in comparison with the number of genes. Indeed, the long lists of differentially expressed genes are replaced by shorter lists of more informative units corresponding to actual biological processes.

2.3 Set-level gene expression analysis

More advanced methods can consider insignificant-but-coordinated changes in the gene sets. Sometimes these methods are called *functional class scoring methods* (FCS) (Khatri et al., 2012). The biological utility of this approach was demonstrated by a study (Mootha et al., 2003) in which a significantly downregulated pathway-based gene set was discovered in type 2 diabetes data despite no significant expression change being detected for any individual gene; note that ORA methods cannot detect any significantly overrepresented gene set in such data.

There are a plethora of FCS methods for the set-level analysis, but all of them follow a relatively simple structure. According to Ackermann and Strimmer (2009), the FCS methods use two general alternative approaches only. In the first approach, the methods compute gene-level statistics (e.g., the t-test), establish a rank transformation of the gene scores (e.g., p-value), compute statistics for the gene sets using the computed rank, and assess their significance (e.g., Mootha et al., 2003, by the Kolmogorov-Smirnov statistics). In the second approach, the methods compute the set-level statistics and significance directly (e.g., Goeman et al., 2004). Results provided by FCS methods can vary significantly depending on methods used for rank transformation, gene set significance assessment, and—most importantly—evaluated null hypothesis (Ackermann and Strimmer, 2009; Goeman et al., 2004).

2.3.1 Current approaches

Several papers consider a performance comparison of the FCS set-level methods in the recent years (Dinu et al., 2008; Liu et al., 2007b; Song and Black, 2008; Tarca et al., 2013). This task is inherently challenging due to the absence of gold standard benchmark data to evaluate the methods; therefore, the key idea behind the evaluation in the mentioned papers is based on an assessment of an expected or known performance on real or simulated datasets. Particularly, for the real data sets, there are usually known gene sets associated with the phenotype. Liu et al. (2007b) compare Global test, ANCOVA Global test, and SAM-GS on simulated data and conclude that all methods have a similar statistical power. Dinu et al. (2008) evaluate methods on real data in which differentially expressed gene sets are predictable biologically from the phenotype. According to their evaluation, SAM-GS, Global test, and ANCOVA Global test perform better than GSEA (Subramanian et al., 2005), sigPATHWAY (Tian et al., 2005), and PLAGE (Tomfohr et al., 2005). Song and Black (2008) assess five methods on real and simulated data and conclude that GSEA[5], Global test, and PCOT2 (Song and Black, 2006) perform similarly, but better than SAFE (Barry et al., 2005), GSEA[6], and sigPATHWAY. The most recent paper by Tarca et al. (2013) compares 16 methods on 42 real microarray data sets with the following best performing methods: Global test, PLAGE, PADOG (Tarca et al., 2012).

[5] Improved implementation of GSEA with different significance assessment procedure.

[6] Standard implementation of GSEA by Subramanian et al. (2005).

Global test was among the best performing methods in all four papers, and conversely, GSEA method, despite its wide popularity among biologists, performed poorly in the all the aforementioned papers which considered this method.

2.3.2 Limitations and alternative approaches

The ORA and FCS methods are primarily limited by considering no dependencies among used gene sets.[7] The methods are also very restricted by how they use the available knowledge since only associations of biological terms with the gene sets are exploited during the analysis. However, some other approaches proceed beyond the simple set-level approach and consider known interactions among genes in pathways and transcriptional regulatory networks (e.g., Alexeyenko et al., 2012; Geistlinger et al., 2011; Rahnenführer et al., 2004; Tarca et al., 2009).

Obviously, the set-level methods give absolutely the same results when a change in topology description occurs, in comparison to net-level methods (which consider the interactions). The set-level methods ignore the differing importance of genes emerging from their place in pathways (placed downstream or upstream in signal transducting cascades). Including but not limited to the following issues, Khatri et al. (2012) point out a low resolution of knowledge bases, their fragmentation, incomplete and inaccurate annotations, missing condition and cell specific information (i.e., experiment conditions, cell type), and the fact that no "existing approach can collectively model and analyze the high-throughput data as a single dynamic system." Other issues arise from the limited quality of the available data. All these deficiencies lead to a limited use of the available information about the gene interactions and make an effective analysis challenging.

2.4 Machine learning fundamentals

Machine learning deals with artificial systems learning from given data in an autonomous manner. Regarding the available mass of gene expression data with their amount and extent ranging beyond the capacity of any living organism, machine learning proposes an interesting alternative for data analysis and knowledge discovery. In this thesis, we focus on a subset of the machine learning framework—supervised machine learning. From the molecular biology viewpoint, it provides a form of data analysis going beyond the mere identification of differentially expressed genes or gene sets; particularly, it provides tools designed to solve the problem of inferring a function from data samples and their labels automatically.

Given input and output random variables X and Y, respectively, the task is defined as fitting a model, $Y = f(X) + \varepsilon$, where Y denotes quantitative or qualitative output, for regression or classification problems, respectively, X is a p-dimensional vector (*feature vector*) of components X_j (*features*), and ε represents noise in the data (this term is typically not used for qualitative output).[8] The goal of supervised

[7] There are a few exceptions (e.g., Goeman and Mansmann, 2008; Tarca et al., 2012).

[8] Definition of the supervised learning problem in this section follows Hastie et al. (2001).

learning is to find a useful approximation of the model function, $\hat{f}$, by learning from examples. For given n observations of an analyzed system (e.g, gene expression data from healthy and diseased patients), we extract inputs and outputs of the observations and assemble a training data set $\mathcal{T} = \{(x_i \in X, y_i \in Y)|i = 1, \ldots, n\}$, and feed a learning algorithm.

The learning process can be seen as a search or optimization problem. Let us assume that there is an unknown joint probability distribution $\Pr(X, Y)$ from which the training examples were drawn, the criterion for choosing f is *expected prediction error* (EPE),

$$\mathrm{EPE}(f) = \mathrm{E}\left(\mathrm{L}(Y, f(X)\right),$$

where $L(Y, f(X))$ defines a loss function for penalizing errors. In what follows we assume that the output variable Y is qualitative, the loose $L(y, y') = 1$ whenever $y \neq y'$ and zero elsewhere, and we call function f a classifier and its EPE a *classification error*. The EPE usually cannot be computed directly, since $\Pr(X, Y)$ is typically not known, but it can be estimated empirically on a given data set S as

$$\mathrm{EPE}_S(f) = \frac{1}{|S|} \sum_{(\mathbf{x},y)\in S} \mathrm{L}(y, f(\mathbf{x})).$$

The complexity of the classifier (i.e., degree if f is a polynomial function) determines its classification error on used training data set (*training error*) and also impacts its generalization (classification error on data used for the evaluation that is *testing error*). An ideal classifier provides low error on training and testing data sets. If the classifier is *too complex* (flexible enough to fit the all training data very well), then the training error is close to zero, but its performance on testing data is poor (e.g., fits noise). If the classifier is *too simple*, then both training and testing error are too large. This complexity issue is called the *bias-variance trade-off*; because, one seeks simultaneously to minimize error over different training data sets (*bias*) and sensitivity of the classifier to small changes in the training data set (*variance*). Two extreme types of classifier misbehavior can occur in given data. The first one (*underfitting*) occurs when the model is not able to fit the data due to its low complexity. The second one (*overfitting*) is caused when the fitted model is too complex and it is not able to generalize beyond the training data.

Other factors can also affect overfitting (and underfitting) of the classifier, two of them are particularly important to the gene expression data: (i) Each classifier needs an adequate number of training samples for its learning phase; in the case of their insufficient number, the learning process can lead to overfitting. Therefore, a high complex "true" function f can be learned only when enough of the training samples are available. (ii) High dimensional input space of X (possible with many irrelevant features) can confuse the learning algorithm and lead to a higher variance of the predicted error which implies the overfitting. In practice, a method for the feature reduction of the input space can be used.

2.5 Machine learning for gene expression analysis

Machine learning techniques have been explored since very early studies on microarray data analysis (Eisen et al., 1998; Golub et al., 1999). Especially successful were supervised methods (for class prediction or regression) and unsupervised algorithms (for clustering). Gene expression data combined with machine learning methods revolutionized cancer classification which had been based solely on morphological appearance. An important milestone was a successful demonstration of cancer classification based solely on high-throughput gene expression data (Golub et al., 1999). Golub et al. (1999) used class discovery and class prediction techniques on *acute myeloid leukemia* and *acute lymphoblastic leukemia* microarray data in order to distinguish between the two cancer types using the data without any additional knowledge and to derive a class predictor able to determine the leukemia class for a new unseen case, respectively. While the clustering on samples in the above mentioned case was used, the clustering on gene level also provides an important insight in analyzed gene expression data; these algorithms have manifested the ability to find groups of co-expressed genes with similar functions which makes the clustering algorithms simple but useful tools for gaining leads to gene functions with missing or unavailable functional description (Eisen et al., 1998).

There is no one-size-fits-all supervised approach. Some methods are particularly suitable for high dimensional problems when $p \gg n$ (Hastie et al., 2001, Ch. 18) where an appropriate algorithm can quickly provide good models despite the low number of samples, abundance of correlated features, and biological or technical noise in data providing platforms (e.g., gene expression microarrays). Other methods can still be considered due to their properties, e.g, a natural way to include biological knowledge which not only improves classifiers interpretability but also can positively affect the bias-variance tradeoff. Generally, forms of the learned classifiers can range from (fast learning and less interpretable) geometrically conceived models such as *Support Vector Machines* (Cortes and Vapnik, 1995), which have been especially popular in the gene expression domain, to (slower learning and easily interpretable) symbolic models such as logical rules or decision trees that have also been applied in this area (Gamberger et al., 2004; Huang et al., 2010; Zintzaras and Kowald, 2010).

A wide range of different approaches implementing various learning models have been proposed for analysis; therefore, Allison et al. (2006) point out that *the need for thoroughly evaluating existing techniques currently seems to outweigh the need to develop new techniques*. The main reason is the absence of a gold standard evaluation technique, which is nearly impossible to resolve, so either simulated or real data with known results are used for evaluation (refer to the evaluation of the state-of-the-art gene-set-based analysis methods in Section 2.3.1). Such solutions are unfortunately difficult and prone to overoptimistic findings (Ioannidis, 2005; Jelizarow et al., 2010). An advantage of the supervised machine learning approach, in comparison to the set-level enrichment methods in Section 2.3.1, consists in a natural way to estimate the performance of learned models which can be implemented, e.g., by the k-fold cross-validation (Hastie et al., 2001, Ch. 7). The only drawback of this approach is the need for a reasonable number of data samples.

Other machine learning approaches have also been used, including association rules mining (a method formerly invented for business transaction data), time series analysis methods, and semi-supervised clustering (Zhang, 2006).

2.6 Machine learning with the set-level approach

The combination of set-level techniques with predictive classification has been suggested (Chen et al., 2008; Liu et al., 2007b; Tomfohr et al., 2005) or applied in specific ways (Bild et al., 2006; Guo et al., 2005; Holec et al., 2009b; Lee et al., 2008; Wong et al., 2008) in the previous studies; however, a focused exploration of the strategy has commenced only recently (Abraham et al., 2010; Hwang, 2012; Mramor et al., 2010; Staiger et al., 2012, 2013).

The set-level framework is adopted in predictive classification as follows. Sample features originally bearing (normalized) expressions of individual genes are replaced by features corresponding to gene sets. Each novel feature aggregates the expressions of genes contained in the corresponding gene set into a single real value; in the simplest case, it may be the average expression of the contained genes. The expression samples are then presented to the learning algorithm in terms of these derived set-level features. Informally, classifiers learned using the set-level features acquire forms such as "predict cancer if pathway P1 is active and pathway P2 is not" (where *activity* refers to an *activity score* computed from expressions of pathway member genes). In contrast, classifiers learned in the standard (gene-level) settings derive predictions from expressions of individual genes where it is usually difficult to find relationships among the genes involved in such models and to interpret them in terms of biological processes. Further motivation for extending the set-level framework to the machine learning (besides the increased interpretability already mentioned) is the possibility to compare learned models straightforwardly on predictive performance (Demšar, 2006). In contrast to the machine learning approach, it is not clear if significant results found by the ORA and FCS methods imply good predictive performance (Abraham et al., 2010).

The main issue of the set-level transformation through aggregation is that the lifting features to the set-level incurs a significant compression of the training data since the number of considered gene sets is typically much smaller than the number of interrogated genes. On the other hand, reducing the number of sample features may mitigate the risk of overfitting and thus, conversely, contribute to higher accuracy. In machine learning terms, introduced in Section 2.4, reformulation of data samples through set-level features increases the *bias* and decreases the *variance* of the learning process (Hastie et al., 2001). Another aspect of transforming features to the set-level is that the prior biological knowledge is channeled into learning through the prior definitions of biologically plausible gene sets.

Method	**References**
Mean (median)	Azuaje et al. (2010); Guo et al. (2005); Liu et al. (2007a); Abraham et al. (2010)
U-statistics	Abraham et al. (2010)
SPCA	Chen et al. (2008),
PCA	Bild et al. (2006); Levine et al. (2006); Liu et al. (2007a); Tomfohr et al. (2005)
CORG	Lee et al. (2008)
PLS	Liu et al. (2007a)
ASSESS	Edelman et al. (2006)
SetSig	Mramor et al. (2010)
Other methods	Breslin et al. (2005); Chuang et al. (2007); Efroni et al. (2007); Rapaport et al. (2007); Taylor et al. (2009)

Table 2.1: Overview of set-level methods applicable in the machine learning settings.

2.6.1 Methods overview

Methods applicable for the gene set activity score computing (set-level aggregation) are summarized in Table 2.1. All of them transform input data into space of new set-level features. Generally, these methods use similar or exactly the same aggregation mechanism, like the FCS methods in Section 2.3.1(e.g., 1st component from the PCA can be used in both approaches). Consequently, these methods share similar problems; they do not consider dependencies among gene sets and interactions among genes.

Some work (Breslin et al., 2005; Chuang et al., 2007; Efroni et al., 2007; Rapaport et al., 2007; Taylor et al., 2009) suggests methods which exploit topology information of the pathways in order to estimate the activity level, but we recall here that these (*net-level*) methods are not considered in this work.

Several papers have analyzed properties of the set-level approach in machine learning settings in recent years (Abraham et al., 2010; Hwang, 2012; Mramor et al., 2010; Staiger et al., 2012, 2013). All the papers (except Hwang, 2012) deal with cancer data and agree on the conclusion that the gene set approach, generally, do not improve the predictive accuracy in comparison to the gene-level-based alternative, but provides rather competitive results. Short descriptions and references to the available methods are in Table 2.1. These studies perform the evaluation mainly on pathways and GO terms (see Section 2.1) obtained from Molecular Signatures Database (MSigDB) (Subramanian et al., 2005) which contains collections of commonly used gene sets for the set-level analysis. MSigDB contains the following gene set collections: *positional gene sets* defined by genes from the same genomic location (C1), *curated gene sets*, e.g., KEGG pathways (Kanehisa et al., 2004) (C2) , *motif gene sets* which represent conserved regulatory elements (C3), *computational gene sets* which were mined from cancer related data (C4), and GO gene sets corresponding to the GO terms.

Abraham et al. (2010) analyze the predictive performance and stability of several set-level aggregation methods on a set of five cancer-related datasets using different gene set collections (MSigDB gene sets C1, C2, C3, C4, and C5). They evaluate three general approaches. (i) An approach where gene expressions of each sample are transformed into one value by simple statistics (mean, median, medoid[9], and a modification of the mean approach based on the t-statistics). (ii) In the second approach, the first principal component (PC) obtained by the PCA represents the aggregated gene expressions for a give gene set. (iii) The last approach is based on a mean rank comparison of genes inside and outside a gene set, this approach computes activity scores using Wilcoxon's rank sum statistics (U-statistics). For classification, they use the centroid classifier with a feature selection strategy based on the highest absolute centroid weight, and validate the results inside each dataset (internal validation) and between each of the other datasets (external validation). The main findings of this experiment are: (i) almost the same predictive accuracy of the set-level approaches as a baseline experiment; particularly, some of the aggregation methods (mean, median, medoid, and set t-statistic) showed similar performance to the baseline based on individual genes, but the first PC and the U-statistic showed statistically significant performance reductions, (ii) more consistent rankings of features within each dataset, and (iii) more stable classifiers across different datasets.

The paper by Mramor et al. (2010) provides a large comparison of set-level methods on 30 cancer datasets using gene sets based on the MSigDB gene sets (C2 and a joint collection of C1 and C5). In comparison with the previous paper, they do the internal validation only, implement no feature selection, and compare the algorithms in a correct unbiased way. For models learning, they use support vector machines, logistic regression, and k-nearest neighbors. Mramor et al. (2010) compare six aggregation methods (SetSig, CORGs, mean, PCA, median, and Assess) with a baseline approach without the set-level aggregation, and use Friedman's non-parametric statistics for models performance comparison, an approach proposed by Demšar (2006). Mramor et al. (2010) conclude that their method, SetSig, performs better than the other method, but worse than the gene-level-based (baseline) approach.

Hwang (2012) analyzes six set-level methods on seven pairs of datasets (each pair of datasets share the same phenotype for performing the external validation) on the KEGG pathways. His experiment compares the following aggregation methods: mean, CORG, ASSESS, PCA, PLS, mean of top 50% of genes[10]. For classification, the author uses SVM with the radial basis function as the kernel for the model learning, and a ranking based on the t-test for the feature selection step. The main paper result is a ranking of the aggregation methods. The best performing approaches are ASSESS and mean top 50%, and on the other side, the worst performing method is mean (mainly for the external validation).

The papers by Staiger et al. (2012, 2013) provide another evaluation of the set- and net level approaches on C2 gene set collection from MSigDB. In the first paper

[9] A representant gene with minimal Euclidean distance to the mean is selected for each sample in this implementation.

[10] A modification of the mean approach where 50% of genes with the highest ranking, according the t-test in each gene set, are used.

(Staiger et al., 2012), the authors perform an evaluation on six breast cancer datasets and aggregation methods by Chuang et al. (2007) and Taylor et al. (2009) for the net-based and CORG for the set-based knowledge. Similar to the previous studies, the authors use the t-test based feature selection and the nearest mean classifier, logistic regression, and 3-nearest neighbors for the machine learning modeling. In addition to the above mentioned papers, the authors inspect a case when the gene set knowledge is randomized. The authors also come to the conclusion that (i) the set-level methods do not outperform the gene-level approach, (ii) the randomization does not result in a performance decrease, and (ii)—contrary to the paper by Abraham et al. (2010)—the stability of the gene-level-based and set-level-based classifiers is similar when a proper correction is performed. The second paper (Staiger et al., 2013) provides a comprehensive framework for the set-level approaches evaluation based on the experimental workflow defined in Staiger et al. (2012), and differs from its predecessor by using a new single large integrated dataset composed of 12 studies (containing data from 1600 patients), and implements the other three network-based methods.

Despite several papers asserting the ability to gain predictive performance by switching to the set-level (see, e.g, discussion in Mramor et al., 2010), the above mentioned studies performed on a large collection of data sets clearly show that none of the mentioned aggregation methods combined the with machine learning can outperform the baseline approach which does not include the set-level aggregation and learn models on the single-gene-based data representation.

Chapter 3

Main hypotheses to test

In this chapter, we define the two central hypotheses of this thesis; we denote them *Hypothesis 1* and *Hypothesis 2* (or shortly H1 and H2, respectively). Both of them are concerned with performance evaluation of the combination of the gene set activity score methods and the machine learning algorithms with dependence on a changing number of features, p, and samples, n, used during the learning phase.

The number of available samples affect the optimal choice of a machine learning algorithm where an insufficient number of the training examples can lead to overfitting (Section 2.4). Unfortunately, due to the nature of gene expression data and the demanding nature of biological experiments, where a limited number of samples is very often available only, the risk of overfitting is high; therefore, simple (e.g., linear or highly regularized) machine learning models are recommended (Hastie et al., 2001). In relation to overfitting, it is an interesting fact that several studies on breast cancer diagnosis provided different sets of signature genes with little or no overlap among them mainly due to a large number of genes slightly correlated with the data phenotype and a strong fluctuation of the correlation of genes when gene expression data were measured on a different subset of patients (Ein-Dor et al., 2005).

The gene set activity score methods implicitly introduce two key data aspects which directly affect the risk of overfitting. The first aspect is *switching to more abstract set-level features* reflecting, i.e., complex processes or activity genes residing near each other on a chromosome. The second aspect is *reduction of the number of features* which in the set-level representation is typically much smaller than the number of genes or probesets.

Here we describe both hypotheses in an abstract way, a more rigorous definition is available in the respective chapters concerning the evaluation of the hypotheses.

3.1 H1—heterogeneous data integration through abstraction

Physicians and scientists analyzing the gene expression data deal with the low number of samples mainly due to the high cost of gene expression profiling, a low number of available tissue samples in some cases (e.g., when only a limited number of patients are in the study), and demanding experimental pipelines. On the other hand, public repositories often provide data originating from different labs, platforms, or even species related to the phenotype of interest and, therefore, give us an opportunity to perform a meta-analysis integrating different studies.

Here we consider the case when additional observations are introduced into an experiment; particularly, when data from different species are integrated. The observations from different platforms or species vary in the type of measured features, we recall that different platforms use different probesets. The abstraction, introduced by the set-level aggregation, provides a means how to unify observations originated from different studies by switching from organism or platform specific features to a common (organism or platform independent) representation. In cases where data come from different studies on the same organism, the more abstract ("unifying") features correspond at least to genes. For cross-species analysis, the generalization can be acquired by switching to features representing evolutionary conserved structures and functions (evolutionary conserved genes or gene sets representing the same biological function); therefore, the new features are activities of pathways or gene sets, for example.

We hypothesize that adding the related samples can help to increase predictive performance. The most important question is whether the combination of the gene set aggregation and machine learning techniques have a such ability to integrate data from multiple species with a positive effect on predictive performance. The other positive effect undoubtedly consists in (i) a possibility to mitigate overfitting due to an increased number of samples (see Section 2.4), (ii) reduction of coincidental feature-to-phenotype correlations because of mixing different subsets of patients measured by different labs, and (iii) potential ability to generalize knowledge beyond a single species. On the other hand, a combination of irrelevant or seemingly relevant samples can make it impossible to learn a desired meta-model and even relevant, but biased (e.g., not properly normalized) samples can also compromise the meta-analysis. We evaluate this hypothesis in Chapter 4.

3.2 H2—performance of models learned from the abstracted data representation

The second hypothesis is connected to the decrease of the data dimensionality by reducing the number of features, p, caused primarily by the data transformation to the abstracted set-level representation, but also by selection of the most informative set-level features.

The decrease is important from both the machine learning and biological viewpoints. For machine learning modeling, the reduction of feature space can lower the risk of overfitting by removing irrelevant features which can confuse learning algorithms and lead to higher variance of the testing error. The biological significance consists mainly in the improved model interpretability since (i) irrelevant features are likely to be removed (features which do not participate in the final model showing a better or equivalent performance) and (ii) the aggregated features are more comprehensible, in comparison with single genes, by their direct association to known biological processes and structures.

We hypothesize that a proper feature selection technique combined with the set-level transformation could possibly boost the predictive performance in comparison with the alternative gene-level representation. Furthermore, this combined approach provides the proper means for the enriched analysis of gene expression data with prior knowledge. We test this hypothesis in Chapter 5, where hypothesis H2 is evaluated for the state-of-the-art gene sets, and Chapter 6, where hypothesis H2 is evaluated for novel experimental gene sets based on transcriptional regulatory networks of prokaryotic bacteria Escherichia coli.

Chapter 4

Cross-species and cross-platform analysis

(Evaluation of Hypothesis 1)

This chapter is dedicated to the evaluation of Hypothesis 1 and is organized as follows. Section 4.1 introduces cross-species and cross-platform analysis. In Section 4.2 we describe methodological ingredients of our approach consisting of data normalization, data integration, predictive classification, and statistical evaluation. Section 4.3 describes and discusses obtained results, and Section 4.4 gives a conclusion of this chapter.

4.1 Background

The recent exponential growth in publicly available microarray data sets aroused interest in the meta-analysis of gene expression data originated from different labs, platforms, or even species.[1] At the same time, there is an obvious room for boosting the analysis of actual samples under study with related samples from a public repository. However, despite the large number of samples, the prohibitive number of experimental conditions and platforms complicates the construction of a representative set of samples measured on the same array and there is the need to allow for heterogeneous platforms to keep the same conditions as in the study of interest. We examine here the hypothesis (Section 3.1) if the integration based on the set-level aggregation combined with machine learning can have a positive effect on performance of the learned machine learning models.

Latest cross-platform methods tackle the problems of matching probes on different microarray platforms (Kuhn et al., 2008) or their proper normalization (Shabalin et al., 2008). There are also dedicated tools facilitating cross-platform analysis of gene expression data (Kim et al., 2011; Lacson et al., 2010). An overview of cross-species

[1] The Gene Expression Omnibus (GEO) (Barrett et al., 2013) is one of the largest public repositories for gene expression data. Currently it stores more than 700,000 expression samples and this number is quickly growing.

analysis of microarray data is given in Lu et al. (2009). This paper distinguishes three strategies for combining microarray data sets from multiple species. The first one is the most straightforward, it takes the same array for all species, the consequent analysis in principle does not differ from the single species oriented. Obviously, this approach is suitable for closely related organisms only. The second strategy takes dedicated platforms for different species, analyzes them independently, and the results combines in a post-processing phase. In a typical scenario, lists of differentially expressed genes are found in the individual platforms, later they are compared for overlap. The third strategy analyzes the microarray data from heterogeneous platforms concurrently. This groups of methods is most relevant to the topic of this chapter which studies how classifiers learned from single-platform data compare in terms of predictive accuracy to those learned from data integrated from heterogeneous platforms. However these methods have primarily been applied to study the cell cycle. The individual exceptions report contradictory conclusions. The first research paper (Warnat et al., 2005) that focused directly on cross-platform analysis of cancer microarray data showed that it improves gene expression based classification of cancer. On the contrary, Bevilacqua et al. (2012) concludes that a better way to improve accurate signature from microarray data sets is to apply a meta-analysis rather than merging all raw data.

Another aspect addressed here is the set-level analysis of gene expression data, as opposed to the more traditional gene-level analysis approaches. Following the approach based on combination of machine learning and the set-level aggregation described in Section 2.6, we firstly group genes into prior determined gene sets based on relevant prior knowledge. For example, such set may correspond to a group of proteins acting as enzymes in a biochemical pathway or be a set of genes sharing a GO term. Naturally, gene sets considered for the analysis may on one hand overlap while on the other hand their union may not exhaust all the genes screened in the expression data. Any gene set may then be assigned descriptive values (such as aggregated expression score, fold change, significance) by aggregation of the analogical values pertaining to its members. Gene sets thus may act as derived sample features replacing the original gene expressions.

In the context of cross-platform and cross-species methods, there were also some efforts to apply gene sets instead of *orthologous genes*[2] (Kumar et al., 2005; Liu et al., 2007b; Manoli et al., 2006). GO terms or pathways were used as set-level features and set-level approaches such as GSEA were used for their selection. However, different platforms and species were analyzed independently and the results were combined in a post-processing phase as in the aforementioned second strategy.

The main contribution of experiments in this chapter is showing that the gene-set-based approach naturally enables to analyze in an integrated manner gene expression data collected from heterogeneous platforms, which may even encompass different organism species. The practical significance of the current contribution is at least twofold. First, microarray experiments are costly, often resulting in num-

[2] DNA sequences of different species are said to be orthologous if they are considered to be descended from a single sequence of their last common ancestor organism.

bers of samples insufficient for reliable modeling. The possibility of systematically integrating the experimenter's data with numerous public expression samples coming from heterogeneous platforms, would obviously help the experimenter. Second, such integrated analysis provides the principal means to discover biological markers shared by different-genome species.

We consider three types of gene set collections. The first type groups genes that share a common GO term. The second type groups genes acting in biological pathways formalized by the KEGG database. The third type represents a further novel contribution of our work and is based on the notion of a *fully coupled flux*, which is a pattern prescribing pathway partitions hypothesized by Notebaart et al. (2008) to involve strongly co-expressed genes. These synergize in single gradually amplified biological functions such as enzymatic catalysis or translocation among different cellular compartments.

4.2 Materials and methods

The *input* of our experimental workflow is a set of gene expression samples possibly measured by different microarray platforms. To each sample are assigned two labels. The first identifies the microarray platform from which the sample originates, the second identifies a sample class (e.g., tissue type). The *output* is a classification model, that is, a model that estimates the sample class given an expression sample and its platform label. The model is obviously applicable to any sample not present in the input ("training") data, as long as its platform label is also known. The remarkable property of the output model in our approach is that it is not a combination of separate models each pertaining to a single platform. Rather, it is a single classifier trained from the entire heterogeneous sample set and represented in terms of activity scores (Section 2.6) of units that apply to all platforms, albeit the computation of these activity scores may be different across platforms. More specifically, the activity score of a gene set (such as a pathway) is calculated using a different gene set in each platform. We now describe the individual steps of the method in more detail.

Normalization

The overview and comparison of cross-platform normalization methods for gene expression data has recently been published by Rudy and Valafar (2011). For experiments presented in this chapter, we applied quantile normalization (Bolstad et al., 2003) adapted to cross-platform utilization in a similar way as described by Lacson et al. (2010). Quantile normalization cannot be applied directly as different platforms represent the individual ortholog genes with probesets of various sizes. To enable quantile normalization, interpolation and aggregation is needed prior to calculating sample quantiles. As a result, all samples independently of the platform exhibit the same distribution of expression values. We conduct these steps using the Bioconductor (Gentleman et al., 2004) software.

Probeset matching

In order to match probes on different microarray platforms we use `annotationTools`, a Bioconductor package (Kuhn et al., 2008). The package provides *orthology tables* which contain definitions of orthologous genes and allow reliable integration of heterogeneous data with satisfactory coverage.

Set construction

During evaluation of H1 we consider three types of prior knowledge. The first type groups genes that share a common GO term. The second type groups genes acting in biological pathways formalized by the KEGG (Kanehisa et al., 2004) database. The third gene set collection type is based on the notion of a *fully coupled flux* (FCF), motivated as follows. Many notable biological conditions are characterized by the activation of only certain parts of pathways (for example, see references Shaw and Filbert, 2009; Sun and Chen, 2008; Weichhart and Säemann, 2008). The notion of pathway score implied by the previous gene set may thus often violate intuition and hinder interpretation. Therefore we extracted all pathway partitions which comply with the graph-theoretic notion of FCF (Notebaart et al., 2008). It is known that the genes coupled by their enzymatic fluxes not only show similar expression patterns, but also share transcriptional regulators, and frequently reside in the same operon in prokaryotes or similar eukaryotic multi-gene units such as the hematopoietic globin gene cluster (Notebaart et al., 2008). FCF is a special kind of network flux that corresponds to a pathway partition in which non-zero flux for one reaction implies a non-zero flux for the other reactions and vice versa. It is the strongest qualitative connectivity that can be identified in a network. The notion of an FCF is explained through an example in Fig. 4.1 (for a detailed definition, see reference Notebaart et al., 2008). Pathway partitions forming FCF's constitute the third gene set collection. Each type of prior knowledge is represented as a gene set collection where, e.g., each pathway is represented as a gene set without considering any type of additional interactions. Since each above mentioned gene set represents a biological property shared among different organisms, a gene set collection can be represented by genes specific for a certain organism using the orthologous genes definitions as a kind of dictionary.

In what follows, gene sets act as features acquiring a real value for each sample. Formally, let $G^{(k)}$ be the set of genes interrogated by given platform $k \in K$ and Σ is a gene set collection of a particular type. We define a mapping:

$$A_{G^{(k)}} : \mathbb{R}^{|G^{(k)}|} \times \Sigma \to \mathbb{R}.$$

For any expression sample $s^{(k)} = (e_1, \ldots, e_{|G^{(k)}|}) \in \mathbb{R}^{|G^{(k)}|}$ obtained from platform k, the mapping $A_{G^{(k)}}(s^{(k)}, \Gamma)$ should collectively quantify the activity score of genes in set $\Gamma \in \Sigma$ in a biological situation (e.g., a tissue type) sampled by $s^{(k)}$. Typically, not all members of Γ will be measured by platform k and the computation of $A_{G^{(k)}}(s^{(k)}, \Gamma)$ will be based on expressions of the genes in $\Gamma \cap G^{(k)}$. For experiments in this chapter

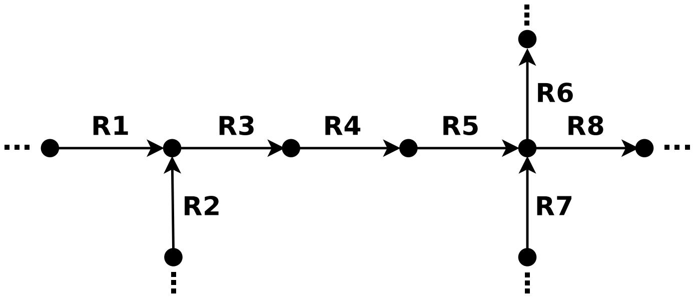


Figure 4.1: Fully coupled fluxes in a simplified network with nodes representing chemical compounds and arrows as symbols for chemical reactions among them. Each arrow can be labeled by a protein. R3, R4 and R5 are fully coupled as a flux in any of these reactions implies a flux in the rest of them. Note that R1 and R3 do not constitute a FCF as a flux in R3 does not imply a flux in R1.

we define $A_{G^{(k)}}(s^{(k)}, \Gamma)$ as the average of expressions measured in s for all genes in $\Gamma \cap G^{(k)}$. Although there are more sophisticated methods to instantiate $A_{G^{(k)}}(s^{(k)}, \Gamma)$ (see Section 2.6), we use the average mainly due to the fact that this method—unlike the other more sophisticated methods—computes set-level activities *independently* to other samples. We consider this approach as sufficient and appropriate mainly because (i) the difference between aggregation by average and the best performing method is not significant on comparable gene sets (Mramor et al., 2010), (ii) the evaluation of H1 requires learning on datasets with relative small number of samples, and (iii) most importantly application of this aggregation approach is straightforward in comparison to other methods which were designed for single-platform data only. However, we admit the re-implementation of the other set-level aggregation methods in cross-platform and cross-species manner could bring more significant results.

Our reasoning above assumes the aggregation of gene expression measurements. Precisely speaking, genes themselves aggregate one or more measurements since multiple probesets can represent the same gene. Here, the expression of a gene is simply defined as the average of the corresponding normalized probeset measurements, despite certain caveats of this approach.[3]

Set-level data integration

The goal of this methodological step is to integrate heterogeneous expression samples into a single-tabular representation (the integrated representation contains samples sharing a common feature set defined by a gene set collection, e.g., a set of pathways) that predictive classification algorithms can process. Formally, we have a set of expression samples from all the platforms $S = \{s_1^{(k)}, s_2^{(k)}, \ldots, s_n^{(k)}\}$ where $k \in K$,

[3] For example, Affymetrix chips contain probesets representing the same gene that cannot be consolidated into unique measures of transcription due to alternative splicing, use of alternative poly(A) signals, or incorrect annotations (Stalteri and Harrison, 2007).

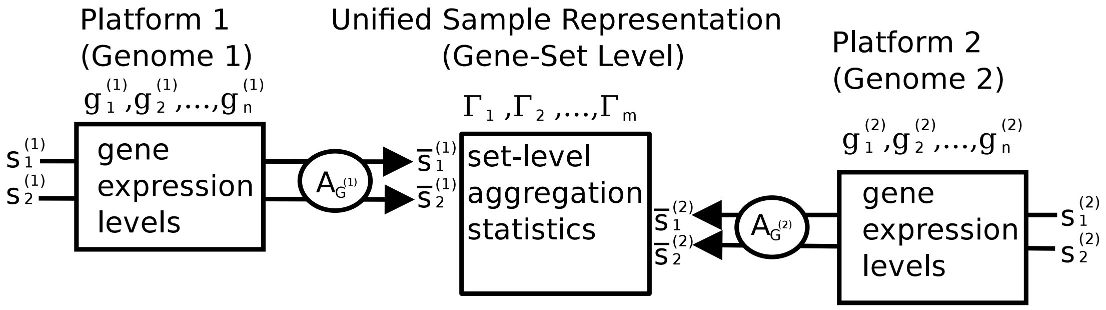


Figure 4.2: Integrating expression data collected from heterogeneous platforms into a unified tabular representation of set-level aggregation statistics. If these platforms pertain to different organisms, we assume that (an ortholog of) each gene set, Γ_i, (e.g., cellular pathway) exists in each of the organisms.

$s_i^{(k)} \in \mathbb{R}^{|G^{(k)}|}$, and n is number of all the measured samples we intend to integrate. We wish to obtain a new representation $\bar{S} = \{\bar{s}_1^{(k)}, \bar{s}_2^{(k)}, \ldots, \bar{s}_n^{(k)}\}$ where each $\bar{s}_i^{(k)} \in \mathbb{R}^{|\Sigma|}$.

This aim is achieved using the above mentioned gene set aggregation concept. Formally, using gene set collection $\Sigma = \{\Gamma_1, \Gamma_2, \ldots, \Gamma_m\}$, for each sample $s_i^{(k)}$ we stipulate:

$$\bar{s}_i^{(k)} = \left(A_{G^{(k)}}(s_i^{(k)}, \Gamma_1), \ldots, A_{G^{(k)}}(s_i^{(k)}, \Gamma_m) \right).$$

Naturally, sample $\bar{s}_i^{(k)}$ then inherits the class label from $s_i^{(k)}$. The integration principle is exemplified in Fig. 4.2. The most straightforward representation based on the ortholog genes is used as the baseline representation to evaluate the effect of set-level features.

Classification and Validation

The final step of the workflow is to employ machine learning algorithms to induce predictive classification models of the integrated samples. As the achieved unified representation $\bar{S}$ can be processed by virtually any machine learning algorithm. Since support vector machine learners are known to cope well with the frequent characteristics of gene expression datasets such as noise and strong disproportion in the amount of genes and samples, we decided to work with them when classification accuracy is the main concern. When direct human or semi-automated inspection or interpretation of the target classifiers is needed we decided to take advantage of decision-tree classifiers. Specifically, we experimented with the SMO support vector machines learner and J48 decision tree learner included the machine learning environment Weka (Witten and Eibe, 2005).

The design of the experiments and the validation protocol for Hypothesis 1 is dictated by the following questions we wish to address empirically.

Q1 How do classifiers learned from *single-platform* data compare to those learned from data integrated from *heterogeneous platforms* in terms of predictive accuracy?

Q2 How do classifiers based on original *single gene* expressions compare in terms of predictive accuracy to those based on activity of biologically meaningful *gene set collections*?

Q3 Do classifiers based on gene set collections improve those based on single genes namely in situations when combining cross-species expression data characterized by only rough similarity in phenotype markers?

Note that while Q1 is evaluated on orthologous genes where we compare the models learned on the integrated and single-platform data, the assessment of Q2 is consisting of the learned model performance evaluation on integrated datasets with different levels of generality; particularly, we compare the aggregated and non-aggregated data (with gene sets and orthologous genes as features, respectively) . In the case of Q3, we are concerned with "difficult domains" only which have the following characteristics: (i) the individual single-platform gene-based models poorly generalize to the samples taken from the alternative platform (e.g., the Mus musculus platform if the original was a Homo sapiens platform), (ii) the general cross-platform model is large in its size (a classification tree with a large number of nodes) in comparison with the single-platform models derived for the same domain. The initial assumption is that gene sets shall increase the robustness of classifiers and help to increase their generality when applied across platforms; therefore, we compare model sizes of classifiers learned on integrated and single-platform data.

We are interested in the insights Q1-Q3 for both the "data-rich" and "data-poor" situation (for both small and large sets of expression samples). Therefore the preferred means of assessment is through *learning curves* which are diagrams plotting an unbiased estimate of the classifier's predictive accuracy against the proportion p of the available data set used for its training. The accuracy estimate for each measured p was obtained by inducing a classifier 100 times with a randomly chosen subset (of proportional size p) of the entire data set and its accuracy is tested on the remaining data not used for training. In each such step, the 100 empirical accuracy results were averaged into the reported value. We let p range from 0.1 to 0.5 to prevent statistical artifacts arising from overly small sets used for training or testing and avoid single-platform tasks with large sample sets that do not ask for cross-platform generalization.

Statistical evaluation

All statistical tests conducted are based on the Wilcoxon signed-rank test (two-sided unless stated otherwise). For pairing, we always relate two experiments equal in terms of all settings except for the one under study. For example, when testing Q1 we gradually fix the way of feature extraction, the domain, the proportion of data taken for training, the learning algorithm, but alter the single-platform and cross-platform way of learning. In this way, a pair of classification accuracies is derived for every setting. In order to avoid dependency among differential accuracies, median differential accuracy is calculated for every domain. Having 20 domains, the Wilcoxon test is applied to the vector of 20 differential accuracies. The hypotheses are tested at a level of significance of $\alpha = 0.05$, Bonferroni correction is used to counteract the problem of multiple comparisons. The stronger t-test is more usual in analysis of predictive accuracy samples in literature but our preliminary normality tests do not justify its application. Given the extent of the collected samples, the Wilcoxon test is sufficient to support the conclusions reported. Besides, the Wilcoxon test is argued to be statistically safer than the t-test for comparing classification algorithms over multiple data sets (Demšar, 2006).

4.3 Results and discussion

Here we show the empirical results obtained by processing the data described below by the method explained in Section 4.2 and comment on their relevance to questions (Q1-Q3) formulated in the same place.

4.3.1 Classification tasks and data

Here we validate our methodology in biological classification tasks. In order to avoid domain bias, we chose not to tackle overly special classification cases, the most problems distinguish between two tissue or cell types. We conducted our experiments on 20 cross-species classification tasks based on gene expression samples downloaded from the GEO. For ease of the experiment design, we take into the account following constraints: (i) each problem contains two species-specific subtasks (e.g., the classification of *Tuberculosis* on Homo sapiens and Mus musculus data), (ii) only two class classification problems are concerned, and (iii) exclusively the gene expression platforms provided by Affymetrix are taken.

Set type	**Total**	**Genes contained**			
		Min	Max	Avg	Median
FCF	**605**	1	48	5.53	3
Pathway	**880**	5	446	44.79	26
GO term	**1454**	6	2045	89.57	26

Table 4.1: Gene set statistics. The numbers in bold are independent of the specific platforms measuring the expression data, being only determined by the respective types of prior knowledge. The “Genes contained” columns capture statistics over all involved platforms.

A detailed overview of the tasks including the number of samples in each class, sample sources, and microarray platforms used for the measurement of the gene expression is in Table 4.2. Table 4.1 shows statistics derived from the application of a priori constructed gene sets onto the collected expression samples.

4.3.2 Reached results

The experiments revealed fundamental differences among the individual domains from the point of view of their learnability and suitability for cross-species analysis. The principal categories of learning curves and corresponding domain types are shown in Figure 4.3. Figure 4.3a represents a domain that is easily learnable even with a small number of training samples, the single-platform models show an excellent accuracy and the accuracy remains good even when the single-platform models are applied to the samples taken from the contrary platform. As expected, the cross-platform model is excellent as well and the performance remains nearly absolute when altering feature extraction methods. This domain is neutral when answering the questions Q1 and Q2 and does not fit the question Q3. Figure 4.3b corresponds to a domain that is learnable, but not trivially. Even though the classifiers less agree across platforms and species, the generalization across platforms and genes helps. The cross-platform classifier outperforms the single-platform ones and gene sets are more accurate than single genes. Figure 4.3c demonstrates a learnable domain that makes a counterexample to the previous one, i.e., it rather suggests negative answers to all the concerned questions. Figure 4.3d shows a domain in which learning is difficult, better than random classifiers can be reached in one organism only. The generalization does not help, the models do not scale across platforms independently of selected features. This domain is again neutral when answering the questions Q1 and Q2 and does not fit the question Q3.

Let us address the individual questions statistically following the detailed methodology given in the last paragraph of Section 4.2. The analysis concerns our pool of domains as unbiased from the point of view of the above mentioned learnability categorization. The domains were constructed purely with the effort to meet the technical criteria such as the agreement in experimental conditions between platforms and species, the representative sample sizes and keeping a limited set of platforms.

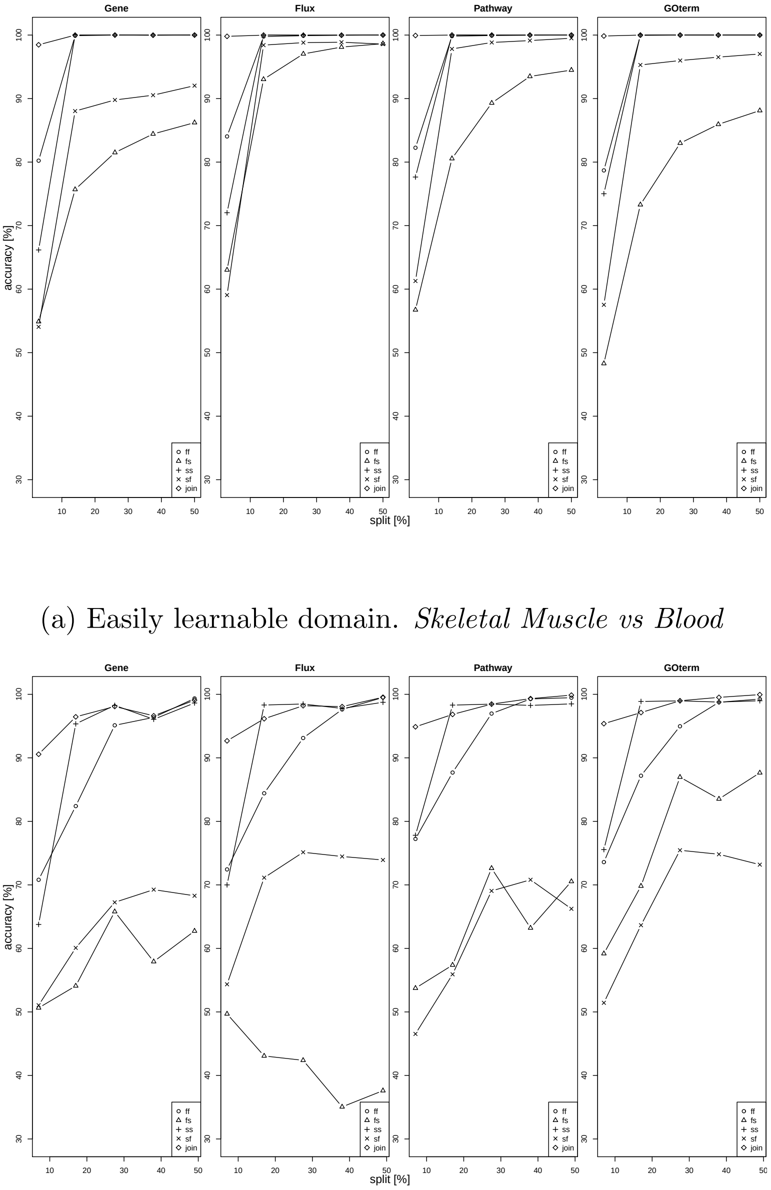
The question Q1 compares the single-platform and cross-platform classifiers. The experiments proved that the cross-platform classifiers significantly outperform the single-platform ones for small proportions of training examples (the p-value $4e^{-6}$ for the 10% proportion of training samples, the p-value 0.02 for the 20% proportion of training samples). For larger training sets, the single-platform classifiers show higher accuracy than their cross-platform counterparts, however the margin is not large and it is statistically insignificant (the mean difference for the largest proportion of 50% samples used for training is 0.8%). The general recommendation is to prefer cross-platform learning when the sample sizes are small (less than a few tens of samples per platform). For larger sample sets, the general model can still be prefered but for other reasons than the classification accuracy itself.

The question Q2 compares the classifiers based on original single gene expressions with those based on activations of biologically meaningful gene sets. All the set of 15 comparative experiments (5 proportions and 3 different ways of gene set construction) resulted in higher accuracy of gene set based classifiers (when averaged over 20 domains). When Q2 is posed without constraining the sample size, the gene sets can be recommended and the recommendation is statistically supported (p-values 0.001 for pathways, 0.0002 for fluxes, and 0.004 for GO terms). When testing separately for different sample sizes, all the gene set collections significantly outperform genes for small sample sizes (the 10% and 20% proportion of training samples), fluxes keep the statistically significant difference even for the largest sample sizes (p-value 0.002 for the 50% proportion of training samples). The general conclusion is to prefer the gene-set-based models, the level of gene set abstraction (see the mean numbers in Table 4.1) shall decrease with the increasing sample sizes.

Finally, to answer Q3 we identified domains where combining cross-species expression data is characterized by only rough similarity in phenotype markers. We found 5 domains (Huntington's Disease A and B, Retinoblastoma, Tuberculosis and Breast Tumors) meeting the technical criteria mentioned in Section 4.2, i.e., the poor generalization to the samples taken from the contrary platform and the large size of gene based cross-platform model. The gene-set-based models met the assumption that their generalization ability helps to improve the classification accuracy of gene-based classifiers, but we could not statistically prove that the increase is larger than the increase observed in an arbitrary domain. The number of domains is not large enough.

4.4 Conclusions

To examine Hypothesis 1, we demonstrate the integration of multi-platform gene expression data for predictive classification. When single-platform samples are rare, integration of related (cross-platform and cross-species) data boosts classification performance which supports the first hypothesis for the limited number of available samples only. In addition, we explored three ways of defining gene sets, including that based on the notion of a fully coupled flux representing a trade-off between very specific genes and general metabolic pathways. In 20 cross-platform classification tasks, we showed that the gene-set-based representation is useful for combining heterogeneous gene expression data. This may be for the sake of assembling a larger sample set or to obtain general biological insights not limited to a particular organism. The gene set features significantly outperform the gene-oriented ones in small sample sets (the training sets containing 10% and 20% of available samples), the fluxes keep this property even for the largest tested sample sets (the training sets containing 50% of available samples). The pathways and GO terms also give higher predictive accuracies than the gene-based features, but the significance of this difference cannot be proved on the selected significance level.

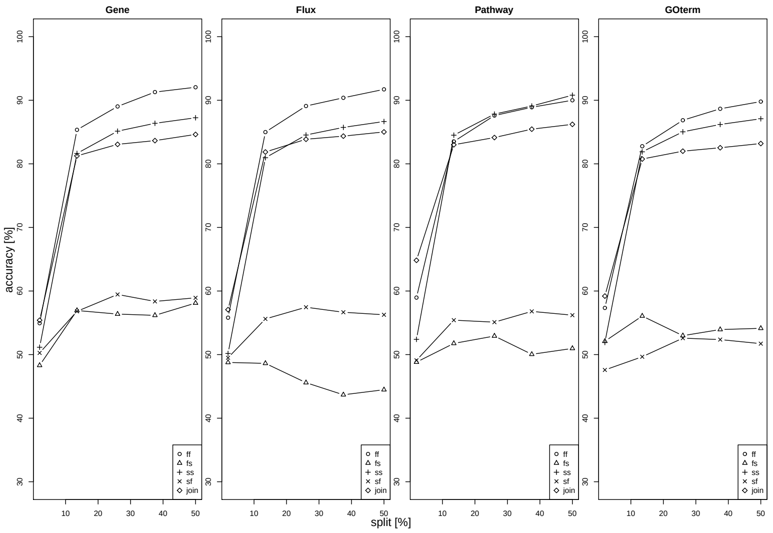


(a) Easily learnable domain. *Skeletal Muscle vs Blood*

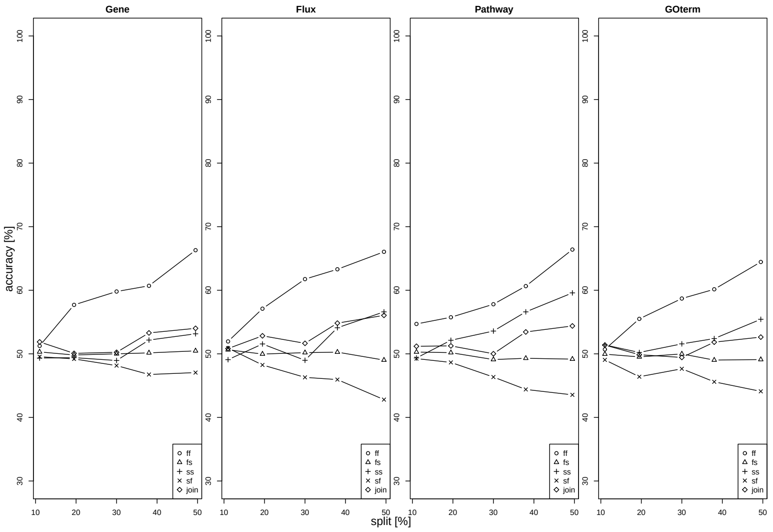
(b) Two co-learnable subproblems. *Skeletal Muscle vs Liver (XGE)*

Platform 1				GPL	Platform 2				GPL	Experiment description
Class 1		Class 2			Class 1		Class 2			
#	dataset id	#	dataset id		#	dataset id	#	dataset id		
13	S10263	13	S10263	1261	12	D2887	10	D2887	570	Huntington's Disease A
10	D3620	8	D3620	1261	12	D2887	10	D2887	570	Huntington's Disease B
6	XGE	11	XGE	1261	41	XGE	6	XGE	96	Skeletal Muscle vs Liver(XGE)
6	XGE	11	XGE	1261	41	XGE	20	XGE	96	Skeletal Muscle vs Brain A(XGE)
108	S29686	24	S29686	1261	75	S29686	4	S29686	570	Retinoblastoma
12	S11199	12	S11199	570	10	S14316	4	S14316	8321	Tuberculosis
65	S27567	28	S27567	1261	131	S27567	31	S27567	570	Breast Tumors
15	XGE	22	XGE	91	41	XGE	20	XGE	96	Skeletal Muscle vs Brain B(XGE)
18	XGE	33	XGE	96	18	XGE	26	XGE	97	Heme vs Stroma (XGE)
54	S15184	38	S15184	1355	12	S5364	9	S5364	96	Lung vs Colon
10	S11291	25	S20465	1261	16	S3307	13	S5364	96	Skeletal Muscle vs Breast
10	S11291	39	D3622	1261	16	S3307	12	S5364	96	Skeletal Muscle vs Lung
10	S11291	25	S20465	1261	16	S3307	14	5364 12630	96	Skeletal Muscle vs Liver
10	S11291	25	S20465	1261	16	S3307	112	S12288	96	Skeletal Muscle vs Blood
25	S20465	25	S20465	1261	14	S5364 & S12630	13	S5364	96	Liver vs Breast
25	S20465	25	S20465	1261	14	S5364 & S12630	112	S12288	96	Liver vs Blood
25	S20465	39	D3622	1261	14	S5364 & S12630	12	S5364	96	Liver vs Lung
39	D3622	25	S20465	1261	12	S5364	13	S5364	96	Lung vs Breast
39	D3622	25	S20465	1261	12	S5364	112	S12288	96	Lung vs Blood
25	S20465	25	S20465	1261	112	S12288	13	S5364	96	Blood vs Breast

Table 4.2: An overview of cross-species domains. Each resource (a platform specific single species subproblem) is described by: *NCBI's GEO repository* unique data sets (GDS) or data series (GSE) description number, the datasets and series are distinguished by prefixes *D* and *S* respectively in the table; the platform description numbers (GPL) provide information about species: 91, 96, 97, 570 for *homo sapiens*, 1261 and 8321 for *mus musculus* and 1355 for *rattus norvegicus*; the numbers of samples available for the individual classes in the individual subproblmes are also given.



(c) Learnable subproblems hardly co-learnable. *Breast Tumors*



(d) Hardly learnable domain. *Huntington's Disease B*

Figure 4.3: The principal categories of learning curves with references to the corresponding particular domains. The legend is as follows: (ff)—the single-platform model learned on the training data from the (f)irst subproblem and tested on the samples taken from the (f)irst subproblem, (fs)— the single-platform model learned on the training data from the (f)irst subproblem and tested on the samples taken from the (s)econd subproblem, (ss) and (sf)—defined likewise for the second subproblem, (join)—the cross-platform model learned and tested on the samples merged from both the subproblems.

Chapter 5

A vanilla approach to descreasing dimensionality

(Evaluation of Hypothesis 2 with state-of-the-art gene sets)

This chapter is dedicated to the evaluation of Hypothesis 2 on the state-of-the-art gene sets commonly used for the set-level analysis.

Initial studies into the predictive performance of set-level classifiers have yielded rather contradictory results (e.g., compare papers by Lee et al., 2008; Mramor et al., 2010). Here we offer a more conclusive evaluation by testing various components of the set-level framework within a large collection of machine learning experiments on different data domains. Section 5.1 gives background for the evaluation of Hypothesis 1 for the state-of-the-art gene sets. Section 5.2 specifies used methods and data. Experimental results are divided into two sections: Section 5.3 contains general performance evaluation of selected set-level aggregation techniques, and Section 5.4 contains additional experiments dealing with an evaluation of combination of machine learning and feature selection either of set-level aggregated or unaggregated data and analyses successful gene sets. In Section 5.5 we conclude and discuss the results.

5.1 Background

Lifting features to the set level incurs a significant compression of the training data since the number of considered gene sets is typically much smaller than the number of interrogated genes, as we already mentioned in Section 2.6. This compression raises the natural question whether relevant information is lost in the transformation, and whether the augmented interpretability will be outweighed by compromised predictive accuracy. On the other hand, reducing the number of sample features may mitigate the risk of overfitting and thus, conversely, contribute to higher accuracy. In machine learning terms, reformulation of data samples through set-level features increases the *bias* and decreases the *variance* of the learning process (see Section 2.4). An objective of experiments in this chapter is to examine the second hypothesis; particularly, to experimentally assess the combined effect of the two antagonistic factors (combining feature selection techniques with the set-level transformation) on the resulting predictive accuracy.

Another aspect of transforming features to the set level is that biological prior knowledge is channeled into learning through the prior definitions of biologically plausible gene sets. Among the goals of this chapter is to assess how significantly such prior knowledge contributes to the performance of learned classifiers. We do this assessment by comparing classification accuracy achieved with genuine curated gene sets against that obtained with gene sets identical to the latter in number and sizes, yet lacking any biological relevance. We also investigate patterns distinguishing genuine gene sets particularly useful for classification from those less useful.

A further objective is to evaluate—from the machine learning perspective—some representative statistical techniques proposed recently in the research on set-level gene expression analysis. We select the Gene Set Enrichment Analysis (GSEA) method (Subramanian et al., 2005) as a representant of the very popular technique among biologists (despite its poor performance), a technique known as the Global test (Goeman and Bühlmann, 2007) which is considered as one of the best performing methods, and the SAM-GS algorithm (Dinu et al., 2007) also considered as a well performing method (see Section 2.3 for the review). Informally, they rank a given collection of gene sets according to their correlation with phenotype classes. The methods naturally translate into the machine learning context in that they facilitate feature selection (Liu and Motoda, 1998), i.e., they are used to determine which gene sets should be provided as sample features to the learning algorithm. We experimentally verify whether these methods work reasonably in the classification setting, i.e., whether learning algorithms produce better classifiers from gene sets ranked high by the mentioned methods than from those ranking lower. We investigate classification conducted with a single selected gene set as well as with a batch of high ranking sets. Furthermore, we test how the three gene-set-specific methods compare to some generic feature selection heuristics (information gain and support vector machine with recursive feature elimination) known from machine learning.

To use a machine learning algorithm, a unique value for each feature of each training sample must be established. Set-level features correspond to multiple expressions and these must therefore be aggregated. We comparatively evaluate three selected aggregation approaches to cover different ways for set-level aggregation (see Section 2.6 for the complete methods review). The first (AVG) simply averages the expressions of the involved genes. The value assigned to a sample and a gene set is independent of other samples and classes. The other two, more sophisticated, methods (SVD, SetSig) rely respectively on the singular value decomposition principle (e.g., Tomfohr et al., 2005) and the so-called gene set signatures, which should perform better than both other approaches (Mramor et al., 2010). In the latter two approaches, the value assigned to a given sample and a gene set depends also on expressions measured in other samples. Let us return to the initial experimental question concerned with how the final predictive accuracy is influenced by the training data compression incurred by reformulating features to the set level. As follows from the above, two factors contribute to this compression: selection (not every gene from the original sample representation is a member of a gene set used in the set-level representation, i.e.

some interrogated genes become ignored) and aggregation (for every gene set in the set-level representation, expressions of all its members are aggregated into a single value). We quantify the effects of these factors on predictive accuracy. Regarding selection, we experiment with set-level representations based on *10 best gene sets* and *one best gene set*, respectively, with both numbers chosen ad-hoc. The two options are applied with all three selection methods (GSEA, SAM-GS, Global). We compare the obtained accuracy to the baseline case where all individual genes are provided as features to the learning algorithm, and to an augmented baseline case where a prior feature-selection step is taken using the information gain heuristic. For each of the selection cases, we further evaluate the contribution of the aggregation factor. This evaluation is done by comparing all the three aggregation mechanisms (AVG, SVD, SetSig) to the control case where no aggregation is performed at all; in this case, individual genes combined from the selected gene groups act as features.

The key contribution of experiments in this chapter is thus a thorough evaluation of a number of aspects and methods of the set-level strategy employed in the machine learning context, entailing the reformulation of various, independently published relevant techniques into a unified framework. Such a contribution is important not only due to the current state of the art in microarray data analysis, where generally the need for methods evaluation seems to be higher than for their development (see Section 2.5), but particularly also due to the inconclusive results of previous, less extensive studies indicating both superiority (e.g., Lee et al., 2008) and inferiority (Mramor et al., 2010, Sec. 4) of the set-level approach to classificatory machine learning, with respect to the accuracy achievable by the baseline gene-level approach.

Our contributions are, however, also significant beyond the machine learning scope. In the general area of set-level expression analysis, it is undoubtedly important to establish a performance ranking of the various statistical techniques for the identification of significant gene sets in class-labeled expression data. This is made difficult by the lack of an unquestionable ranking criterion—there is in general no ground truth stipulating which gene sets should indeed be identified by the tested algorithms. The typical approach embraced by comparative studies such as (Dinu et al., 2007) is thus to appeal to intuition (e.g. *the p53 pathway should be identified in p53-gene mutation data*). However legitimate such arguments are, evaluations based on them are obviously limited in generality and objectivity. We propose that the predictive classification setting supported by the cross-validation procedure for unbiased accuracy estimation, as adopted here, represents exactly such a needed framework enabling objective comparative assessment of gene set selection techniques. In this framework, results of gene set selection are deemed good if the selected gene sets allow accurate classification of new samples. Through cross-validation, the accuracy can be estimated in an unbiased manner.

5.2 Materials and methods

Here we first specify the methods adopted for gene set ranking, gene expression aggregation, and for classifier learning. Next we present the datasets used as benchmarks in the comparative experiments. Lastly, we describe the protocol followed by the experiments in this chapter.

5.2.1 Gene set ranking

Three statistics-based methods (GSEA, SAM-GS, and Global test) are considered for ranking gene sets in this chapter. As inputs, all of the methods take a set $G = \{g_1, g_2, \dots g_p\}$ of interrogated genes, and a set S of n expression samples where for each $s_i \in S$, $s_i = (e_{1,i}, e_{2,i}, \dots e_{p,i}) \in \mathbb{R}^p$ where $e_{j,i}$ denotes the (normalized) expression of gene g_j in sample s_i. The sample set S is partitioned into phenotype classes $S = C_1 \cup C_2 \cup \dots \cup C_o$ so that $C_i \cap C_j = \{\}$ for $i \neq j$. To simplify experiments in this chapter, we assume binary classification, i.e., $o = 2$. A further input is a collection of gene sets Σ such that for each $\Gamma \in \Sigma$ it holds $\Gamma \subseteq G$. In the output, each of the methods ranks all gene sets in Σ by their estimated power to discriminate samples into the predefined classes.

Next we give a brief account of the three methods and refer to the original sources for a more detailed description. In experiments, we used the original implementations of the procedures as provided or published by the respective authors.

Gene set enrichment analysis (GSEA) (Subramanian et al., 2005)

This method tests a null hypothesis that gene rankings in a gene set Γ, according to an association measure with the phenotype, are randomly distributed over the rankings of all genes. It first sorts G by correlation with binary phenotype. Then it calculates an enrichment score (ES) for each $\Gamma \in \Sigma$ by walking down the sorted gene list, increasing a running-sum statistic when encountering a gene $g_i \in \Gamma$ and decreasing it otherwise. The magnitude of the change depends on the correlation of g_i with the phenotype. The enrichment score is the maximum deviation from zero encountered in the random walk. It corresponds to a weighted Kolmogorov-Smirnov-like statistic. The statistical significance of the ES is estimated by an empirical phenotype-based permutation test procedure that preserves the correlation structure of the gene expression data. GSEA was one of the first specialized gene-set analysis techniques. It has been reported to attribute statistical significance to gene sets that have no gene associated with the phenotype, and to have less power than other recent test statistics (see Section 2.3.1).

SAM-GS (Dinu et al., 2007)

This method tests a null hypothesis that the mean vectors of the expressions of genes in a gene set do not differ by phenotype. Each sample s_i is viewed as a point in an p-dimensional Euclidean space. Each gene set $\Gamma \in \Sigma$ defines its $|\Gamma|$-dimensional subspace in which projections s_i^Γ of samples s_i are given by coordinates corresponding to genes in Γ. The method judges a given Γ by how distinctly the clusters of points $\{s_i^\Gamma | s_i \in C_1\}$ and $\{s_j^\Gamma | s_j \in C_2\}$ are separated from each other in the subspace induced by Γ. SAM-GS measures the Euclidean distance between the centroids of the respective clusters and applies a permutation test to determine whether, and how significantly, this distance is larger than that obtained if samples were assigned to classes randomly.

The global test (Goeman and Bühlmann, 2007)

The global test, analogically to SAM-GS, projects the expression samples into subspaces defined by gene sets $\Gamma \in \Sigma$. In contrast to the Euclidean distance applied in SAM-GS, it proceeds instead by fitting a regression function in the subspace, such that the function value acts as the class indicator. The degree to which the two clusters are separated then corresponds to the magnitude of the coefficients of the regression function.

5.2.2 Expression aggregation

Three methods (SetSig, SVD, and AVG) are considered for assigning a value to a given gene set Γ for a given sample s_i by aggregation of expressions of genes in Γ.

Averaging (AVG)

The first method simply produces the arithmetic average of the expressions $e_{j,i}$ of all Γ genes $1 \leq j \leq p$ in sample s_i. The value assigned to the pair (s_i, Γ) is thus independent of samples s_j, $i \neq j$. Several authors utilized this approach (Abraham et al., 2010; Azuaje et al., 2010; Guo et al., 2005; Liu et al., 2007a).

Singular value decomposition (SVD)

A more sophisticated approach was also employed by many authors (Bild et al., 2006; Levine et al., 2006; Liu et al., 2007a; Tomfohr et al., 2005). Here, the value assigned to (s_i, Γ) depends on expressions $e_{j,i}$ measured in sample s_i but, unlike in the averaging case, also on expressions $e_{j,k}$ measured in samples s_k, $k \neq i$. In particular, all samples in the sample set S are viewed as points in the $|\Gamma|$-dimensional Euclidean space induced by Γ the same way as explained in Section *Gene set ranking*. Subsequently, the specific vector in the space is identified, along which the sample points exhibit maximum variance. Each point $s_k \in S$ is then projected onto this vector. Finally, the value assigned to (s_i, Γ) is the real-valued position of the projection of s_i on the maximum-variance vector in the space induced by Γ.

Gene set signatures (SetSig)

Similarly to the SVD method, the SetSig (Mramor et al., 2010) method assigns to (s_i, Γ) a value depending on expressions both in sample s_i as well as in other samples s_k, $k \neq i$. However, unlike in the previous two aggregation methods, here the value also depends on the class memberships of these samples. In particular, SetSig confines to two-class problems and the value ('signature') assigned to (s_i, Γ) can be viewed as the Student's unpaired t-statistic for the means of two populations of the Pearson correlation coefficients. The first (second) population studies correlation of s_i with the samples from the first (second) class in the space induced by Γ. Intuitively, the signature is positive (negative) if the sample correlates rather with the samples belonging to the first (second) class.

5.2.3 Machine learning

We experimented with five diverse machine learning algorithms (1-nearest neighbor, 3-nearest neighbors, naive Bayes, decision tree, and support vector machine) to avoid dependence of experimental results on a specific choice of a learning method. These algorithms are explained in depth for example by Hastie et al. (2001). In experiments, we used the implementations available in the WEKA software (Hall et al., 2009) with the default settings. None of the methods below is in principle superior to the others, although the first one prevails in predictive modeling of gene expression data and is usually associated with high resistance to noise in data.

Support vector machine

Samples are viewed as points in a vector space with coordinates given by the values of its features. A classifier is sought in the form of a hyperplane that separates training samples of distinct classes and maximizes the distance to the points nearest to the hyperplane (i.e. maximizing the *margin*) in that space or in a space of extended dimension into which the original vector space is non-linearly projected.

1-nearest neighbor

This algorithm is a simple form of classification proceeding without learning a formal data model. A new sample is always predicted to have the same class as the most similar sample (i.e. the nearest neighbor) available in training data. We use the Euclidean metric to measure the similarity of two samples.

3-nearest neighbors

This method is similar to 1-Nearest Neighbor, except that class is determined as one prevailing among the three, rather than one, most similar samples in training data. This method becomes superior to the previous one as noise in data exceeds a certain threshold amount. The threshold value (and thus the optimal number of considered neighbors) is in general not known.

Naive Bayes

A sample is classified into the class that is most probable given the sample's feature values, according to a conditional probability distribution learned from training data on the simplifying assumption that, within each class, all features are mutually independent random variables. Gene expression data usually deviate from this assumption and consequently the method becomes suboptimal.

Decision tree

A tree-graph model enables to derive a class prediction for a sample by following a path from the root to a leaf of the tree, where the path is determined by outcomes of tests on the values of features specified in the internal nodes of the tree. The tree model is learned from training data and can also be represented as a set of decision rules.

5.2.4 Expression and gene sets

We conduct our experiments using 30 public gene expression datasets, each containing samples categorized into two classes. This collection contains both hard and easy classification problems (see Figure 5.1). The individual datasets are listed in Table 5.1 and annotated in more detail in the supplemental material[1].

Besides expression datasets, we utilized a gene set database consisting of 3272 manually curated sets of genes obtained from the Molecular Signatures Database (MSigDB v3.0). These gene sets have been compiled from various online databases (e.g. KEGG, GenMAPP, BioCarta).

For control experiments, we also prepared another collection of gene sets that is identical to the latter in the number of contained sets and the distribution of their cardinalities. However, the contained sets are assembled from random genes and have no biological significance. The particular method used to obtain the randomized gene sets is as follows. For sampling, we consider the set $\mathcal{G}$ of all genes occurring in some of the genuine gene sets, formally $\mathcal{G} = \{g | g \in \Gamma, \Gamma \in \Sigma\}$. Then, for each genuine gene set Γ, we sample $|\Gamma|$ genes without replacement uniformly from $\mathcal{G}$ to constitute the counterpart random gene set Γ'.

5.2.5 Experimental protocol

Classifier learning in the set-level framework follows a simple workflow. Its performance is influenced by several factors, each corresponding to a particular choice from a class of techniques (such as for gene set ranking). We evaluate the contribution that these factors make to the predictive accuracy of the resulting classifiers by repeated executions of the learning workflow with varying the factors.

The learning workflow is shown in Fig. 5.3. Given a set of binary-labeled training samples from an expression dataset, the workflow starts by ranking the provided set of a priori-defined gene sets according to their power to discriminate sample classes. The resulting ranked list is subsequently used to select the gene sets which form set-level sample features. Each such feature is then assigned a value for each training sample by aggregating the expressions in the gene set corresponding to the feature. An exception to this pattern is the *None* alternative of the aggregation factor, where expressions are not aggregated, and features correspond to genes instead of gene sets. This alternative is considered for comparative purposes. Figure 5.2 illustrates the resulting sample representation for four combinations of the selection and aggregation

[1] The material is available at `http://ida.felk.cvut.cz/CESLT`.

Dataset	*Genes*	*Class 1*	*Class 2*	*Source*	*Reference*
Adenocarcinoma	14023	8	29	GDS2201	Laiho et al. (2007)
ALL/AML	10056	24	24	Broad institute	Armstrong et al. (2002)
Brain/muscle	13380	41	20	–	Holec et al. (2009b)
Breast tumors	14023	16	27	GDS1329	Farmer et al. (2005)
Clear cell sarcoma	14023	18	14	GDS1282	Cutcliffe et al. (2005)
Colitis and Crohn 1	14902	42	26	GDS1615	Burczynski et al. (2006)
Colitis and Crohn 2	14902	42	59	GDS1615	Burczynski et al. (2006)
Colitis and Crohn 3	14902	26	59	GDS1615	Burczynski et al. (2006)
Diabetes	13380	17	17	Broad institute	Mootha et al. (2003)
Heme/stroma	13380	18	33	–	Holec et al. (2009b)
Gastric cancer	5664	8	22	GDS1210	Hippo et al. (2002)
Gender	15056	15	17	Broad institute	Subramanian et al. (2005)
Gliomas	14902	26	59	GDS1975	Freije et al. (2004)
Gliomas 2	31835	23	81	GDS1962	Sun et al. (2006)
Lung cancer Boston	5217	31	31	Broad institute	Bhattacharjee et al. (2001)
Lung cancer Michigan	5217	24	62	Broad institute	Beer et al. (2002)
Lung cancer – smokers	14023	90	97	GDS2771	Spira et al. (2007)
Melanoma	14902	18	45	GDS1375	Talantov et al. (2005)
p53	10101	33	17	Broad institute	Subramanian et al. (2005)
Parkinson 1	14902	22	33	GDS2519	Scherzer et al. (2007)
Parkinson 2	14902	22	50	GDS2519	Scherzer et al. (2007)
Parkinson 3	14902	33	50	GDS2519	Scherzer et al. (2007)
Pheochromocytoma	14023	38	37	GDS2113	Dahia et al. (2005)
Pleural mesothelioma	14902	10	44	GDS1220	Gordon (2005)
Pollution	37804	88	41	–	Líbalová et al. (2010)
Prostate cancer	14023	18	45	GDS1390	Best et al. (2005)
Sarcoma and hypoxia	14902	15	39	GDS1209	Yoon et al. (2006)
Smoking	5664	18	26	GDS2489	Carolan et al. (2006)
Squamous-cell carcinoma	9460	22	22	GDS2520	Kuriakose et al. (2004)
Testicular seminoma	9460	22	14	GDS2842	Gashaw et al. (2005)

Table 5.1: Number of genes interrogated and number of samples in each of the two classes of each dataset.

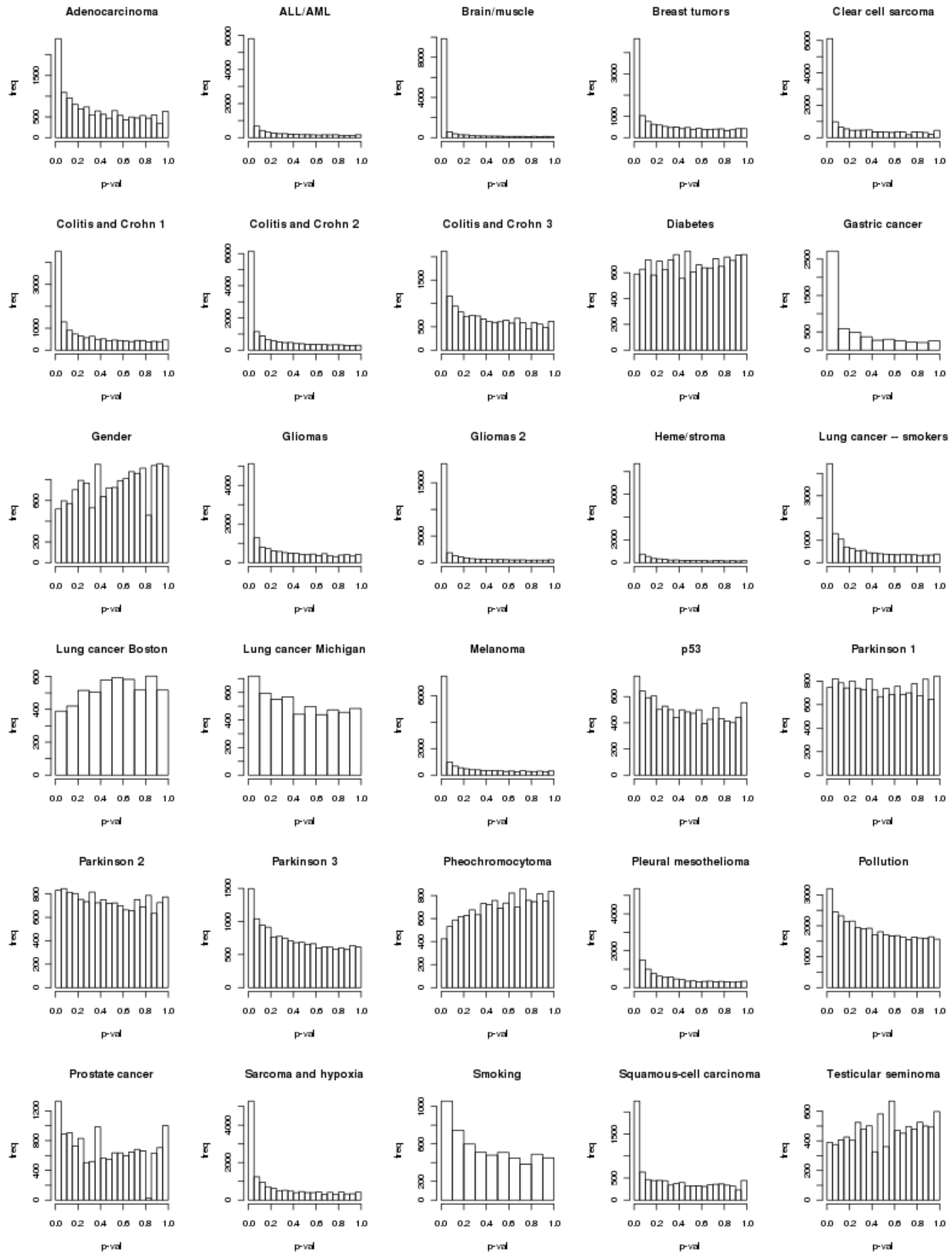


Figure 5.1: Histograms of differential gene expression suggest the difficulty of the individual domains. An easy domain is supposed to have a strongly left-skewed histogram, while the difficult domains rather show a flat histogram. There is one plot for each of 30 domains, x axis shows the p-value of differential expression, the y axis gene frequency.

alternatives. Next, a machine learning algorithm produces a classifier from the reformulated training samples. Finally, the classifier's predictive accuracy is calculated as the proportion of samples correctly classified on an independent testing sample fold. For compatibility with the learned classifier, the testing samples are also reformulated to the set level prior to testing, using the same selected gene sets and aggregation mechanism as in the training phase.

Seven factors along the workflow influence its result. The alternatives considered for each of them are summarized in Table 5.2. We want to assess the contributions of the first four factors (top in table). The remaining three auxiliary factors (bottom in table) are employed to diversify the experimental material and thus increase the robustness of the findings. Factor 7 (testing fold) is involved automatically through the adoption of the 10-fold cross-validation procedure (see, e.g., Chap. 7 in Hastie et al., 2001). We execute the workflow for each possible combination of factor alternatives, obtaining a factored sample of 792,000 predictive accuracy values.

While the measurements provided by the above protocol allow us to compare multiple variants of the set-level framework for predictive classification, we also want to compare these to the baseline gene-level alternative usually adopted in predictive classification of gene expression data. Here, each gene interrogated by a microarray represents a feature. This sample representation is passed directly to the learning algorithm without involving any of the pre-processing factors (1-4 in Table 5.2). The baseline results are also collected using the 5 different learning algorithms, the 30 benchmark datasets and the 10-fold cross-validation procedure (i.e., Factors 5-7 in

<table>
<tr><th><i>F3</i></th><th><i>F4</i></th><th><i>Example row</i></th></tr>
<tr><td>1</td><td>avg</td><td><table><tr><td>Feature 1</td></tr><tr><td>$\text{avg}\{e_1^1, \ldots e_{|\Gamma_1|}^1\}$</td></tr></table></td></tr>
<tr><td>1</td><td>none</td><td><table><tr><td>Feature 1</td><td>...</td><td>Feature $|\Gamma_1|$</td></tr><tr><td>e_1^1</td><td>...</td><td>$e_{|\Gamma_1|}^1$</td></tr></table></td></tr>
<tr><td>1:10</td><td>avg</td><td><table><tr><td>Feature 1</td><td>...</td><td>Feature 10</td></tr><tr><td>$\text{avg}\{e_1^1, \ldots e_{|\Gamma_1|}^1\}$</td><td>...</td><td>$\text{avg}\{e_1^{10}, \ldots e_{|\Gamma_{10}|}^{10}\}$</td></tr></table></td></tr>
<tr><td>1:10</td><td>none</td><td><table><tr><td>Feature 1</td><td>...</td><td>Feature $\sum_{i=1}^{10} |\Gamma_i|$</td></tr><tr><td>$e_1^1$</td><td>...</td><td>$e_{|\Gamma_{10}|}^{10}$</td></tr></table></td></tr>
</table>

Figure 5.2: Examples of sample representation generated with four combinations of alternatives of factors 3 and 4 from Table 5.2. Shown for one sample (i.e. header + one row) with e_i^j denoting the expression of the i-th member of the j-ranked gene set Γ_j. Non-exemplified combinations of the two factors are analogical to the cases shown. The remaining considered factors do not influence the structure of sample representation.

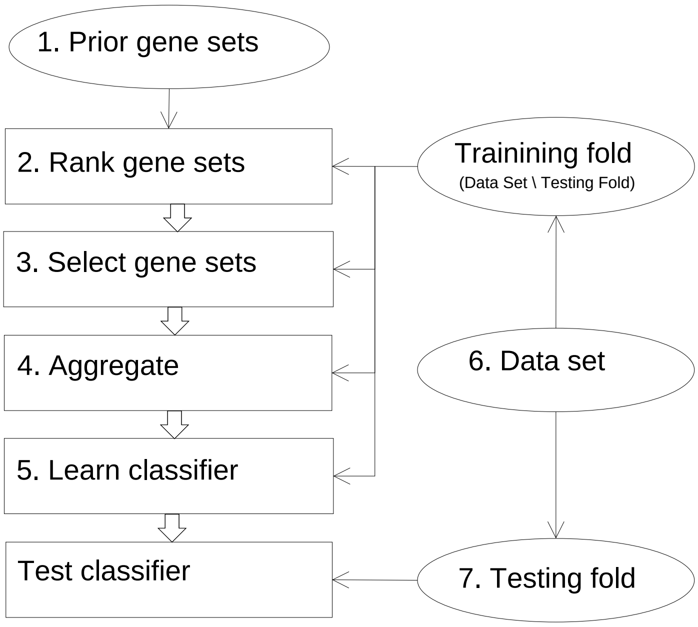


Figure 5.3: The workflow of a set-level learning experiment conducted multiple times with varying alternatives in the numbered steps. For compatibility with the learned classifier, testing fold samples are also reformulated to the set level. The reformulation is done using gene sets selected in Step 3 and aggregation algorithm used in Step 4. The diagram abstracts from this operation.

Table 5.2 are employed). As a result, an additional sample of 1,500 predictive accuracy values is collected for the baseline variant.

Finally, to comply with the standard application of the cross-validation procedure, we average the accuracy values corresponding to the 10 cross-validation folds for each combination of the remaining factors. The subsequent statistical analysis thus deals with a sample of 79,200 and 150 measurements for the set-level and baseline experiments, respectively, described by the predictive accuracy value and the values of the relevant factors.

All statistical tests we conduct are based on the paired Wilcoxon test (two-sided unless stated otherwise). For pairing, we always relate two measurements equal in terms of all factors except for the one investigated. The stronger t-test is more usual in analysis of predictive accuracy samples in literature but our preliminary normality tests did not justify its application. Given the extent of the collected samples, the Wilcoxon test was sufficient to support the conclusions reported. Besides, the Wilcoxon test is argued (Demšar, 2006) to be statistically safer than the t-test for comparing classification algorithms over multiple data sets.

5.3 Main results

We first verified whether gene sets ranked high by the established set-level analysis methods (GSEA, SAM-GS, Global) indeed lead to construction of better classifiers by machine learning algorithms, i.e. we investigated how classification accuracy depends on Factor 3 in Table 5.2. In the top panel of Fig. 5.4, we plot the average accuracy for Factor 3 alternatives ranging 1 to 10 (top 10 gene sets), and $n - 9$ to n (bottom 10). The trend line fitted by the least squares method shows a clear decay of accuracy as lower-ranking sets are used for learning. The bottom panel corresponds to Factor 3 values 1:10 (left) and $n - 9 : n$ (right) corresponding to the situations where the 10 top-ranking and the 10 bottom-ranking (respectively) gene sets are combined to produce a feature set for learning. Again, the dominance of the former in terms of accuracy is obvious.

Given the above, there is no apparent reason why low-ranking gene sets should be used in practical experiments. Therefore, to maintain relevance of the subsequent conclusions, we conducted further analyses on the set-level experimental sample only with measurements where Factor 3 (gene set rank) is either 1 or 1:10.

We next addressed the hypothesis that genuine gene sets constitute better features than random gene sets, i.e. we investigated the influence of Factor 1 in Table 5.2. Classifiers learned with genuine gene sets exhibited significantly higher predictive accuracies ($p = 1.4 \times 10^{-4}$, one-sided test) than those based on random gene sets.

Analyzed factors	*Alternatives*	*#Alts*
1. Gene sets (Sec. 5.2.4)	Genuine, Random	2
2. Ranking algo (Sec. 5.2.1)	GSEA, SAM-GS, Global	3
3. Set(s) forming features[*]	$1, 2, \ldots 10,$ $n-9, n-8, \ldots n,$ 1:10, $n-9:n$	22
4. Aggregation (Sec. 5.2.2)	SVD, AVG, SetSig, None	4
Product		528

Auxiliary factors	*Alternatives*	*#Alts*
5. Learning algo (Sec. 5.2.3)	svm, 1-nn, 3-nn, nb, dt	5
6. Dataset (Sec. 5.2.4)	$d_1 \ldots d_{30}$	30
7. Testing Fold	$f_1 \ldots f_{10}$	10
Product		1500

[*] Identified by rank, n corresponds to the lowest ranking set, i:j denotes that all of gene sets ranking i to j are used to form features.

Table 5.2: Alternatives considered for factors influencing the set-level learning workflow. The number left of each factor refers to the workflow step (Fig. 5.3) in which it acts.

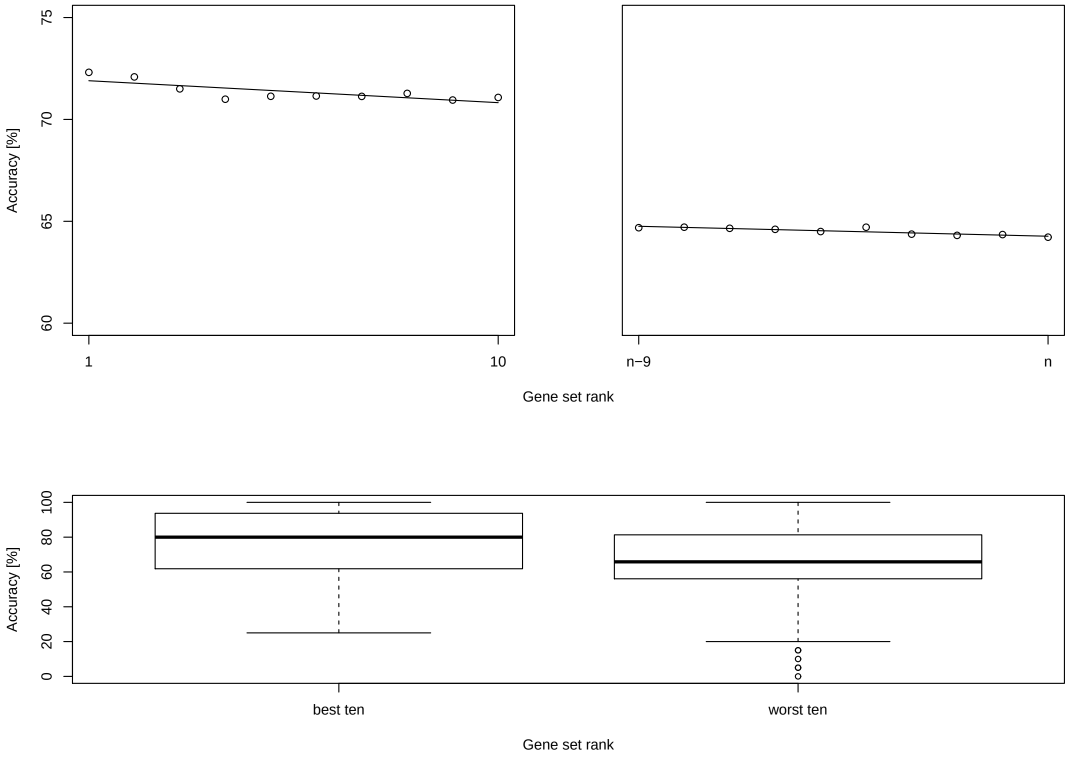


Figure 5.4: The top panels show the plots for the average accuracy of Factor 3 alternatives ranging 1 to 10, and $n - 9$ to n. Average predictive accuracy tends to fall as lower-ranking gene sets are used to constitute features (see text for details). The trend lines shown in the top panels are the ones minimizing the residual least squares. The bottom panel gives the accuracy boxplot for the batch experiments. 10 highest-ranking and the 10 lowest-ranking (respectively) gene sets are combined to produce a feature set for learning. Again, the dominance of the former in terms of accuracy is obvious. Each point in the top panels and each box plot in the bottom panel follows from 16,000 learning experiments.

Given this result, there is a clear preference to use genuine gene sets over random gene sets in practical applications. Once again, to maintain relevance of our subsequent conclusions, we constrained further analyses of the set-level sample to measurements conducted with genuine gene sets.

Working now with classifiers learned with high-ranking genuine gene sets, we revisited Factor 3 to assess the difference between the remaining alternatives 1 and 1:10 corresponding respectively to more and less compression of training data. The 1:10 variant where sample features capture information from the ten best gene sets exhibits significantly (p=3.5×10^{-5}) higher accuracy than the 1 variant using only the single best gene set to constitute features (that is, a single feature if aggregation is employed).

We further compared the three dedicated gene-set ranking methods, i.e. evaluated the effect of Factor 2 in Table 5.2. Since three comparisons are conducted in this case (one per pair), we used the Bonferroni-Dunn adjustment on the Wilcoxon test result. The Global test turned out to exhibit significantly higher accuracy than either SAM-GS ($p = 0.0051$) or GSEA ($p = 0.0039$). The difference between the latter two methods was not significant.

Concerning the aggregation method (Factor 4 in Table 5.2), there are two questions of interest: whether there are significant differences in the performance of the individual aggregation methods (SVD, AVG, SetSig), and whether aggregation in general has a detrimental effect on performance. As for the first question, both SVD and SetSig proved to outperform AVG ($p=0.011$ and $p=0.03$, respectively), while the difference between SVD and SetSig is insignificant. The answer to the second question turned out to depend on Factor 3 as follows. In the more compressive (1) alternative, the answer is affirmative in that all the three aggregation methods result in less accurate classifiers than those not involving aggregation ($p = 0.0061$ for SVD, $p = 0.013$ for SetSig and $p=1.1 \times 10^{-4}$ for AVG, all after Bonferroni-Dunn adjustment). However, the detrimental effect of aggregation tends to vanish in the less compressive (1:10) alternative of Factor 3, where only the AVG alternative in comparison to None yields a significant difference ($p=0.011$). Table 5.3 summarizes the main findings presented above.

Factor	*Alternatives*	
	Better	*Worse*
1. Gene sets	Genuine	Random
2. Ranking algo	Global	SAM-GS, GSEA
3. Sets forming features	high ranking	low ranking
	1:10 (best ten sets)	1 (best set)
*4. Aggregation**	SetSig, SVD	AVG

* Difference not significant if Factor 3 is 1:10.

Table 5.3: See Section 5.3 for details on how the conclusions were determined.

The principal trends can also be well observed through the ranked list of methodological combinations by median classification accuracy, again generated from measurements not involving random or low-ranking gene sets. This is shown in Table 5.4. Position 17 refers to the baseline method where sample features capture expressions of all genes and prior gene set definitions are ignored. In agreement with the statistical conclusions above, the ranked table clearly indicates the superiority of the Global test for gene-set ranking, and of using the 10 best gene sets (i.e., the 1:10 alternative) to establish features rather than relying only on the single best gene set. It is noteworthy that all four combinations involving the Global test and the 1:10 alternative (i.e., ranks 1, 2, 4, 5) outperform the baseline method.

Rank	*Methods*			*Accuracy*			
	Sets	*Rank. algo*	*Aggrgt*	*Median*	*Avg*	*σ*	*Iqr*
1	1:10	Global	SVD	89.2	79.5	18.9	33.2
2	1:10	Global	None	88.3	81.0	17.7	31.3
3	1	Global	None	87.8	80.7	17.5	31.0
4	1:10	Global	SetSig	87.4	81.1	16.5	26.1
5	1:10	Global	AVG	85.6	78.7	18.4	32.6
6	1:10	SAM-GS	SetSig	85.4	79.9	17.1	30.2
7	1:10	SAM-GS	None	84.6	80.1	17.3	30.7
8	1	Global	SVD	83.8	77.9	20.1	34.3
9	1:10	GSEA	SetSig	83.4	78.3	16.7	26.3
10	1:10	GSEA	None	82.3	80.0	16.8	30.4
11	1:10	SAM-GS	SVD	79.9	77.1	18.0	32.1
12	1:10	GSEA	SVD	79.2	77.2	17.7	31.7
13	1:10	GSEA	AVG	79.1	76.4	16.9	31.9
14	1	SAM-GS	None	78.3	76.0	15.3	26.3
15	1	Global	SetSig	77.5	75.9	15.1	23.5
16	1	GSEA	None	76.7	75.6	16.3	29.5
17		*baseline (all genes used)*		75.5	76.6	18.4	33.5
18	1	SAM-GS	SetSig	75.0	74.7	14.2	18.9
19	1	Global	AVG	72.7	73.8	17.6	31.1
20	1:10	SAM-GS	AVG	72.5	73.8	15.9	26.0
21	1	GSEA	SetSig	70.2	72.6	17.0	26.8
22	1	GSEA	AVG	69.6	68.1	12.8	22.4
23	1	GSEA	SVD	69.5	71.9	16.3	28.2
24	1	SAM-GS	SVD	69.0	69.5	15.7	21.3
25	1	SAM-GS	AVG	67.3	67.0	11.4	15.5

Table 5.4: Ranking of combinations of gene set methods by median predictive accuracy achieved on 30 datasets (Section 5.2.4) with 5 machine learning algorithms (Section 5.2.3) estimated through 10-fold cross-validation (i.e. 1,500 experiments per row). The columns indicate, respectively, the resulting rank by median accuracy, the gene sets used to form features (1—the top ranking set, 1:10—the top ten ranking sets), the gene set selection method, the expression aggregation method (see Section 5.2 for details on the latter 3 factors), and the median, average, standard deviation and interquartile range of the accuracy.

While intuitive, rankings based on median accuracy over multiple datasets may, according to Demšar (2006), be problematic as to their statistical reliability. Therefore, we offer in Table 5.5 an alternative ranking of the 19 methods that avoids mixtures of predictive accuracies from different datasets. Here, the methods were sub-ranked on each of the 150 combinations of 30 datasets and 5 learning algorithms by cross-validated predictive accuracy achieved on that combination. The 150 sub-ranks were then averaged for each method, and this average dictates the ranking shown in the table. In this ranking, the baseline strategy improves its rank to Position 5. The superiority of classifiers learned from 10 gene sets selected by the Global test, as formerly noted for Table 5.4, continues to hold in the alternative ranking underlying Table 5.5.

5.4 Additional analyses

5.4.1 Generic feature selection

In the set-level classification framework, gene sets play the role of sample features. Therefore the three gene-set ranking methods (GSEA, SAM-GS, Global) are employed for feature selection conducted in the learning workflow. While the latter three methods originate from research on gene expression analysis, generic feature selection methods have also been proposed in machine learning research (Liu and Motoda, 1998). It is interesting to compare the latter to the gene-expression-specific methods. To this end, we consider two approaches. *Information Gain* (IG) (Mitchell, 1997) is a feature-selection heuristic popular in machine learning. In brief, IG measures the expected reduction in class-entropy caused by partitioning the given sample set by the values of the assessed feature. One of the main disadvantages of IG is that it disregards potential feature interactions. *Support Vector Machine with Recursive Feature Elimination* (SVM-RFE) (Guyon et al., 2002) is a method that ranks features by repetitive training of a SVM classifier with a linear kernel while gradually removing the feature with the smallest input classifier weight. This approach does not assume that features are mutually independent. On the other hand, it naturally tends to select a feature set that maximizes the accuracy of the specific kind of classifier (SVM). For computational reasons (large number of runs and genes), we removed several features at a time ($F \times 2^{-i}$ features in the i-th iteration, where F is the original number of features). Guyon et al. (2002) mention such a modification with the caveat that it may be at the expense of possible classification performance degradation.

In the present context, generic feature selection can be applied either on the gene level or on the set level. We explored both scenarios.

The gene-level application produces a variant of the baseline classifier (position 17 in Table 5.4, position 5 in Table 5.5) where, however, the learning algorithm only receives features corresponding to genes top-ranked by the feature selection heuristic, rather than all measured genes. The selection is thus based only on the predictive power of the individual genes and ignores any prior definitions of gene sets. The

Rank	*Methods*			*Avg Subrank*
	Sets	*Rank. algo*	*Aggrgt*	
1	1:10	Global	None	15.3
2	1:10	Global	SetSig	15.7
3	1	Global	None	16.3
4	1:10	GSEA	None	16.7
5	*baseline (all genes used)*			16.8
6	1:10	Global	SVD	17.0
7	1:10	SAM-GS	None	17.2
8	1:10	SAM-GS	SetSig	17.6
9	1:10	Global	AVG	18.6
10	1	Global	SVD	19.4
11	1:10	GSEA	SetSig	19.9
12	1:10	GSEA	SVD	20.1
13	1:10	SAM-GS	SVD	20.8
14	1:10	GSEA	AVG	22.1
15	1	Global	SetSig	22.2
16	1	SAM-GS	None	23.0
17	1	SAM-GS	SetSig	23.8
18	1	GSEA	None	23.9
19	1	Global	AVG	24.6
20	1:10	SAM-GS	AVG	25.5
21	1	GSEA	SVD	26.7
22	1	GSEA	SetSig	26.8
23	1	SAM-GS	SVD	28.3
24	1	SAM-GS	AVG	30.3
25	1	GSEA	AVG	30.9

Table 5.5: Ranking of all combinations of methods in terms of average subrank. Subranking is done on each of the 150 combinations of 30 datasets and 5 learning algorithms by cross-validated predictive accuracy. Column descriptions are as in Table 5.4.

question of how many top-ranking genes should be used for learning is addressed as follows. We want to make the resulting predictive accuracy comparable to that obtained in the main (set-level) experimental protocol, in particular to the 1 and 1:10 alternatives of Factor 3. The median of the number of unique genes present in the selected gene sets in the 1 (1:10, respectively) alternative is 22 (228). Therefore we experiment respectively with 22 and 228 genes top-ranked by generic feature selection. The results are shown in Table 5.6. Comparing the latter to Tables 5.4 and 5.5, we observe that both variants improve the baseline and in fact produce the most accurate classifiers (IG outperforms the set-level approaches, SVM-RFE is comparable with the Global test). SVM-RFE does not outperform IG in general, but it does so in the special case when SVM is used as the learning algorithm.

While the gene-level application of feature selection results in accurate classifiers, the obvious drawback of this approach is that the genes referred in such produced classifiers cannot be jointly characterized by a biological concept. This deficiency is removed if feature selection is instead applied on the set level, i.e. to rank apriori-defined gene sets. This way, the selection methods essentially become the fourth and fifth alternative of Factor 2 (see Table 5.2) up to the following nuance. While the dedicated gene-set methods (GSEA, SAM-GS, Global) score a feature (gene set) by the expressions of its multiple member genes, IG and SVM-RFE score a feature by the single real value assigned to it, i.e., by the aggregated expressions of the member genes. Therefore, when using the generic feature selection, the aggregation step in the experimental workflow (Figure 5.3) must precede the ranking step. The results of applying IG and SVM-RFE on the set level are shown in Table 5.7. Comparing again to Tables 5.4 and 5.5, both IG and SVM-RFE are outperformed by the Global test (Wilcoxon test, p=0.017).

5.4.2 Successful gene sets

We also explored patterns distinguishing gene sets particularly useful for classification from other employed gene sets sourced from the Molecular Signatures Database. To

# Method	*# Selected Genes*	*Accuracy*				*Avg Subrank*
		Median	*Avg*	*σ*	*Iqr*	
IG	22	90.2	81.5	18.1	30.7	15.0
IG	228	89.8	82.0	17.9	30.3	14.5
SVM-RFE	228	88.3	82.3	16.7	28.5	16.4
SVM-RFE	22	88.0	82.1	17.2	30.4	16.2

Table 5.6: Performance of the baseline classification method equipped with a feature-selection step prior to learning. Features (genes) are ranked by the information gain and SVM-RFE heuristics. The number of selected top-ranking genes (22 and 228, respectively) corresponds to the mean number of unique genes acting in gene sets selected in the 1 and 1:10 (respectively) alternatives of the set-level workflow.

Sets	*Methods*		*Accuracy*				*Avg Subrank*
	Selection	*Aggrgt*	*Median*	*Avg*	σ	*Iqr*	
1:10	SVM-RFE	SVD	88.3	80.6	17.3	33.0	17.6
1:10	IG	SVD	87.0	79.0	18.7	31.6	17.4
1:10	IG	AVG	84.6	78.2	18.6	33.4	18.7
1:10	SVM-RFE	AVG	84.4	79.2	17.1	31.2	19.2
1:10	SVM-RFE	SetSig	82.5	78.7	17.0	31.2	19.4
1	IG	SVD	80.8	76.3	17.7	33.1	22.5
1:10	IG	SetSig	80.0	77.1	17.4	33.2	20.8
1	SVM-RFE	SetSig	71.8	73.7	15.8	26.4	23.3
1	SVM-RFE	SVD	71.5	74.4	17.4	30.3	23.0
1	IG	AVG	70.9	74.0	18.6	33.1	24.1
1	SVM-RFE	AVG	70.8	72.5	15.4	26.6	24.4
1	IG	SetSig	66.2	68.8	16.2	25.0	28.9

Table 5.7: Performance of the set level classification strategy using the information gain and SVM-RFE heuristics for ranking gene sets. Column descriptions are as in Table 5.4.

this end, we defined three groups of gene sets. The first group referred to as *full* comprises the entire set of 3028 gene sets obtained from the database (gene sets containing fewer than 5 or more than 200 genes were discarded). The second group referred to as *selected* consists of the 900 gene sets ranked high (1^{st} to 10^{th}) by any of the three selection methods for any of the dataset. The third group referred to as *successful* is a subset of the *selected* group and contains the 210 gene sets acting in classifiers that outperformed the baseline.

We investigated two kinds of properties of the gene sets contained in the three respective groups. First, we considered the gene set type as defined in the Molecular Signatures Database. The gene sets belonging to the category of chemical and genetic perturbations (CGP) were more frequently *selected* and also more frequently appeared in the *successful* group than the gene sets representing canonical pathways (CP) (full: CGPs 73%, CPs 27%, selected: CGPs 88%, CPs 12%, successful: CGPs 88%, CPs 12%). Second, we considered four possible notions of gene set *size*: i) nominal size (the gene set cardinality), ii) effective size (number of genes from the gene set measured in the dataset), iii) number of PCA coefficients capturing 50% of expression variance in the gene set, iv) as in iii) but with 90% variance. As follows from Table 5.8, the *successful* group contains smaller gene sets than the other two groups, and this trend is most pronounced for the Global test ranking method (Mann-Whitney U test, the *successful* group versus the *full* group, Bonferroni adjustment: Effective size p=0.084, PCA 90% p=0.0039)

Group	*Selection*	*Statistic*	*Nominal size*	*Effective size*	*PCA 50%*	*PCA 90%*
Full	None	mean	71.7±1.7	40.9±0.7	4.4±0.03	16.7±0.14
		median	37.0	28.1	4.1	15.3
Selected	all	mean	62.5±2.7	47.8±1.9	3.8±0.08	15.1±0.35
		median	33.5	27.0	3.4	13.4
	Global	median	32.0	25.5	3.3	12.8
	GSEA	median	34.0	27.0	3.4	13.7
	SAM-GS	median	40.5	28.0	3.7	14.3
Successful	all	mean	56.9±4.4	39.2±2.9	4.3±0.14	14.7±0.56
		median	31.0	21.0	3.9	12.6
	Global	median	22.0	18.5	3.8	11.7
	GSEA	median	37.0	27.5	4.3	14.2
	SAM-GS	median	30.5	22.5	4.0	12.7

Table 5.8: Mean and median sizes of gene sets partitioned into three groups (see Section *Successful gene sets* for details.

5.5 Conclusions

We have established the following main conclusions by executing various experiments on 30 gene expression data classification problems.

1. State-of-the-art gene set ranking methods (GSEA, SAM-GS, Global test) perform reasonably as feature selectors in the machine learning context in that high ranking gene sets outperform (i.e., constitute better features for classification than) those that are low ranking.

2. Genuine curated gene sets from the Molecular Signature Database outperform randomized gene sets. Smaller gene sets and sets pertaining to chemical and genetic perturbations were particularly successful.

3. For gene set selection, the Global test (Goeman and Bühlmann, 2007) outperforms SAM-GS (Dinu et al., 2007), GSEA (Subramanian et al., 2005) as well as the generic information gain heuristic (Mitchell, 1997) and the SVM-based recursive feature elimination approach (Guyon et al., 2002).

4. For aggregating expressions of set member genes into a unique feature value, both SVD (Tomfohr et al., 2005) and SetSig (Mramor et al., 2010) outperform arithmetic averaging (e.g., Holec et al., 2009b).

5. Using the top ten gene sets to construct features results in better classifiers than using only the single best gene set.

6. The set-level approach using the top ten genuine gene sets as ranked by the Global test outperforms the baseline gene-level method in which the learning algorithm is given access to expressions of all measured genes. However, it is

outperformed by the baseline approach if the latter is equipped with a prior feature selection step.

Conclusion 1 is rather obvious and was essentially meant as a preparatory check.

The first statement of Conclusion 2 is not obvious, since constructing randomized gene sets in fact corresponds to the machine learning technique of stochastic feature extraction (Ho, 1998) and as such may itself contribute to learning good classifiers. Nevertheless, relevant prior knowledge resting in the prior definition of biologically plausible gene sets contributes further to increasing the predictive accuracy. Conclusions 3 and 4 are probably the most significant for practitioners in set-level predictive modeling of gene expression as so far there has been no clear guidance for making the right choice.

Concerning Conclusion 3, the advantages of the Global test were argued in Goeman and Bühlmann (2007) but not supported in terms of the predictive power of the selected gene sets. As for conclusion 4, the SetSig technique was introduced and tested in Mramor et al. (2010), appearing superior to both averaging and a PCA-based method which is conceptually equivalent to the SVD method (Tomfohr et al., 2005). However, owing to the limited experimental material in Mramor et al. (2010), the ranking was not confirmed by a statistical test. Here we confirmed the superiority of SetSig with respect to averaging; however, the difference between the performance of SetSig and SVD was not significant.

A further remark concerns the aggregation methods mentioned. All three of them are applicable to any kind of gene set collections, whether these are derived from pathways, the GO or other sources of prior knowledge. The downside of this generality is that substantial information available for specific kinds of gene sets is ignored.

Conclusion 5 is not entirely surprising. Relying only on a single gene set entails too large an information loss and results in classifiers less accurate than those using the ten best gene sets. Note that in the single gene set case, when aggregation is applied (SVD, AVG or SetSig), the sample becomes represented by only a single real-valued feature and learning essentially reduces to finding a threshold value for it. To verify that more than one gene set should be taken into account, we tested the 10-best-sets option and indeed it performed better.

A straightforward interpretation of Conclusion 6 is that the set-level framework is not an instrument for boosting predictive accuracy, and—therefore—we can reject Hypothesis 2 for the used state-of-the-art gene set collections, data, and methods. However, set-level classifiers have a value per se, just as set-level units are useful in the standard differential analysis of gene expression data. In this light, it is important that with a suitable choice of techniques, set-level classifiers do achieve an accuracy competitive with conventional gene-level classifiers.

Chapter 6

A more advanced approach to decreasing dimensionality

(Evaluation of Hypothesis 2 with originally designed gene sets)

This chapter is dedicated to the evaluation of Hypothesis 2 on bacteria, whose hereditary information and structural complexity is much simpler than complexity of prokaryotic cells, we primarily focus here on different specific type of gene set collections with potential to boost predictive accuracy.

The chapter is organized as follows. Section 6.1 specifies background for examination of this hypothesis for the novel gene set collections. In Section 6.2 we give an overview of used data, prior knowledge, and methods. Sections 6.3 and 6.4 provide results and conclude experiments, respectively.

6.1 Background

Given the implied reduction in sample dimensionality, the set-level approach should lead to a decreased risk of overfitting potentially resulting in improved accuracy of induced predictive models. In the previous chapter, we have analyzed this reduction using the traditional gene set collections based on pathways and GO terms, and concluded that we can reject Hypothesis 2 for this type of gene set collections, which is also in concordance with the latest published results (Abraham et al., 2010; Mramor et al., 2010; Staiger et al., 2012). Specifically, it was shown that reducing the dimensionality through a standard feature (gene) selection method applied on the original gene-level representation leads to accuracies higher than those achieved with the gene set representation.

We presume that the lack of predictive accuracy improvements observed in the previous studies was due to the adoption of unsuitable gene set definitions.

To define biologically relevant gene sets, we rely on the following two observations:

1. Expression of individual genes does not correlate very well with the amount of the coded proteins.

2. High correlation can be expected between expressions of genes which share several activating regulatory proteins (transcription factors).

As for the first observation, that there is—generally—only a modest correlation between the amounts of mRNAs on one hand and corresponding proteins on the other hand, this is hypothesised to be mainly due to the presence of post-transcriptional and post-translational modifications and other regulatory interactions (see, e.g., the review in Vogel and Marcotte, 2012). For bacteria, post-transcriptional rather than post-translational regulatory mechanisms control cellular mRNA to protein abundance ratios (Maier et al., 2011). Furthermore, the correlation of the abundance of proteins and the expression measured by microarrays is even smaller because microarrays are not a quantitative platform; however, the measured mRNA expression is highly correlated with the amount of present transcripts (Canales et al., 2006).

The second observation is a rather obvious implication of gene regulatory mechanisms but was also confirmed experimentally in specific studies concerning operons. Operons are contiguous clusters of genes which are transcribed into mRNA as a single unit. Thus, genes in the same operon share some of their regulatory proteins and, by the second assumption above, they should be highly correlated. This is supported, for example, by an improvement of gene expression estimations using operon structure (Xiao et al., 2006) and relatively higher consistence of genes contained in the same operon as opposed to gene groups defined by a common GO term or KEGG pathway membership (Tintle et al., 2012). We recall here that the GO terms and KEGG pathways represents very popular gene set collections previously used for gene-set-level predictive classification (e.g., Abraham et al., 2010; Mramor et al., 2010; Staiger et al., 2012), and this type of prior knowledge is also frequently used in other tasks like gene enrichment analysis (Huang et al., 2009; Líbalová et al., 2012), gene functional clustering (Krejník and Kléma, 2012; Mitra and Ghosh, 2012), and pattern mining (Kléma et al., 2006; Leyritz et al., 2008).

6.2 Methods

This section is organized as follows. First, we explain the gene sets used in our study. Then we briefly describe the machine learning scenarios which we use to assess the performance of the new types of gene sets, and also explain how gene sets are used to aggregate features of gene expression samples. After that, we describe the procedure we conducted to prepare training data. Finally, we describe our simple two-step experimental protocol.

6.2.1 Gene set collections

We introduce several novel types of gene sets which will be based on the following two assumptions. First, the expression of genes which are regulated by the same sets of activating and repressing regulatory genes should be more correlated than random sets of genes. Second, their aggregated expression should be correlated with

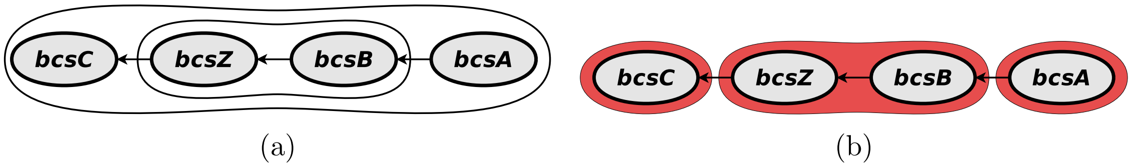


Figure 6.1: Example of operon based gene sets: (a) The operon *bcsABZC* contains genes *bcsC, bcsZ, bcsB, bcsA*, and contains transcription units *bcsABZC* and *bcsBZ*. (b) OPRgs chunks are consecutive set of genes in operons which are always co-transcribed.

the presence of the regulatory gene products in their active form, which, in turn, should be associated with some phenotype. In order to test the extent to which it is necessary to consider sets of genes with exactly the same *regulation*, we define several types of gene sets in which this condition is relaxed in various ways.

There are two possible ways to construct gene sets according to the described logic. The first possibility is to exploit the known operon structure of prokaryotic genome, for which we do not need to know the exact regulatory network. The second possibility is to use information about the regulatory network.

The actual instantiation of the gene sets that we use is derived from the transcriptional regulation network of Escherichia coli K-12 as described in RegulonDB (ver. 8.2) (Salgado et al., 2013).

Gene sets based on the structure of operons

The gene sets based on the operon structure are based on the following biological intuition based on the way DNA is transcribed into mRNA which is specific for prokaryotes. In prokaryotes, genes are organized into contiguous clusters called *operons* which are transcribed into mRNA as a single units. Some operons may also have multiple promoters, possibly located even inside the operon (which means that sometimes only a partial group of genes of the operon may be transcribed).

Operon genes (OPRgs). We can expect genes in an operon to be more correlated than randomly selected genes and, therefore, it makes sense to try to use them as so called *operon* gene sets.

Transcriptional unit genes (TUgs). Similar gene sets based on organization of genes in prokaryotic genomes are *transcription unit* (TU) gene sets which actually include also all operons. A transcription unit is a group of genes transcribed from a single promoter. Unlike operons, transcription units may overlap and one transcription unit can be contained in another transcription unit (Fig. 6.1a). Similarly, as for operons, we may expect the expression of genes in a transcription unit to be more correlated than for a random group of genes.

Non-interrupted subsequences of operons (OPRgs chunks). However, it may still be the case that only a part of a transcription unit is transcribed into mRNA (already because operons are also transcription units). Therefore, we also consider a third kind of this type of gene sets: *operon chunks* which are contiguous subsets of operons such that there is no promoter or terminator located between two genes contained in an operon chunk (Fig. 6.1b). It follows that the expression of genes in an operon chunk should be highly correlated.

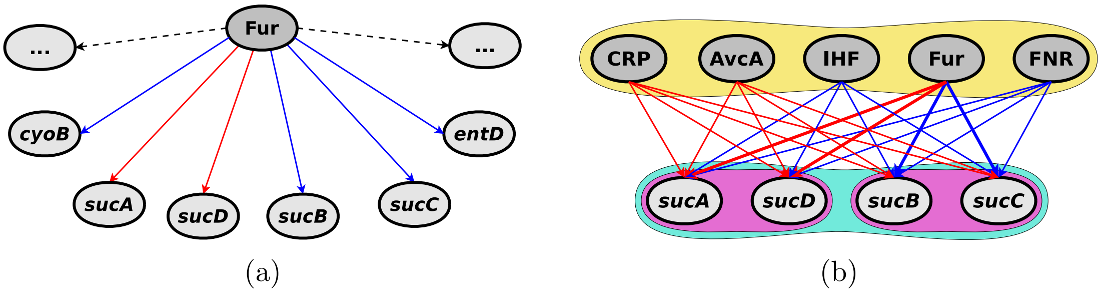


Figure 6.2: Example of transcription factor based gene sets: (a) A transcription factor *Fur* regulates altogether 130 genes including positively regulated genes (e.g., *sucA, sucD*) and negatively regulated genes (e.g., *cyoB, sucB, sucC, entD*); all these regulated genes constitute a gene set. (b) A complex regulon defined by genes *sucA, sucD, sucB, and sucC* can be divided into two strict regulons defined by two pairs of genes *(sucA, sucD)* and *(sucB, sucC)*. The all three mentioned regulons are regulated only by a common set of transcription factors *CRP, ArcA, IHF, Fur, and FNR*.

Gene sets based on transcription factors

The following gene sets are based directly on characterizing groups of genes *with the same regulation*.

Transcription factor regulated genes (TFgs). The simplest of this type of gene sets are *TF gene sets* (Fig. 6.2a) which are composed of sets of genes having a regulating transcription factor in common.

Gene sets based on regulons (REGgs). Another type of gene sets are *Regulon gene sets* which are based on the notion of *regulon* from Maas (1964). A regulon is a set of genes where each of its element is regulated by the same set of transcription factors. Regulon gene sets are gene sets corresponding to regulons. Using description of the regulatory network, we group together genes regulated by the same set of transcription factors, the type of relations is ignored here (Fig. 6.2b).

Strict REGgs (Strict REGgs). The only difference between the regular and *strict* REGgs gene sets is that the latter group only such genes which share the same type of the regulatory interaction with each of the transcription factors (Fig. 6.2b). This formulation is based on the definition of strict regulons from Gutiérrez-Ríos et al. (2003)

One may expect quite reasonably that the regulons and strict regulons perform better than TF gene sets. Let us see why on an example. Let us have genes g_1, g_2 and g_3 which are all regulated by a transcription factor t which is an activator for all of these three genes. This means that g_1, g_2 and g_3 form a TF gene set according to our definition. If t was the only gene regulating any of the three genes then there would be basically no problem and we could expect that the presence of a sufficient amount of proteins coded by t in their active form would imply increased expression of all the three genes. However, if g_1 was also regulated by a repressor gene t_2 and the repressor t_2 was present in its active form then g_1 would not probably have increased expression. This issue is solved by using regulons and strict regulons. There may still be problems when using non-strict regulons as gene sets because non-strict regulons do not take into account the role of the regulatory genes (whether they are activators or repressors). For example, if we wanted to aggregate expression of genes in a regulon by averaging them then we might hit upon a problem similar to the following one. Let us have a regulon corresponding to genes regulated by only one transcription factor t'. Let us suppose that this regulon consists only of two genes g_1' and g_2' where g_1' is repressed by t' whereas g_2' is activated by t'. Now, if t' is present in its active form then we might expect the expression of g_1' to be low and the expression of g_2' to be high. So we cannot expect their average expression to be very correlated with the presence of t' in its active form. However, this potential problem can be solved by using strict regulons instead of non-strict regulons, as strict regulons take the role of the regulatory genes into account.

Gene sets based on KEGG pathways and GO terms. The GO terms and KEGG pathways are often used as specifications of gene sets in order to incorporate prior knowledge into analysis of gene expression data. Here they mainly serve as reference in order to compare with the results reached with the operon-based and TF-based gene sets. We extracted these gene sets from the R package *Genome wide annotation for Escherichia coli strain K12* (version 2.9.0).

Randomized gene sets In addition to gene sets based on organization of prokaryotic genomes or gene sets based on structure of the prokaryotic regulatory networks, we defined also their randomized counterparts (denoted by the prefix *Fake*). For a given type of gene sets (e.g., transcription-unit gene sets), genes in the randomized gene sets are shuffled among all gene sets, thus, proportions of the gene sets remain unchanged. The reason why we introduce these fake gene sets is to see whether the knowledge present implicitly in our gene sets is the actual factor influencing accuracy of the classification method exploiting these gene sets. If randomly selected gene sets with the same cardinalities performed equally well or even better then this would mean that what determines accuracy of the method the most is not the implicitly captured information about structure of the regulatory network.

	# sets	# genes	Min.	Median	Mean	Max.
baseline on TU genes		4524				
baseline on TF genes		1685				
baseline on GO+KEGG genes		2734				
TUgs	3213	4524	1	1	1.685	16
OPRgs	2649	4524	1	1	1.708	16
OPRgs chunks	3164	4524	1	1	1.430	12
GO+KEGG	260	2734	1	12	31.830	847
TFgs	186	1685	1	7	24.720	534
REGgs	459	1685	1	2	3.671	61
Strict REGgs	541	1685	1	2	3.115	51

Table 6.1: Gene set collection types properties.

6.2.2 Learning and set-level aggregation methods

In order to evaluate usefulness of the gene sets that we propose in this chapter, we performed machine learning experiments in which a machine learning algorithm was used to learn a classifier for predicting phenotype from measured gene expressions. We used support vector machine learning algorithm (Cortes and Vapnik, 1995). For a support vector machine, samples are viewed as points in a vector space with coordinates given by the values of the sample's features. A classifier is sought in the form of a hyperplane that separates training samples of distinct classes and maximizes the distance to the points nearest to the hyperplane (i.e., maximizing the margin) in that space. We used implementation from the *R package e1071, version 1.6-1.*

Normally, features of a sample would be the expressions of the individual genes. However, in set-level methods, features are aggregates of expressions of genes in prespecified gene sets. Thus, an important component of set-level methods is data aggregation which computes a single real number representing the *aggregated* expression of genes in a gene set, i.e., the input to an aggregation procedure is a vector of expressions of genes in a gene set and its output is one real number which *should represent* the expression of the genes in the gene set, similarly like in the previous chapter. There are many aggregation methods (Section 2.6). Due to very low number of samples in the datasets with which we work, we represent the aggregated expression of a gene set by *arithmetic average.* We should also note that aggregation using arithmetic average is very well suited for the types of gene sets that we defined because the expression of genes in these gene sets is expected to be positively correlated

A technique which is often used to reduce the number of features used by learnt classifiers is *feature selection.* In principle, feature selection may increase predictive accuracy but it may also lead to decorrelation of the error estimated by cross-validation and the true error (Hanczar et al., 2007). We performed experiments with feature selection in which features are either expression of individual genes or aggregated expressions of gene sets or a combination of both. We used SVM-RFE feature selection (Guyon et al., 2002) which is especially suited for use in conjunction with support vector machines. We used the faster version of SVM-RFE which is also described in Guyon et al. (2002). When performing the leave-one-out cross-validation, the feature selection procedure was always performed only on the training folds.

6.2.3 Gene expression data and prior knowledge

The results presented here are based on a large experiment involving more than 70 small microarray gene expression datasets measured in the bacteria Escherichia coli. We selected this popular model organism for the following reasons. First, it is estimated that about $^2/_3$ of its transcriptional regulatory proteins and most of their targets are already known and described in the publicly available database RegulonDB[1] (Salgado et al., 2013). Second, there is a reasonable number of gene-expression datasets for Escherichia coli available in the Gene Expression Omnibus database (Edgar et al., 2002). Third, organization of genomes and transcriptional networks is much simpler in prokaryotic organisms such as Escherichia coli than in eukaryotic organisms and, therefore, it seems more reasonable to try to devise new methods for prokaryotic organisms and then to try to extend them to eukaryotes rather than the reverse way.

We downloaded 10 largest series of gene expression data for Escherichia coli K12 with as many measured samples as possible (as available in the GEO in April 2013). To have homogeneous data, we limited ourselves to Affymetrix microarray platforms only; particularly, *GeneChip® E. coli Antisense Genome array* and *GeneChip® E. coli Genome 2.0 array*. The series were downloaded from the GEO (see Table 6.2 for series identifiers). Two of the series could be possibly confounding, because they were used for the development of the RegulonDB; therefore, we excluded them. We checked that the remaining series were not used in the development of RegulonDB.

Each of the series contain samples corresponding to several phenotypes (Table 6.2 for number of phenotypes in the series). In order to obtain datasets for learning problems with two classes, we generated a set of *non-overlapping datasets* from the available series in the following way. For each of the series, we considered all pairs of phenotypes as candidate datasets. From this basic pool of candidate datasets we randomly selected 71 datasets which have unique phenotypes (no phenotype appears more than once in these 71 datasets and thus no sample can appear in two different datasets), we call this set of datasets *non-overlapping datasets* (N/O datasets). We also generated an *auxiliary* set of 100 possibly overlapping datasets (AUX) not contained in the N/O datasets which is used only for facilitating unbiased preselection of gene set types to be used on the final datasets. Due to the nature of the data, all datasets contain only 6-10 samples.

[1] 201 from the 314 predicted (Perez-Rueda and Collado-Vides, 2000) regulatory proteins are currently available.

In order to see how the tested methods perform on different kinds of datasets, depending on the number of genes which are critical for the phenotypes, we also divided the datasets in the N/O datasets into two groups: datasets which contain a phenotype induced by knockouting some genes (KO datasets) and datasets which do not contain any such phenotype-class (non-KO datasets). It turns out that all the datasets in the set N/O which contain knockouted genes originate from the series GSE6836.

Series id	Platform id	# phenotypes
GSE6836	GPL199	62
GSE33147	GPL199	30
GSE10160-1	GPL199	9
GSE10160-2	GPL3154	4
GSE35371	GPL3154	20
GSE21869	GPL199	5
GSE17505	GPL199	10
GSE34023*	GPL3154	7
GSE7398*	GPL199	8
GSE4778	GPL199	4

Table 6.2: List of used gene series. The series marked with * were omitted due their use in the RegulonDB.

6.2.4 Experimental protocol

The experimental protocol used in this chapter consists of two steps. In the first step, we pre-select the best performing gene set types—one type from the three types of operon-based gene sets and one type from the three regulatory-network-based gene sets. For the pre-selection, we use average accuracy estimated by leave-one-out cross-validation (LOO CV) on the datasets in the set AUX. Then, in the second step, we evaluate performance of the pre-selected gene sets on the datasets from the set N/O. Note that the datasets contained in AUX and N/O are independent as they do not overlap. The rationale for this two-step procedure is that the available datasets are insufficient for statistically significant comparison of multiple gene set types. This is resolved by the two-step procedure because it enables us to evaluate statistical significance of the results only for the gene sets which are selected in an unbiased way on a different set of datasets. As a consequence, no overoptimistic bias should be introduced by our evaluation procedure, but higher statistical power can be achieved than what we could expect to obtain if, for instance, we compared all gene set types using the Friedman's rank-sum test.

All statistical tests performed were based on the one-sided paired Wilcoxon test (unless stated otherwise). For evaluation of three or more methods to get preference rank which method behaves better than the others, we use sum of ranks from the

Friedman's rank-sum test. This approach is also utilized for a graphic representation of results.

6.3 Results and discussion

We performed experiments in which we assessed accuracy of classifiers based on the various gene sets and also of classifiers based on individual genes. We performed experiments using complete feature sets and experiments using feature-selection techniques to select a handful of informative features. The results presented in this section clearly indicate that carefully defined gene sets do not decrease accuracy and are distinguishable from their fake version as was the case for gene sets which were previously studied in literature (Abraham et al., 2010; Mramor et al., 2010; Staiger et al., 2012, 2013). In fact, accuracy is improved in most cases by use of gene sets either alone or in combination with individual genes in the case when feature selection is used.

6.3.1 Preselection of gene sets on independent data

First, we performed the preselection of the gene set types which should be used in the next step for statistical evaluation, in order to retain reasonable statistical power of the tests. Recall that otherwise there would be several gene set types to compare for both transcriptional-network-based gene sets and for the operon-based gene sets, which would force us to use tests with significantly less power such as the Friedman's rank-sum test. This might be acceptable if the datasets were not so small and high-dimensional at the same time and if high fraction of them was not trivial in the sense that most classifiers obtain 100% accuracy on them, which is unfortunately the case for the available datasets we use.

For the preselection, we used sum of ranks obtained by the classifiers based on the specified gene set types. The results obtained on the set AUX consisting of 100 datasets are shown in Table 6.3. As the result of this auxiliary experiment, we selected OPRgs chunks as the representatives of operon-based gene sets and REGgs as the representatives of transcription-network-based gene sets. Although the differences in average accuracies in Table 6.3 may seem very small, it is necessary to recall that the averages are computed over 100 datasets. Moreover, all three methods obtained 100% accuracy in 56 out of 100 cases for the operon-based gene sets and in 51 out of 100 cases for the transcription-network-based gene sets. This explains why the obtained differences among average accuracies are rather small.

6.3.2 Classification on complete set of features

We evaluated predictive accuracy of the classifiers based on gene set types (OPRgs chunks and REGgs) preselected on the set of datasets AUX on the independent set of 71 non-overlapping datasets (N/O). We compared them against their randomized ("fake") versions and against classifiers built on the complete sets of genes. When comparing against classifiers built on complete sets of genes, we used all genes in the

gene set type	mean accuracy [%]	sum of ranks
TUgs	82.17	198.50
OPRgs chunks	**83.00**	**205.50**
OPRgs	81.50	196.00
REGgs	**82.67**	**211.50**
Strict REGgs	82.50	209.50
TFgs	78.69	179.00

Table 6.3: Preselection results. Table contains results on the the 100 auxiliary datasets; columns contain mean accuracy and sum of ranks over the datasets, higher rank indicates better performance. Here, the best ranked units are *OPRgs chunks* and *REGgs*.

case of OPRgs chunks and those genes for which at least one regulatory transcription factor is known in the case of REGgs. The rationale for this is to provide the same amount of information to both competing methods under comparison (note that every gene of Ecoli K12 is associated to an operon which is why we use all genes in the case of operon-based gene sets).

We compared the results of classifiers based on OPRgs chunks and REGgs gene sets to the classifiers based on the respective full sets of individual genes. We used the one-sided paired Wilcoxon test. As a result, we got that classifiers based on both types of gene sets are significantly better than the classifiers based on *all* individual genes (REGgs $p = 0.038$ and OPRgs chunks $p = 0.018$). We also compared results obtained by classifiers based on OPRgs chunks and REGgs gene sets against classifiers based on the respective randomized counterparts of these gene sets. Both OPRgs chunks and REGgs turned out to be significantly better than their randomized counterparts (REGgs $p = 0.019$ and OPRgs chunks $p = 0.030$). It's also worthy of notice that, despite only one third of all genes appear in at least one transcription factor, the performances on the all genes and all transcription-factor-related genes are also indistinguishable ($p = 0.635$, two sided test).

At first sight, the obtained results may seem to be in disagreement with the results obtained by Mramor et al. (2010) or in Chapter 5) where gene-set-based classifiers which represented the aggregated expression of genes by a single number did not perform better than classifiers built on the complete set of genes whereas they perform significantly better in our present work. However, there are two differences between our approach presented here and the work of Mramor et al. or in the previous chapter. First, we work with gene expression of bacteria whereas the two mentioned studies used human gene expression data. Second, the gene sets which we use are constructed so that the genes in them would be correlated and so that the average expression of these genes would be correlated also to presence of respective regulatory proteins in their active form. The gene sets used in the previous studies were based on pathways from the KEGG database and on gene sets from the GO (GO+KEGG). In order to check whether the better results which we obtained indeed follow as a consequence from higher quality of our gene sets, we performed also experiments with the gene sets from KEGG and the GO. Classifiers based on these gene sets turned out to be significantly worse in comparison to classifiers based on all genes which appear in at least one gene set ($p = 0.003$) and indistinguishable to their randomized counterpart ($p = 1$, two sided test). This suggests that the reason why gene-set-based techniques did not perform well in some of the previous works is that the gene sets which they used consisted of genes which were not correlated. One could still argue that the superiority of OPRgs chunks and REGgs over GO+KEGG is mainly caused by the fact that the average sizes of OPRgs chunks or REGgs are small in comparison to average sizes of gene sets in GO+KEGG, however, then we would expect at least that GO+KEGG would achieve higher predictive accuracy than their randomized (fake) counterparts, which was not the case. This leads us to believe that carefully selected gene sets may increase predictive accuracy of machine learning methods while badly selected gene sets may decrease it substantially.

6.3.3 Classification on selected the top n features

We also evaluated the utility of the operon-based and transcription-factor-based gene sets in the context of predictive classification where feature selection is used to select features for the classifier learning phase. First, we performed experiments in which we evaluated predictive accuracy of SVM classifiers learnt on the features selected from the complete set of genes as a function of the number of features. The average accuracies are shown in Figure 6.3 for all datasets in the set N/O, for the KO datasets (the datasets from N/O where at least one class contains knock-outed genes) and for the non-KO datasets (datasets with no knock-outed genes). We can see that the highest accuracy is obtained with just a handful of features (genes) in the case of KO datasets whereas a bigger number of features is needed to obtain the maximum accuracy on average in the case of non-KO datasets. This is in accordance with the reasonable assumption that a change in expression of genes induced by knock-outing a gene should be detectable either from the expression of the gene itself or from the expression of just a few genes regulated by the respective knock-outed gene. Since the phenotype-classes in the non-KO datasets usually correspond to reactions to changes in the environment, it is also natural to expect that the reaction should often exhibit itself through change in expression of a bigger number of genes.

We performed experiments in which we used the same protocol with feature sets consisting entirely of the gene sets (either OPRgs chunks or REGgs) or of a combination of the gene sets and individual genes (called hybrid OPRgs chunks and hybrid REGgs, respectively). We also performed these experiments with the gene sets from GO and KEGG. The results are displayed in Figure 6.4. We can see several trends. First, addition of individual genes improves accuracy for smaller numbers of selected features which can be seen from the better performance of hybrid OPRgs chunks and hybrid REGgs. This could be expected especially in the case of KO datasets where the gene sets may be sometimes too coarse-grained and the expression of certain individual genes may be therefore expected to improve the predictive accuracy. Second, gene sets alone tend to dominate over individual genes and over the combination of in-

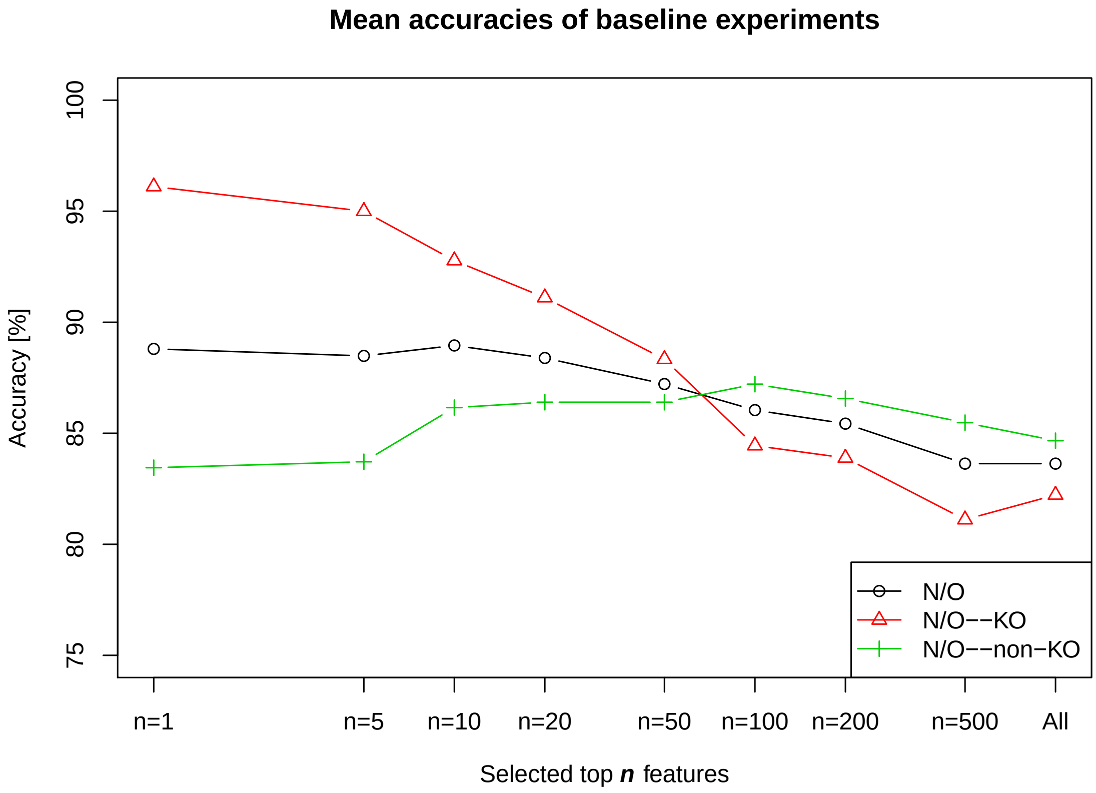
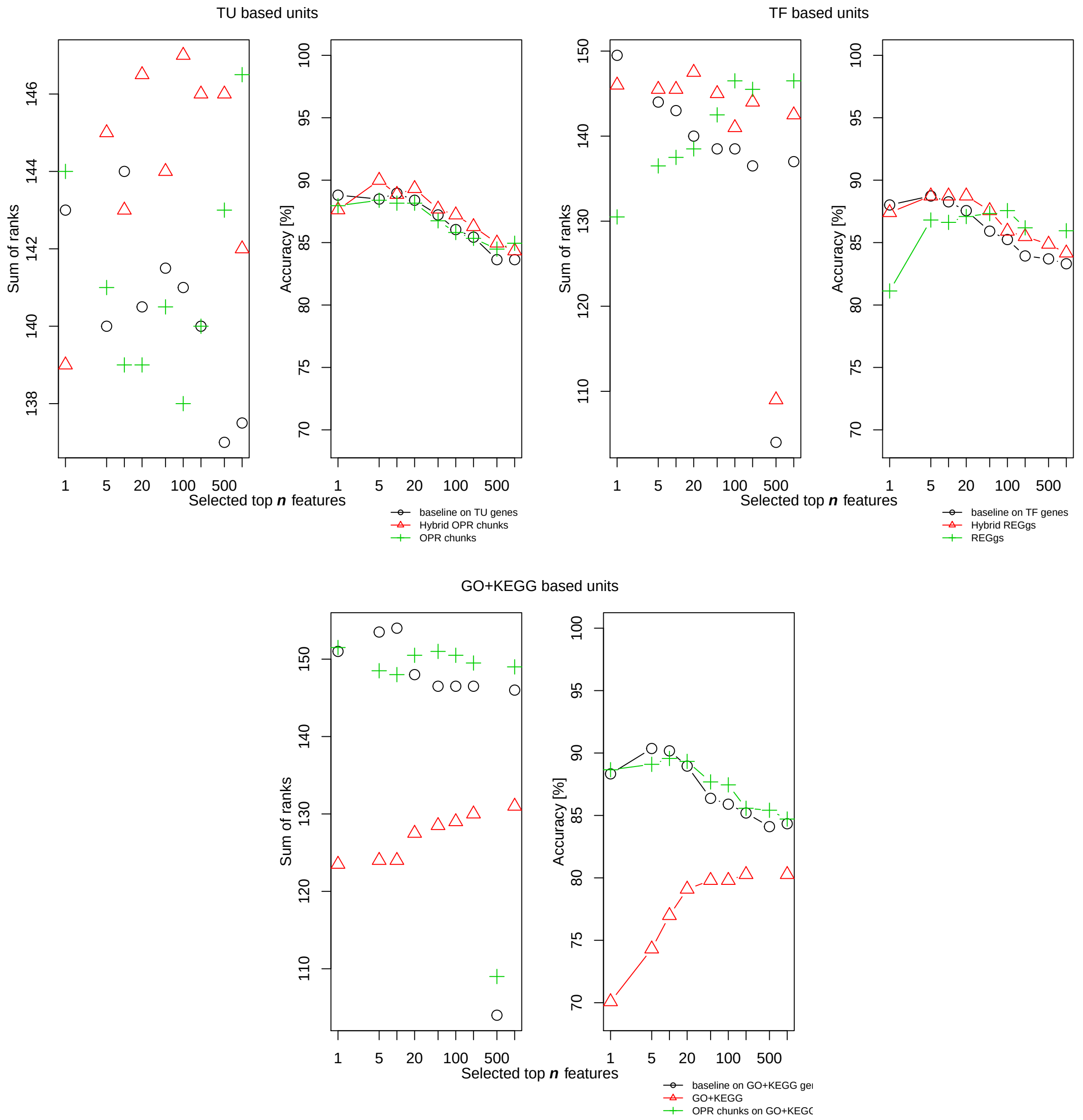


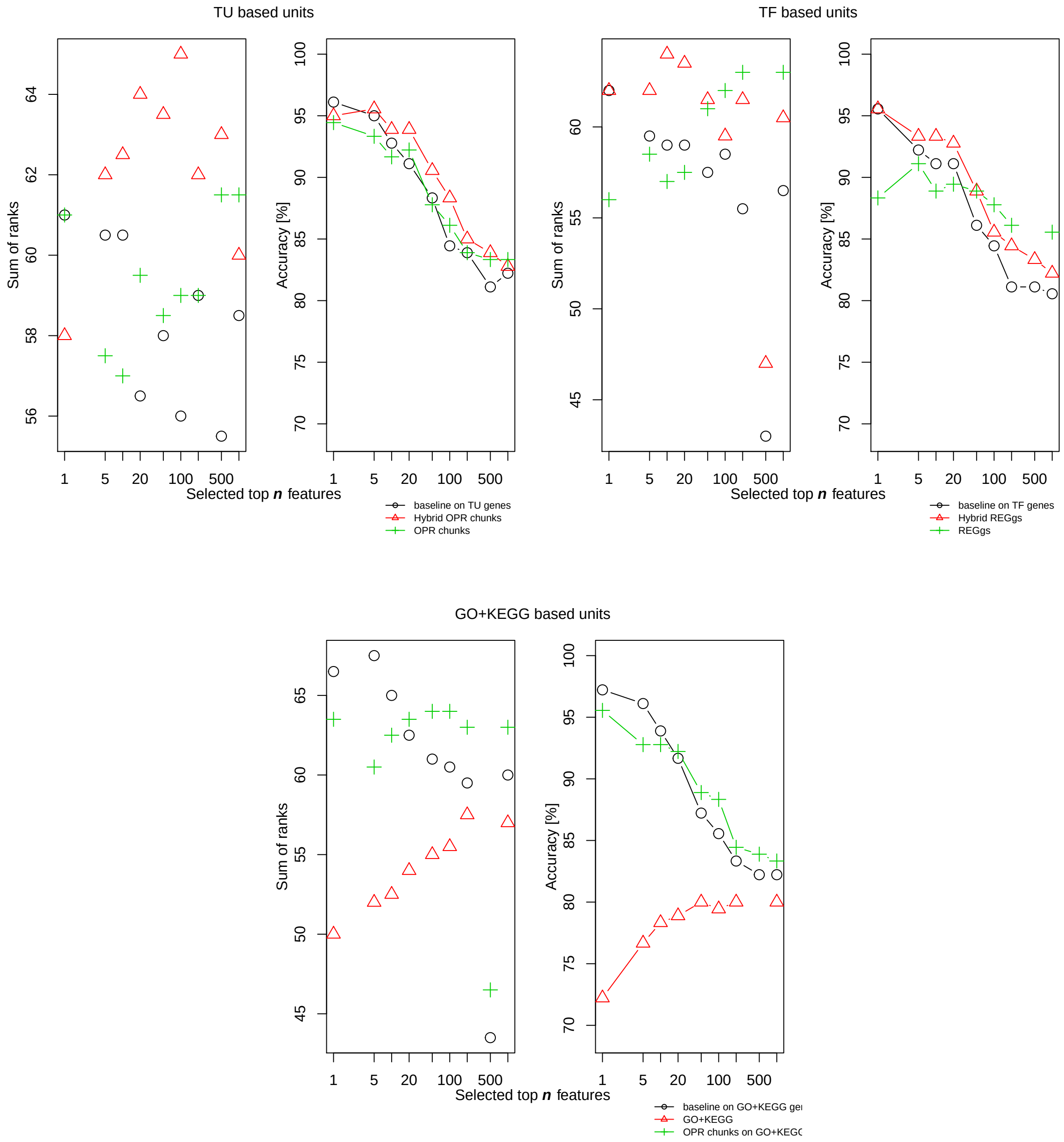
Figure 6.3: The N/O set of datasets is divided into two subsets, the first, *N/O—KO*, contains N/O datasets from the series GSE6836 only, and the second, *N/O—non-KO*, contains the rest of the datasets. According the series annotation, only the series GSE6836 contain samples with knock-outed genes.

dividual genes when the number of selected features is higher. This can be attributed to the fact that the overall number of our gene sets is smaller than the number of individual genes and therefore there is a somewhat smaller risk of overfitting stemming from feature selection on a large set of features. Third, the accuracies obtained by classifiers based on GO and KEGG gene sets tend to be substantially worse than the accuracies obtained by either the classifiers based on selected individual genes or selected OPRgs chunks. In the latter case, the OPRgs chunks were constructed only from genes which were contained in at least one gene set from GO or KEGG in order to allow fair comparison so that the OPRgs chunks would not use more information than available to GO+KEGG gene sets. The results for GO+KEGG agree to the findings reported previously in literature (Tintle et al., 2012). When we compare them to the results for OPRgs chunks and REGgs gene sets, we can see that what matters is the quality of the gene sets, i.e., how much genes in the given gene sets are correlated.

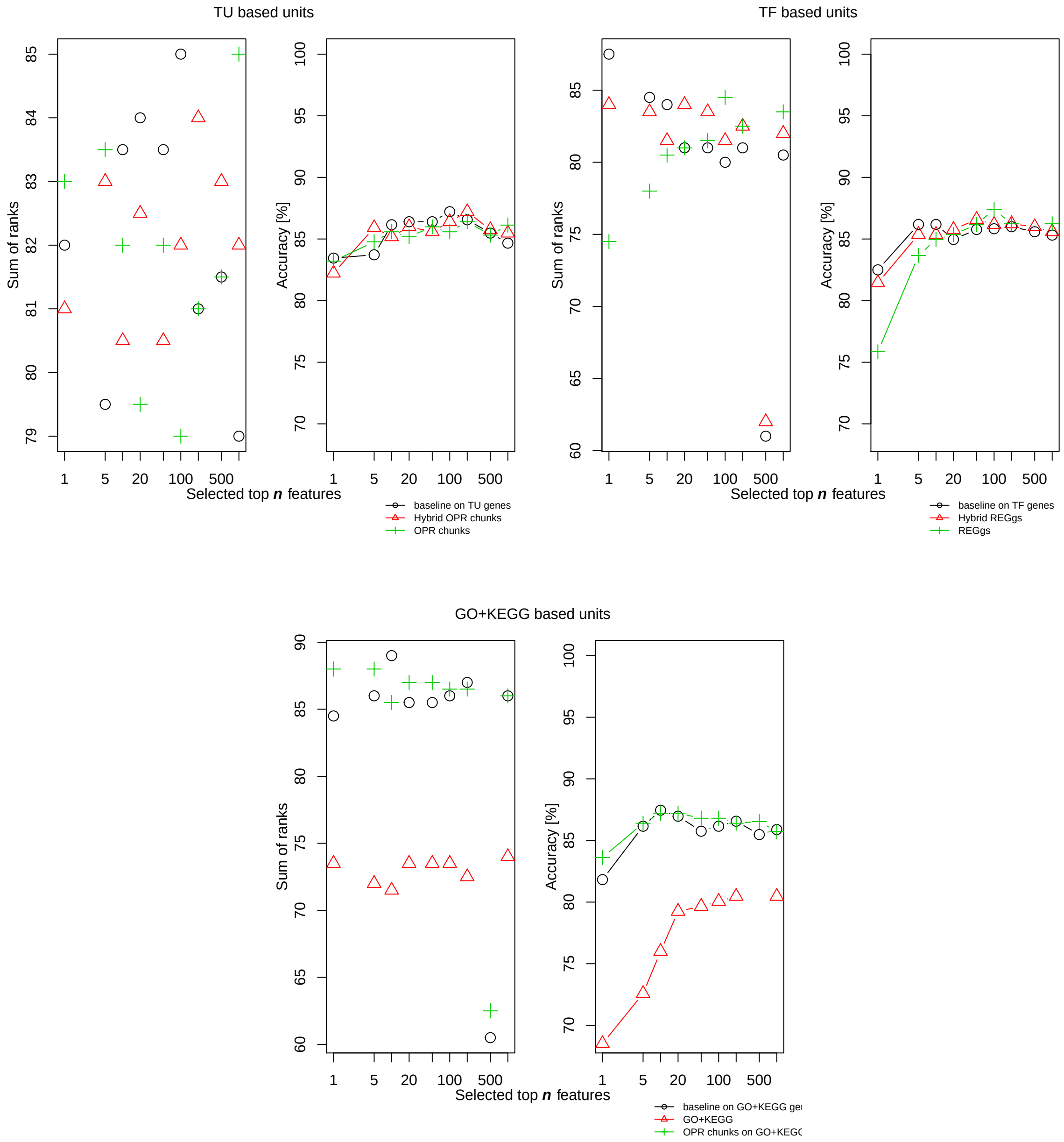
Figure 6.4: Plots of sum of ranks (left panels) and mean accuracy plots (rights panels) on different types of datasets. Note: While the mean accuracies in the (a) can be computed as the average of (b) and (c), this property does not hold for the plots of sum of ranks.



(a) Results on N/O datasets.



(b) Results on N/O—KO subset of datasets.



(c) Results on N/O—non-KO subset of datasets.

gene set type		N/O	N/O—GSE6836	N/O except GSE6836
	ORPgs chunks	4-0-5	7-0-2	3-2-4
	REGgs	4-0-4	4-0-4	4-0-4
Hybrid	ORPgs chunks	7-1-1	8-0-1	5-1-3
Hybrid	REGgs	8-0-1	8-1-0	6-0-3

Table 6.4: An overview of performance from the Figure 6.4. Each value in the table denotes a triplet, *higher-equal-lower*, with performance of a gene set type against its baseline.

6.3.4 Correlation of gene sets

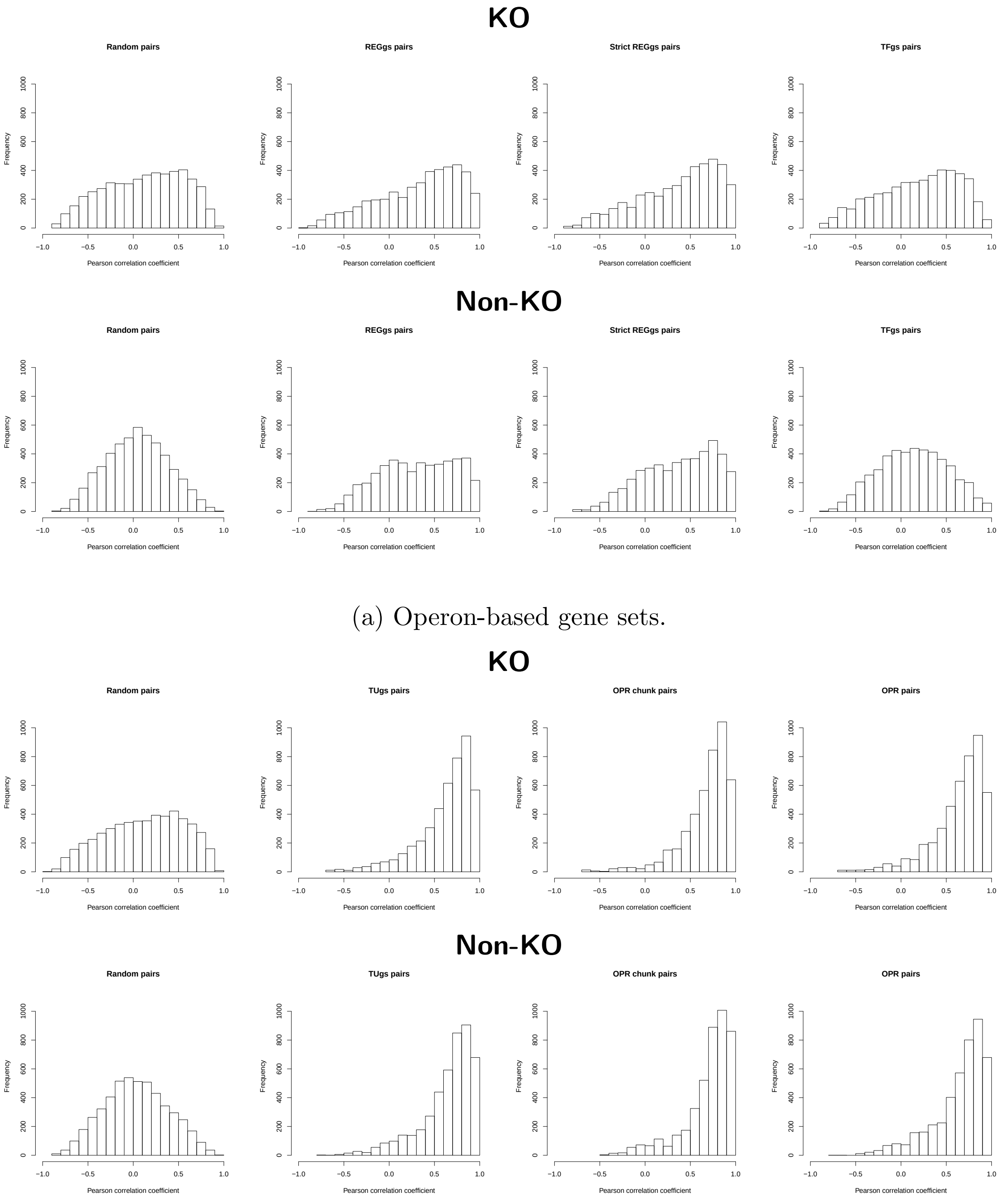
Since we believe that the main reasons for the observed good performance of our gene sets based on operon structure of prokaryotic genomes or on the structure of regulatory networks are: (i) correlation of the average expression of genes in the gene sets with the presence of the respective regulatory proteins in their active form in the bacteria and (ii) the correlation of genes in a gene set. Since the extent to which the first reason is valid is not directly measurable without expensive experiments, we at least assessed validity of the second claim. The correlations are generated in the following way. For the random correlations—without considering any gene set—firstly, we select a gene expression series (note the samples are in columns and rows are identified by genes), subsequently, we randomly chose n pairs of rows of the series without replacement, where $n = 5000$, and compute n Pearson's correlation coefficients for the selected pairs. For the correlations defined on gene sets, the following two-step process on the same data as above is repeated n times: (i) We randomly select a gene set, s, with probability corresponding to number of 2-combinations of its size. (ii) Using the selected gene set, s, we randomly select pair of two distinct genes and use them for computing their correlation coefficient. In order to avoid bias to small gene sets, we omit single-gene sets.

On Figure 6.5 are the computed correlations for the two largest series where the first represents KO datasets and the second non-KO datasets. Both our preferred gene set types, REGgs and OPRgs chunks, show better than random correlation, but this is not the case of the GO+KEGG gene sets (Fig. 6.5C) where even smaller gene sets do not show similar correlation like the REGgs and OPRgs chunks gene set types.

6.4 Conclusions

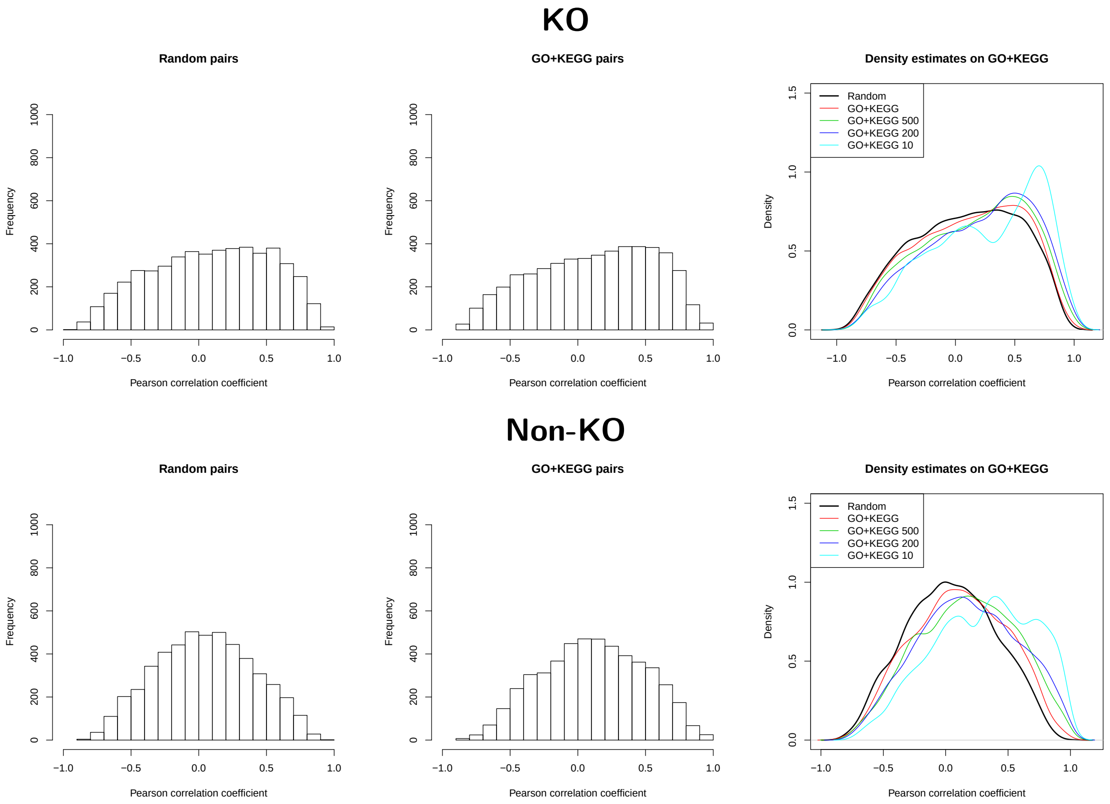
We evaluated the performance of gene sets based on the structure of transcription-regulation networks and on the operon structure of bacterial genomes using machine learning and gene set aggregation. All the gene sets are new in the context of predictive classification. For classification using the all-features scheme, we conclude that using prior knowledge in the form of gene sets can significantly improve predictive accuracy. This finding is not in contradiction with the conclusion of the paper by Mramor et al. (2010) as the gene sets used here differed from the gene sets used by

Figure 6.5: Correlation histograms (and density estimates) of randomly chosen gene pairs with and without considering the used gene set types. The KO data are represented by the series GSE6836 and the non-KO data by the series GSE33147.



(a) Operon-based gene sets.

(b) TF-based gene sets.



(c) GO+KEGG-based gene sets. Note the density estimates on GO+KEGG gene sets limited to maximally 10, 200, and 500 genes per each gene set).

Mramor et al. When we tested the performance of the set-level method with the same type of gene sets used by Mramor et al., we obtained worse results than when no gene sets were used which is in agreement with not only Mramor et al. but also other recent works (Staiger et al., 2012, 2013). For the feature selection case, we cannot assert significantly the dominance of the novel gene sets, or the combined versions where gene sets and individual genes were used together over the baseline due to a very high variance of accuracy typical for this type of experiment and a very low power of available statistical tests for the case when we have multiple statistically dependent points on every plot of sum of ranks. However the hybrid gene sets—despite their higher susceptibility to overfitting—perform better than individual genes for most of the points on the plots of sum of ranks.

The main conclusion is that methods based on an aggregation of gene sets are able to improve predictive accuracy when provided with suitable gene sets. When inappropriate gene sets are used, e.g., when one uses GO terms or KEGG pathways, then the accuracy may actually drop significantly. Therefore, we concede Hypothesis 2 when suitable gene set collection is chosen.

Chapter 7

Software implementation of the set-level methods

Here we present two public freely available tools for gene-expression data analysis with prior knowledge. This chapter is organized as follows. Section 7.1 describes an older tool, XGENE.ORG, and Section 7.2 describes a new tool, miXGENE, which is currently in development. In the future, the latter tool should also gain most of the XGENE.ORG functionality.

7.1 XGENE.ORG tool: a basic description

The web application, XGENE.ORG, is a tool for cross-platform cross-genome analysis of gene expression data with several predefined types of prior knowledge in form of gene sets collecitons.

7.1.1 Introduction

The tool XGENE.ORG (available at `http://xgene.org`) is designed particularly for integration of large volumes of raw gene expression measurements with another huge body of available genomic information in form of gene set collections. XGENE.ORG offers additional functionality resulting from a data-fusion strategy based a priori defined gene sets (similar to the described in Section 4). In particular, the main resulting feature of the present tool is that it enables to analyze gene expression data collected from heterogeneous platforms in an integrated manner. The heterogeneous platforms may pertain to different organism species. The significance of this contribution is at least twofold. First, microarray experiments are costly, often resulting in numbers of samples insufficient for reliable modeling. The possibility of systematically integrating the experimenter's data with numerous public expression samples coming from heterogeneous platforms, would obviously help the experimenter. Second, such integrated analysis provides the principal means to discover biological markers shared by different-genome species.

XGENE.ORG explicitly implements various set-level features (*working units*) and determines their activity score. The activity of a superior (more abstract) working unit is calculated from the known (measured) activity of a set of inferior (less general) working units. For example, it selects all the probesets that are annotated by the same gene identifier and computes gene activity. Likewise, all the genes whose products act in a single pathway are used to compute pathway activity. XGENE.ORG works with a various types of working units on different levels of generality, and use them to perform cross-genome and cross-organism analysis as there are working units that generalize beyond individual platforms and species. Furthermore, in addition to standard statistical analyses, it applies machine learning techniques to develop interpretable models that distinguish among user-defined classes.

Let us exemplify some of the available types of working units. The first type that enables cross-platform analysis aggregates measurements that share a common GO term. The second type aggregates measurement units acting in the same biological pathways formalized by the KEGG database. The third type represents is based on the notion of the fully coupled flux.

To sum up, analyses and models based solely on *measurement units* defined by the individual probesets whose expression is immediately measured by microarrays suffer from the inherent microarray noise and often fail to identify subtle patterns, give a large room to overfitting and prove hard to interpret and apply. Genomic prior knowledge makes it possible to introduce and analyze alternative working units that avoid the bottlenecks mentioned above and provide improved interpretation power and statistical significance of analysis results. At the same time, different platforms and/or species deal with different sets of measurement units that cannot be directly matched. Consequently, multi-platform analyses cannot be performed without working units whose meaning is general enough to be defined in each platform and whose activity can unambiguously be evaluated in each sample independently of its platform type. Working units then serve as markers (or features) to distinguish between user-supplied sample classes.

7.1.2 System description

The main goal of XGENE.ORG tool is to analyze a wide range of publicly accessible heterogeneous gene expression samples. The tool provides an interface to search available measurements whose annotation is relevant to the studied biological topic. Typically, a set of relevant measurements straddles various microarray platforms and organisms. There are two principal reasons to allow for their integration (Chapter 4). The technical reason concerns the sufficiency of sample sets for reliable modeling. The more platforms accessed, the larger number of samples is at hand. The scientific reason pertains to the relevance of the outcomes. Combining multi-platform input data contributes to the generality of any knowledge discovered.

The tool operates in three basic phases:

1. define sample classes of interest; search and collect existing measurements representing these classes,
2. compute the activity scores of various working units with respect to the collected samples,
3. apply statistical, machine learning and visualization methods to obtain models distinguishing between the defined classes, with the pre-computed activity scores of working units acting as sample features.

XGENE.ORG implements this workflow, facilitating all three phases above. The architecture of the tool is depicted in Figure 7.1. XGENE.ORG integrates data from several publicly accessible databases.

Regarding the first phase above, our tool provides an interface to the GEO. XGENE.ORG enables a keyword-based search and filtering of individual gene expression measurements as illustrated in Fig. 7.2. The interaction with the GEO is supervised by the user. The measurements are normalized and saved in the internal format that simplifies subsequent integration of the expression data with data capturing biological process structure (pathways) and relational information (the Gene ontology).

Secondly, XGENE.ORG accesses the databases that provide prior knowledge required to define and interpret the predefined set of working unit types (they are discussed in detail thereunder). The individual microarray platforms are annotated by the Bioconductor packages (Gentleman et al., 2004). Bioconductor packages also provide annotations by the GO terms. The prior knowledge on pathways and fluxes is taken directly from KEGG database. The prior knowledge management is fully automated and carried out without user interventions. The tool downloads all the packages and datasets needed to analyze the measurements currently selected by the user and stores them in the internal representation.

The critical step is to fuse the collected measurements and prior knowledge into unified cross-platform data subsequently accessed by the statistical and machine learning tools. Within this fusion, working units are computed across samples taken from various platforms and organisms. The resulting unified representation consists of a single matrix in which rows correspond to samples, columns correspond to working units and the respective matrix cells express the activity of a given unit within a given sample as a real value. Each working unit subsequently serves as a statistical variable for tasks such as fold change analysis, or a *sample feature* for machine learning algorithms.

Three kinds of analysis results are supported:

- a classifier that estimates the sample class given an expression sample and its platform label
- a list of working units significantly differentially expressed in classes
- a scatterplot that shows class distribution in a (transformed 2D) space of working units.

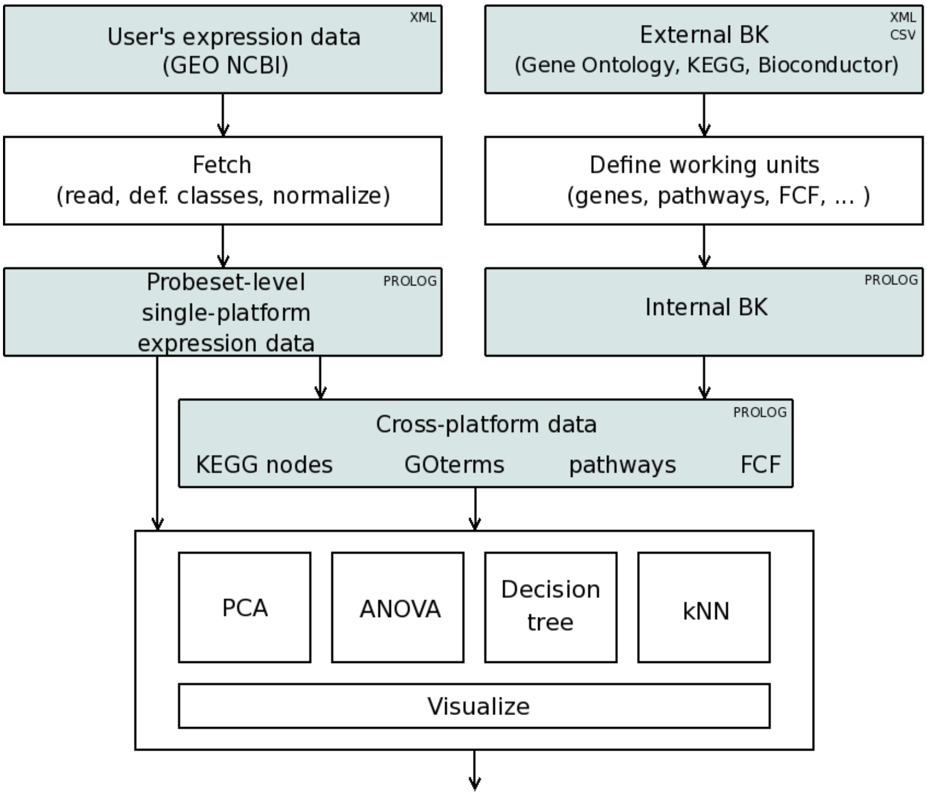


Figure 7.1: XGENE.ORG architecture

The results are provided to the user in the form of hypertext, including links pointing to detailed descriptions working units employed in the displayed result.

The interaction with the user who starts a new experiment consists of the following steps:

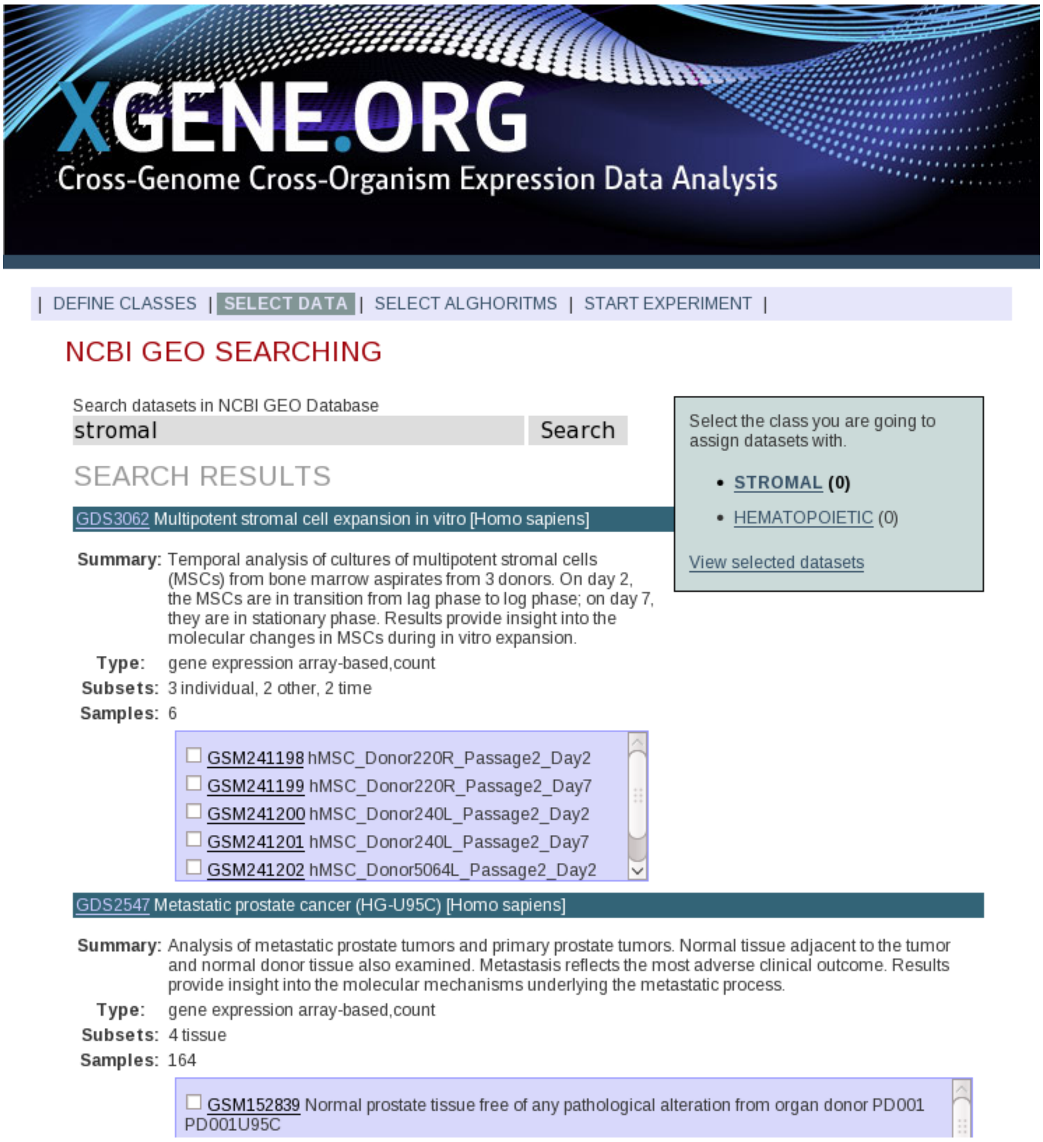
1. The user logs to his/her personal account. This account stores the user's previous experiments and their results.

2. The user creates a new experiment. The experiment can be entirely new (the interaction proceeds by the following step) or it can be derived from a previous experiment (the experiment then inherits the classes and datasets defined earlier and thus skips the two following steps).

3. The user creates and entitles two or more of sample classes. These classes contain no measurement samples at this stage.

4. The user fills each of the defined classes with a set of relevant GEO expression samples. The samples are preselected via keyword-based search and then finely filtered by the user on the basis of experimental annotations (see Figure 7.2),

5. The user selects (possibly repeatedly) proper working units, platform types and algorithms and starts the experiment.


XGENE.ORG
Cross-Genome Cross-Organism Expression Data Analysis
| DEFINE CLASSES | SELECT DATA | SELECT ALGHORITMS | START EXPERIMENT |
Search

NCBI GEO SEARCHING

Search datasets in NCBI GEO Database

stromal

SEARCH RESULTS

Select the class you are going to assign datasets with.

- **STROMAL (0)**
- HEMATOPOIETIC (0)

View selected datasets

GDS3062 Multipotent stromal cell expansion in vitro [Homo sapiens]

Summary: Temporal analysis of cultures of multipotent stromal cells (MSCs) from bone marrow aspirates from 3 donors. On day 2, the MSCs are in transition from lag phase to log phase; on day 7, they are in stationary phase. Results provide insight into the molecular changes in MSCs during in vitro expansion.

Type: gene expression array-based,count

Subsets: 3 individual, 2 other, 2 time

Samples: 6

□ GSM241198 hMSC_Donor220R_Passage2_Day2
□ GSM241199 hMSC_Donor220R_Passage2_Day7
□ GSM241200 hMSC_Donor240L_Passage2_Day2
□ GSM241201 hMSC_Donor240L_Passage2_Day7
□ GSM241202 hMSC_Donor5064L_Passage2_Day2

GDS2547 Metastatic prostate cancer (HG-U95C) [Homo sapiens]

Summary: Analysis of metastatic prostate tumors and primary prostate tumors. Normal tissue adjacent to the tumor and normal donor tissue also examined. Metastasis reflects the most adverse clinical outcome. Results provide insight into the molecular mechanisms underlying the metastatic process.

Type: gene expression array-based,count

Subsets: 4 tissue

Samples: 164

□ GSM152839 Normal prostate tissue free of any pathological alteration from organ donor PD001 PD001U95C

Figure 7.2: XGENE.ORG: collecting relevant samples from NCBI GEO. Clicking on a sample identifier ('GSMxxxxx') opens a detailed description of that sample.

6. The system collects the necessary prior knowledge, computes the working units defined above and applies the selected algorithms.

7. The computation begins and the user can log out. (S)he are informed by email as soon as the results are ready to be shown.

8. The user views the results. A result-filter helps user's orientation if a large number of result types has been requested in step 5.

7.1.3 Methods

This section describes the methodological elements of our approach. It gives an overview of working units and shows the way in which their activity is estimated and evaluated. It specifies the statistical methods serving to identify differentially expressed working units. It also gives a summary of currently implemented machine learning methods.

Working units—types and activity

XGENE.ORG consider two principal knowledge sources in order to define working units—the GO databaseand the KEGG database. The Bioconductor annotation packages serve to translate among the identifiers used by the microarray manufacturers (only Affymetrix is supported), and the two mentioned prior knowledge databases. The widely spread EntrezIds (gene identifiers) introduced by NCBI play the role of intermediate translation identifiers. The hierarchy of working units as implemented in XGENE.ORG is shown in Figure 7.3. The ultimate working units correspond to the measurement units, i.e., the probesets. Their activity in the individual samples is directly reported in the GEO input files. A single GEO file corresponds to a single microarray sample, a whole sample is represented by a probeset activity vector. The set of measured probesets is platform dependent, i.e., the vectors taken form different platforms cannot be directly matched. The more general units are gradually inferred from their subordinate units. For example, the list of probesets that are annotated by the same gene identifier makes up the *gene* working unit. The list of genes linked to a pathway node makes up the *pathway node* working unit. The activity of a pathway is computed by aggregating the activity of all probesets corresponding to genes which in turn correspond to nodes contained in the given pathway. Obviously, this mapping is platform dependent; pathways have different probeset interpretations in different platforms. At the same time, this mapping is organism dependent and thus we have to deal with organism orthologs of pathways, like in the Section 4.

The process of computation of KEGG node activity in a sample set that originates from two different platforms is exactly the same as in Figure 4.2. The only difference is that XGENE.ORG uses pathways, particularly pathway nodes, for mapping between different platforms instead of the orthology tables used in the experiments in Chapter 4.

Analysis algorithms

After the collection of all data needed for a defined experiment, *normalization* is conducted separately for each involved platform to consolidate same-platform samples. Quantile normalization (Bolstad et al., 2003) ensures that the distribution of expression values across such samples is identical. As a second step, scaling provides means to consolidate the measurements across multi-platform samples. We subtract the sample mean from all sample components, and divide them by the standard deviation within the sample. As a result, all samples independently of the platform

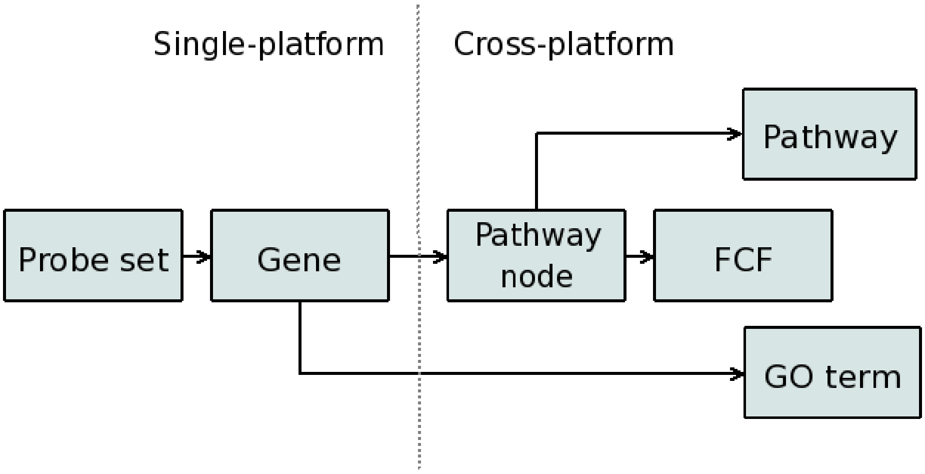


Figure 7.3: The hierarchy of working units. An arrow from X to Y denotes that unit Y refers to a set of X units. This relation is transitive and thus all units can ultimately be represented as families of probesets.

exhibit zero mean and unit variance. We conduct these steps using the Bioconductor (Gentleman et al., 2004) software.

After normalization, the most basic type of analysis that may be generated on user's request is *fold change* analysis whose goal is to rank the ability of the individual working units to distinguish among the user-defined classes. For this sake, we apply the one-way ANOVA or also implement Kruskal-Wallis test, which is a non-parametric equivalent of the one-way ANOVA. For the set-level analysis, we implement Global test (Goeman et al., 2004)

Having a single-tabular representation in which activity of a set of working units is computed across samples, a wide-scale of machine learning algorithms can be applied. The most interesting appear to be such algorithms that allow for direct human interpretation of the resulting models and still keep a good predictive power. Specifically, we included the J48 decision tree learner provided by the machine learning environment WEKA (Hall et al., 2009). The k-nearest neighbor (kNN) algorithm from the same environment has also been included.

Finally, principal component analysis (PCA) is used for the purpose of dimensionality reduction in a space of working units with subsequent visualization of samples (Holter et al., 2000). PCA is known to retain those characteristics of the dataset that contribute most to its variance. In XGENE.ORG it helps to exhibit class distribution in 2D and visually assess the potential of a set of working units to distinguish among classes.

7.1.4 Case studies

Here we demonstrate our methodology in two biological case studies. We address general tasks of tissue type classification. The first experiment focuses on distinct features of blood-forming (*hematopoietic*) and supportive (*stromal*) cellular compartments in the bone marrow. The second assesses differences in brain, liver and muscle tissues. Both experiments are of biological significance as they tackle novel challenges in understanding of cellular behavior: the former in the complex functional unit termed hematopoietic stem cell niche, where inter-dependent hematopoietic and stromal cell functions synergize in the blood-forming function of the bone marrow; the latter in comparison of cell fate determined by the tissue origin from the separate layers of the embryo: ectoderm (brain), endoderm (liver) and mesoderm (muscle). While of general character, the chosen tasks are not just random biological exercises as these studies may illuminate cellular functions determined by gene expression signatures in complex cell system seeded by cell-type-heterogeneous undifferentiated populations (hematopoietic and stromal stem cells in the cell niche), and in the cell-type-homogeneous differentiated tissues (brain, liver and muscle), respectively.

The significance tests at gene level identified elevated expression of genes canonical for the specific tissue studied, such as myelin basic protein in brain, isocitrate dehydrogenase in liver, tropomyosin in muscle and differential expression of integrin beta 5 inhematopoietic and stromal cell populations of the bone marrow.

The experiments with machine learning algorithms proved that working units applicable across platforms clearly distinguish among classes in both studies. The resulting models are compact, easy to interpret and accurate. Fig. 7.4 exemplifies the application of the decision tree learner J48 on the level of FCFs in the brain/liver/muscle study. The model tested by 10-fold cross-validation reaches the classification accuracy nearly 98%, it misclassifies 3 out of 131 samples. The tree has only 2 internal nodes (2 activity tests that put into use two FCFs) and 3 leaves (one leaf per class).

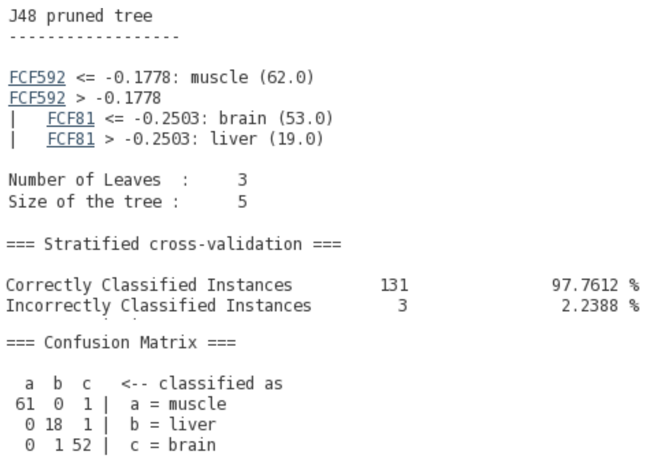
A similar conclusion follows from PCA visualizations (Fig. 7.5). The activity of working units tends to share the same pattern within classes as well as within the same platforms or the same laboratories. However, the class pattern is strong enough to clearly distinguish among classes independently of platform.

The complete overview of results is available via the XGENE.ORG webpage.

7.1.5 Discussion

XGENE.ORG is a web tool for the analysis of gene expression data collected from heterogeneous (multi-platform) microarray platforms under the presence of genomic prior knowledge. The integration of multi-platform data is conducted automatically by using the available genomic prior knowledge to define candidate working units general enough to be quantified in any sample regardless of the platform on which it was measured. The heterogeneous data are transformed into a single-tabular representation which summarizes the activity of the working units for all the collected samples. Such a unified representation lends itself to various types of analysis provided by XGENE.ORG based on statistical or machine learning methods.

The contribution of this tool is at least twofold. First, microarray experiments are costly, often resulting in numbers of samples insufficient for reliable modeling. The possibility of systematically integrating the experimenter's data with numerous public expression samples coming from heterogeneous platforms, would obviously help



```
J48 pruned tree
------------------

FCF592 <= -0.1778: muscle (62.0)
FCF592 > -0.1778
|   FCF81 <= -0.2503: brain (53.0)
|   FCF81 > -0.2503: liver (19.0)

Number of Leaves  :     3
Size of the tree :      5

=== Stratified cross-validation ===

Correctly Classified Instances         131              97.7612 %
Incorrectly Classified Instances         3               2.2388 %

=== Confusion Matrix ===

  a  b  c   <-- classified as
 61  0  1 |  a = muscle
  0 18  1 |  b = liver
  0  1 52 |  c = brain
```

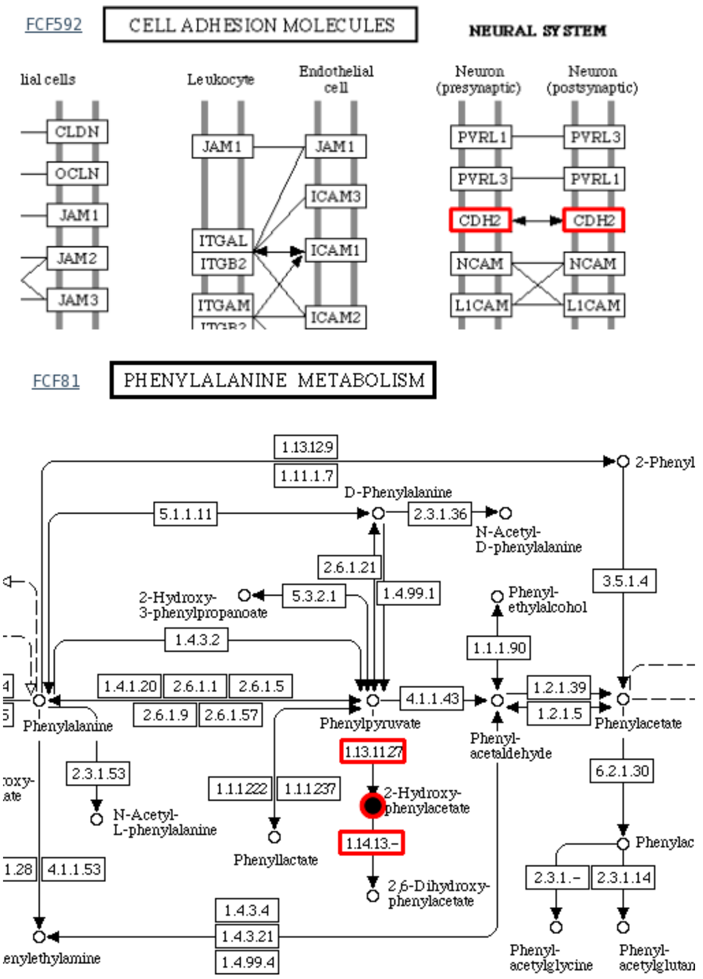


Figure 7.4: The flux-based cross-platform decision tree for the brain/liver/muscle study. The tree is very compact, the class is determined by two activity thresholds on two fluxes, the fluxes are visualized using KEGG pathway maps (in bold).

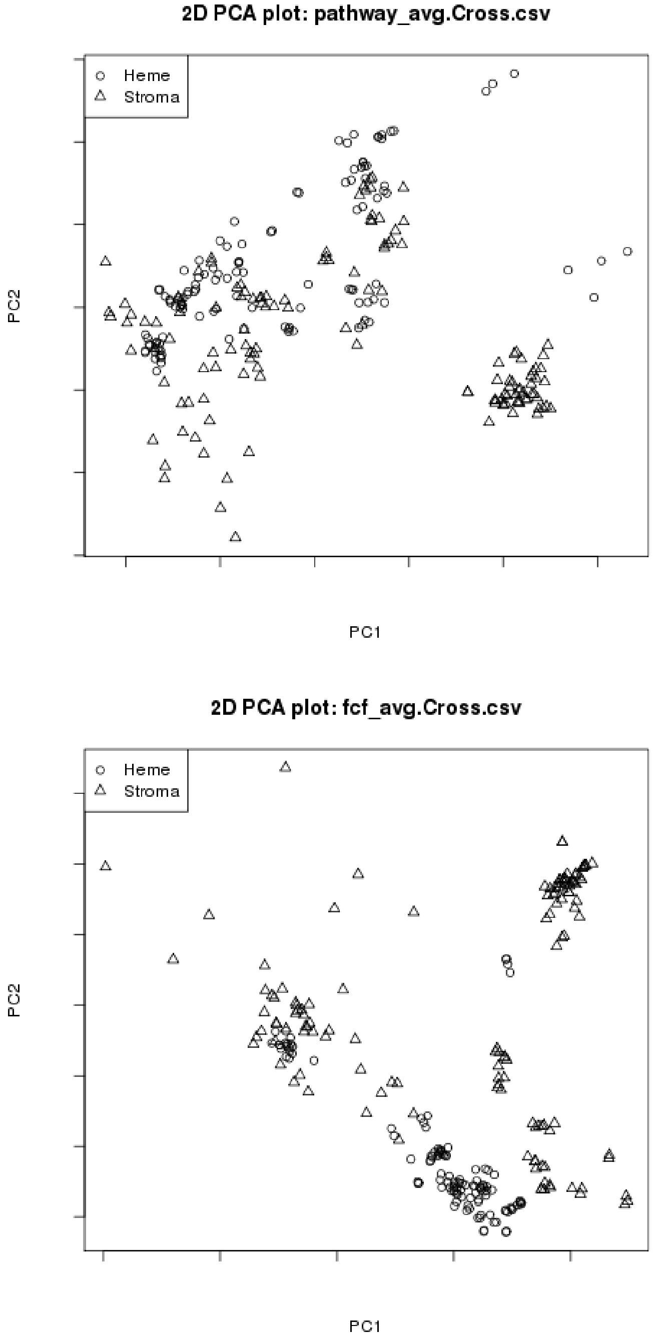


Figure 7.5: PCA in the hematopoietic/stromal study. The first subfigure shows cross-platform PCA in the space of pathways, the second subfigure uses FCFs instead. FCFs seem to better separate the classes (which is also confirmed by a higher classification accuracy if FCFs are used).

the experimenter. Second, such integrated analysis provides the principal means to discover biological markers shared by different-genome species.

7.2 miGENE—a new extension to XGENE.ORG

Here we describe a new bioinformatic tool which extends functionality of XGENE.ORG. It is a response to challenges related to new data sources of genetic and epigenetic data on gene level.

7.2.1 Introduction

In the last 15 years, microarrays have become well established technology. Also other novel high-throughput technologies emerged lately (e.g., RNA-seq, microRNA arrays, and methylation arrays). The same progress is present in quality and coverage of knowledge-related biological databases. Despite this progress, still an integrative analysis of the data remains a challenge. Here we describe a tool which addresses this issue.

One of the most critical obstacles in the analysis process dwells in use of the prior knowledge which is typically not related to the measured transcriptional activity, but concerns more abstract biological phenomena. This complication consists in a moderate-only correlation between mRNA and protein levels which is caused by post-transcriptional and post-translational modifications (Vogel and Marcotte, 2012) which makes the use of the protein-level knowledge difficult (Staiger et al., 2013). These modifications not only play an important role in cell development and many diseases (e.g., Nugent, 2014; Peng et al., 2009) but also negatively affects the use of prior knowledge which is *not* directly based on transcriptional regulation interactions (e.g., protein-protein interaction networks) (Staiger et al., 2013). This *lack of correlation* can be mitigated if we take into account other data sources; particularly, microRNA expression, since the small non-coding RNAs (microRNAs) play a role in translational and post-transcriptional regulation of gene expression and often result in gene silencing, and epigenetic data measuring level of DNA methylation and explaining unexpected transcriptional irregularities.

Here we presents the web tool miXGENE freely available at `http://mixgene.felk.cvut.cz/`. It is designed mainly for joint enrichment analysis of mRNA, microRNA and DNA methylation data. miXGENE also contains an interface to the database repository GEO and some other prior knowledge related databases (the GO, KEGG, MSigDB, miRWalk (Dweep et al., 2011), miRBase (Kozomara and Griffiths-Jones, 2011)). miXGENE allows the user to create their own analytic pipeline using an interactive workflow editor and offers a spectrum of methods designed for data visualisation and analysis of mRNA, microRNA and DNA methylation profiles.

7.2.2 System description

miXGENE is representative of the mashup technology that fuses data from several publicly available sources (NCBI raw profiles and platform annotations, R Bioconductor libraries (Gentleman et al., 2004) and MSigDB. The tool (Figure 7.6) can be split into three parts: (i) graphical user interface (task definition, presentation of results), (ii) workflow management (task decomposition and its global planning in terms of the individual plugins) and (iii) computational plugins (implementation of the individual analytical methods such as data normalization, feature extraction, learning of classifiers, etc.). Web interface and storage management are implemented in the web application framework Django, workflow management is implemented in JavaScript and the computational plugins are mainly implemented in Python and R (R Core Team, 2013).

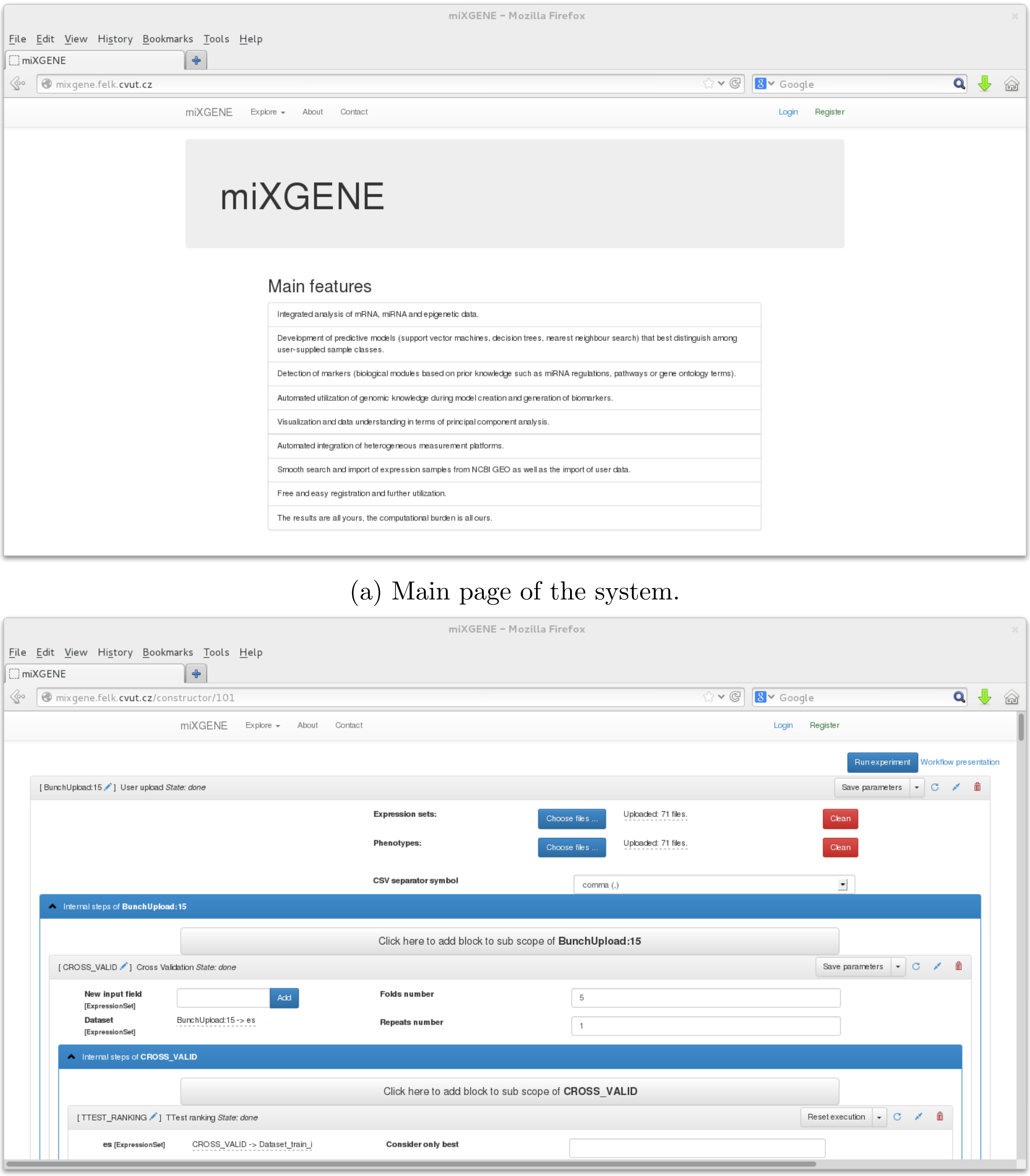
miXGENE as a workflow management system

miXGENE as a workflow management system are a growing area of research (Barker and Hemert, 2008). The main reasons for deployment of WMSs are: (i) an effort to make computational biology accessible to researchers who are not expert programmers, (ii) to enable tracking of experimental history and offer an easy-to-use tool for testing different settings, and (iii) the possibility to exchange the scientific workflows (Barker and Hemert, 2008). All these reasons are motivated by a goal to improve reproducibility, transparency and, therefore, mitigating experimental mistakes. There are many general frameworks or tools (both stand-alone and web-based) designed to represent bioinformatic or data-analytic workflows; e.g., BioBike (Elhai et al., 2009), Taverna (Wolstencroft et al., 2013), Galaxy (Goecks et al., 2010) and Anduril (Ovaska et al., 2010). miXGENE can be seen as a specialized bioinformatics workflow management system. Despite the fact the mentioned WMSs (mainly the Galaxy) already implement some tools (several statistical test) and interfaces (GEO) we require, the WMSs are too general for our purposes. Therefore, in order to facilitate maintenance (e.g., keeping our system up-to-date and as specific for the joint analysis as possible), we implemented our own WMS.

With miXGENE, all experiments are built from components called blocks using interactive workspace. Each block represents one meaningful step in the experiment e.g., providing a source dataset, creating a machine learning model, visualisation. Nevertheless, each block usually contains—in contrast to the more general systems mentioned above—a few atomic activities such as downloading input data, preprocessing and diagnostic visualization in the *source dataset providing block*. The execution order is inferred from the data flow defined by binding the corresponding output and input ports of the consecutive blocks. miXGENE enables the block structured pattern (La Rosa et al., 2011) and it does not allow cycle dependency and conditional execution.

miXGENE building blocks and types

miXGENE defines two types of blocks: "basic" *blocks* and *meta-blocks* where the latter serve as containers of other blocks or meta-blocks. The meta-blocks generate their own scope of possible input variables and, therefore, improve simplicity and clarity of workflows. This structure allows powerful and clear representation of machine learning workflows. Currently miXGENE enables blocks with the following functionality: (i) **data input** (access to gene expression data and knowledge from local user files or to selected public repositories), (ii) **data preprocessing** (tools


miXGENE – Mozilla Firefox
File Edit View History Bookmarks Tools Help
miXGENE
mixgene.felk.cvut.cz
Google
miXGENE Explore About Contact Login Register
miXGENE

Main features

- Integrated analysis of mRNA, miRNA and epigenetic data.
- Development of predictive models (support vector machines, decision trees, nearest neighbour search) that best distinguish among user-supplied sample classes.
- Detection of markers (biological modules based on prior knowledge such as miRNA regulations, pathways or gene ontology terms).
- Automated utilization of genomic knowledge during model creation and generation of biomarkers.
- Visualization and data understanding in terms of principal component analysis.
- Automated integration of heterogeneous measurement platforms.
- Smooth search and import of expression samples from NCBI GEO as well as the import of user data.
- Free and easy registration and further utilization.
- The results are all yours, the computational burden is all ours.

(a) Main page of the system.

miXGENE – Mozilla Firefox
File Edit View History Bookmarks Tools Help
miXGENE
mixgene.felk.cvut.cz/constructor/101
Google
miXGENE Explore About Contact Login Register
Run experiment
Workflow presentation
Save parameters
Choose files ...
Clean
Add
Reset execution

[BunchUpload:15] User upload *State: done*

Expression sets: Uploaded: 71 files.

Phenotypes: Uploaded: 71 files.

CSV separator symbol: comma (,)

Internal steps of **BunchUpload:15**

Click here to add block to sub scope of **BunchUpload:15**

[CROSS_VALID] Cross Validation *State: done*

New input field [ExpressionSet]

Dataset [ExpressionSet] BunchUpload:15 -> es

Folds number: 5

Repeats number: 1

Internal steps of **CROSS_VALID**

Click here to add block to sub scope of **CROSS_VALID**

[TTEST_RANKING] TTest ranking *State: done*

es [ExpressionSet] CROSS_VALID -> Dataset_train_i

Consider only best

(b) Interactive workflow editor.

Figure 7.6: miXGENE tool.

for working with missing data and normalisation), (iii) **data manipulation** (simple data concatenation in case of compatible datasets; integrating different datasets, e.g., from different platforms measuring mRNA expression and joint analysis for data from mRNA, microRNA and methylation platforms; see Section 7.2.3 for details), (iv) **analysis** (various machine learning and statistical methods; see Section 7.2.3 for details), (v) **visualization** (results in human readable form, e.g., graphs, tables, textual descriptions of models, (vi) **performance evaluation (meta-block)** (evaluation schemes like k-fold cross validation or leave-one-out cross-validation), (vii) **multiple datasets evaluation (meta-block)** (for performing the same experiment on two or more datasets).

miXGENE operates with predefined complex data types rather than with a combination of atomic types like integer, string, and array. Such an approach allows a required combination of data and meta-data for the desired level of workflow abstraction. E.g., the *Expression set* type contains gene expression data defined by a matrix dataset, phenotype description and platform annotation. Meta-data contains useful information about object content like data provider, used data type, properties of source tissue, etc. Data content is an object stored in fixed structure. Since data content may consume a great amount of memory, the complex data types allow serialization into the storage system.
List of implemented complex data types: (i) **expression set** (represents gene-expression data from a micro-array experiment including all necessary information), (ii) **gene set** (structure for representation of sets of genes, e.g., GO terms), (iii) **machine learning model** (learned model/classifier for the given data), (iv) **result table** (generic table in which each row represents features analysed during an experiment and each column represents different properties, metadata section contains a description of the column properties and working units), (v) **array container** (array of objects with the same structure, the cell structure description is stored in the metadata section).

Workflow construction

The main point of interaction between a user and the system is an experiment workspace with a block toolbox where the user defines an experimental workflow and executes it. The user constructs the new experiment from the empty workspace by adding appropriate blocks from the toolbox. To define data-flow, the user assigns input ports to outputs of the appropriate blocks. Then (if needed) the user sets mandatory or optional block parameters. When all the blocks in the experiment are configured correctly, the user can either execute each block by hand or run an automatic execution of the all blocks at the same time. The user will be notified about experiment's successful completion or will be pointed to occurred errors. The interactive nature of the experiment workspace allows the user to add more blocks anytime and continue the experiment with all the acquired results. Depiction of a machine-learning experiment based on a comparison of two alternative methods for analysis of mRNA and microRNA data is available via the miXGENE webpage. The

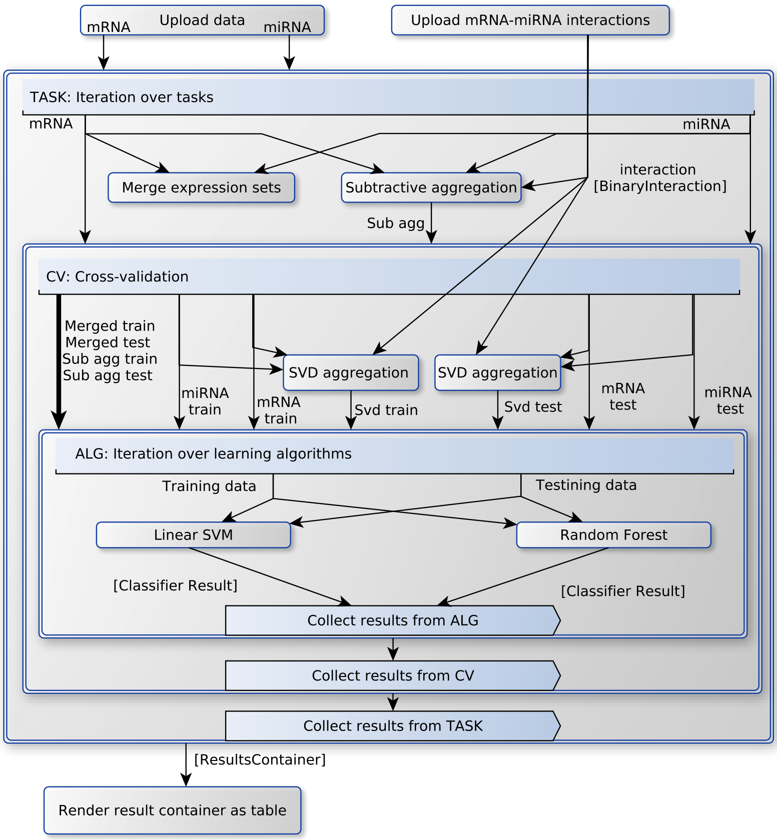


Figure 7.7: Workflow schema from the second case study. The first step is data and knowledge upload. In the nested meta-blocks, different types of data aggregation techniques are used. This workflow represents a cross-validation performance estimate of different machine learning algorithms over different data phenotypes (tasks). (From Gologuzov, 2014)

shown workflow produces an estimate of accuracy of both methods and also final models based on the mentioned methods.

7.2.3 Methods

This section describes the methodological elements of our approach. The implemented WMS is primarily designed to support analysis using attribute-value machine-learning methods. These methods take input in the form of matrix where samples are in columns and features (e.g., probesets or genes) are in rows; each column contains a gene expression profile from one sample. In the case of supervised learning methods there is another vector with an assignment of each sample to a class of samples (e.g. healthy or cancerous tissue). The unsupervised learning methods do not include such a vector; instead, it makes its own classification using data properties. As input data, miXGENE currently supports a few human and mouse mRNA and microRNA platforms provided by Affymetrix and Illumina GoldenGate methylation assays. A complete list of the supported platforms is available on the miXGENE web.

Aggregating methods for knowledge enrichment

The aggregating methods (alternatively set-statistics methods) can incorporate prior knowledge in the form of gene-sets using a direct transformation which also produces matrix data representation. For example, there is the pathway p which is represented as a set of genes $g_1, g_2, \ldots, g_n$ and matrix with gene expression profiles where each row contains GEs for a gene g_j for the all samples. The aggregating methods transform the gene-expression matrix induced by the genes in the pathway p into a row vector which represents aggregated expressions for all the samples; such a vector is typically denoted by the name of the geneset p. The current miXGENE version supports the following methods: simple statistics as mean, median, PCA based transformation, and SetSig (Mramor et al., 2010). Thanks the flexible representation of workflows, the miXGENE does not impose any restriction on the gene-sets' definition; therefore, it is possible to use these aggregation functions anytime there are appropriate gene-sets which can define the transformation from the former to the new representation.

Data integration approaches

Integrated analysis of gene expression data from different platforms and organisms. The integration is based on an assumption that it is generally possible to transform different data on the same common scale. For the integration of different MA platforms it can be mapping to the same genes and for different species it can be in evolutionary conserved elements like orthologous proteins. Generally, any common functionality describing gene sets like pathways or the GO terms can be used (Holec et al., 2009a).

Joint analysis of mRNA microRNA and methylation profiles. Presently, miXGENE supports two joint-analysis approaches. The first one is the "naive" approach proposed in Lanza et al. (2007) which is implicitly accessible due to the flexibility of the workflow designer tool and power of the machine learning methods. It joins all of the types of datasets by columns; from the three datasets with mRNA, microRNA and methylation profiles which are represented by three matrices with features $F_{microRNA}$, $F_{microRNA}$ and F_{methyl} the new "joint" dataset contains the set of features $F_{join} = F_{microRNA} \cup F_{microRNA} \cup F_{methyl}$. The second approach is based on a correction of mRNA expressions using microRNA expression profiles and known microRNA targets which describe the regulatory effect of microRNAs on mRNAs. miXGENE implements two versions of this approach; the substractive and the SVD-based method, which are suitable only for mRNA and microRNA data (Kléma et al., 2014).

Other methods

miXGENE also implements other well established and state-of-the-art methods for analysis on single data with and without prior knowledge and on joint mRNA and microRNA datasets as referential standards. Different analytical approaches typically offer definite different solution due to the presence of alternative solutions (e.g., marker genes can point not on to disease causing genes but erroneously to genes related to a consequence of the disease) or unstable nature of the methods (Michiels et al., 2005). Moreover, the lack of gold standard data makes it impossible to compare alternative methods thoroughly; therefore, there is a need for the referential methods in order to problem being scrutinized to the depths necessary. For the prior knowledge-enriched analysis of mRNA expression, miXGENE integrates the global test (Goeman et al., 2004). As an alternative to the joint analysis methods we have implemented the algorithm based on generation context specific microRNA regulation modules based on GO terms (Zhou et al., 2013).

7.2.4 Case studies

Here we demonstrate miXGENE functionality in two biological case studies. A concise overview of results is available via the miXGENE webpage. The first study follows experiments from Chapter 6.

The first cas study focused on an evaluation of the hypothesis "gene set aggregation methods improve predictive accuracy if we use gene sets based on the structure of transcription regulation networks and on the operon structure of bacterial genomes" and was conducted solely on mRNA gene expression data. Recent studies reject this hypothesis for gene sets based on the GO terms and KEGG pathways (Mramor et al., 2010; Staiger et al., 2013). We evaluated this hypothesis on 71 small microarray GE datasets measured in the bacteria. The results on the bacterial data indicate that methods based on aggregation of gene sets are able to improve predictive accuracy when provided with suitable gene sets. When inappropriate gene sets are used, e.g., when one uses GO terms or KEGG pathways, then the accuracy may actually drop significantly.

In the second case study, we evaluated our novel feature extraction and data integration method for the accurate and interpretable classification of biological samples based on their mRNA and microRNA expression profiles. The main idea was to use the knowledge of microRNA targets and better approximate the actual protein amount synthesized in the sample. The raw mRNA and microRNA expression features become enriched or replaced by new aggregated features that model the mRNA-microRNA regulation instead. The underlying hypothesis is that "the sample profile presumably gets closer to the phenotype being predicted". The proposed subtractive aggregation method (SubAgg) directly implements a simple mRNA-microRNA interaction model in which mRNA expression is modified using the expression of its targeting microRNAs. This method works with the simplifying assumption of the equal weight of the individual microRNAs suitable for small sample sizes where learning of their proper weights may lead to overfitting. Its SVD-based modification

(SVDAgg) enables different subtractive weights for different microRNAs learned by SVD. The two proposed knowledge-based subtractive methods were compared with their straightforward counterparts for obtaining the integrated mRNA and microRNA data through concatenating two respective datasets. We classified germ cell tumors patients under various experimental settings and compared the concatenation with SubAgg and SVDAgg. The results suggest that the knowledge-based approaches dominate the concatenation benchmark, and the features resulting from the mRNA-microRNA target relation can improve classification performance.

7.2.5 Conclusion

miXGENE represents a web tool for automated learning from heterogeneous genomic measurements that makes use of prior knowledge. The resulting models and markers match the actual measurements as well as the relationships among biological entities recorded in curated biological databases. The contribution of this tool is as follows. First, it provides the principal means for the user-friendly discovery of dedicated models in particular domains. Second, it is the platform for assembly, development, comparison and eventual dissemination of the methods for joint analysis of omics data. When compared with the traditional learning and statistical tools such as WEKA, RapidMiner, Orange or R/Bioconductor, it offers web interface with the possibility to easily fetch NCBI data and implements specific learning methods, currently SubAgg and SVDAgg proposed in Kléma et al. (2014). When compared with the bioinformatics WMSs such as Galaxy, it is focused on the specific task of learning from heterogeneous expression data. In particular, it facilitates the access both to the expression data and prior knowledge on their interaction, it provides specific learning methods and suggests sample workflows relevant to the given task.

Future work lies in further development and implementation of dedicated integration tools. We plan to continue with the development of our own methods as well as to employ the existing state-of-the-art algorithms. At the moment, there are no integration methods available for methylation and other epigenetic data available in miXGENE. We intend to improve miXGENE tool itself too, namely its graphical user interface and visualisation tools that serve for the presentation of results.

Chapter 8

Conclusions

In this thesis we examined several aspects of the gene expression data analysis in terms of priori defined gene sets and their impact on predictive performance. This approach provides more compact and interpretable results than those produced by traditional methods that rely on individual genes. The potential performance improvement rests in a correct feature-space reduction caused by the transformation of gene-level features into set-level features. Moreover, the analysis of the transformed data can be accomplished with traditional and well established supervised machine learning algorithms and offers a natural way to compare the performance of different methods.

The Conclusion is structured in three sections. Section 8.1 summarizes all the obtained results. Section 8.2 further discusses some of the salient experimental observations in light of some recent relevant papers. Section 8.3 indicates possible ways of future development.

8.1 Results

The core of this work is based on the thorough evaluation of different ratios between the number of samples and knowledge-enriched features in a large amount of classification experiments. Particularly, we evaluated two hypotheses (Hypothesis 1 and 2), where the first being related to the possibility of integrating observations from heterogeneous sources (originating even from different species), and the second addressing the question of whether gene expression data transformed from the space of genes into the space of gene sets lead to better results than the original non-transformed data. The entire evaluation is based on a comparison of predictive accuracies.

Moreover, we developed two publicly available tools, XGENE.ORG and miXGENE, where the former is dedicated to cross-platform and cross-species analysis and the latter provides the complete platform for the integrated analysis of mRNA, miRNA, and DNA methylation data with prior knowledge.

8.1.1 Hypothesis 1

To examine the first hypothesis, we demonstrate the integration of multi-platform gene expression data for predictive classification. When single-platform samples are rare, integration of related (cross-platform and cross-species) data boosts classification performance which supports Hypothesis 1 for the limited number of available samples only. In addition, we explored three ways of defining gene sets, including that based on the notion of the fully coupled flux representing a trade-off between very specific genes and general metabolic pathways. In 20 cross-platform classification tasks, we showed that the gene-set-based representation is useful for combining heterogeneous gene expression data. This may be for the sake of assembling a larger sample set or to obtain general biological insights not limited to a particular organism. The gene set features significantly outperform the gene-oriented ones in small sample sets (the training sets containing 10% and 20% of available samples), the fluxes keep this property even for the largest tested sample sets (the training sets containing 50% of available samples). The pathways and GO terms also give higher predictive accuracies than the gene-based features, but the significance of this difference cannot be proved on the selected significance level.

8.1.2 Hypothesis 2

We divided the examination of the second hypothesis into the two specific cases, the first is the general evaluation on Homo sapiens data using state-of-the-art gene sets (pathways and GO terms), and the second is performed entirely on Escherichia coli bacteria and focuses primarily on new, originally designed, gene set collections based on bacterial transcriptional regulatory networks.

Results on the state-of-the-art gene sets

We have established the following main conclusions by executing various experiments on 30 gene expression data classification problems.

1. State-of-the-art gene set ranking methods (GSEA, SAM-GS, Global test) perform reasonably as feature selectors in the machine learning context in that high ranking gene sets outperform (i.e., constitute better features for classification than) those that are low ranking.

2. Genuine curated gene sets from the Molecular Signature Database outperform randomized gene sets. Smaller gene sets and sets pertaining to chemical and genetic perturbations were particularly successful.

3. For gene set selection, the Global test (Goeman and Bühlmann, 2007) outperforms SAM-GS (Dinu et al., 2007), GSEA (Subramanian et al., 2005) as well as the generic information gain heuristic (Mitchell, 1997) and the SVM-based recursive feature elimination approach (Guyon et al., 2002).

4. For aggregating expressions of set member genes into a unique feature value, both SVD (Tomfohr et al., 2005) and SetSig (Mramor et al., 2010) outperform arithmetic averaging (e.g., Holec et al., 2009b).

5. Using the top ten gene sets to construct features results in better classifiers than using only the single best gene set.

6. The set-level approach using the top ten genuine gene sets as ranked by the Global test outperforms the baseline gene-level method in which the learning algorithm is given access to expressions of all measured genes. However, it is outperformed by the baseline approach if the latter is equipped with a prior feature selection step.

Conclusion 1 is rather obvious and was essentially meant as a preparatory check.

The first statement of Conclusion 2 is not obvious, since constructing randomized gene sets in fact corresponds to the machine learning technique of stochastic feature extraction (Ho, 1998) and as such may itself contribute to learning good classifiers. Nevertheless, relevant prior knowledge resting in the prior definition of biologically plausible gene sets contributes further to increasing the predictive accuracy. Conclusions 3 and 4 are probably the most significant for practitioners in set-level predictive modeling of gene expression as so far there has been no clear guidance for making the right choice.

Concerning Conclusion 3, the advantages of the Global test were argued in Goeman and Bühlmann (2007) but not supported in terms of the predictive power of the selected gene sets. As for Conclusion 4, the SetSig technique was introduced and tested in Mramor et al. (2010), appearing superior to both averaging and a PCA-based method which is conceptually equivalent to the SVD method (Tomfohr et al., 2005). However, owing to the limited experimental material in Mramor et al. (2010), the ranking was not confirmed by a statistical test. Here we confirmed the superiority of SetSig with respect to averaging; however, the difference between the performance of SetSig and SVD was not significant.

A further remark concerns the aggregation methods mentioned. All three of them are applicable to any kind of gene set collections, whether these are derived from pathways, the GO or other sources of prior knowledge. The downside of this generality is that substantial information available for specific kinds of gene sets is ignored.

Conclusion 5 is not entirely surprising. Relying only on a single gene set entails too large an information loss and results in classifiers less accurate than those using the ten best gene sets. Note that in the single gene set case, when aggregation is applied (SVD, AVG or SetSig), the sample becomes represented by only a single real-valued feature and learning essentially reduces to finding a threshold value for it. To verify that more than one gene set should be taken into account, we tested the 10-best-sets option and indeed it performed better.

A straightforward interpretation of Conclusion 6 is that the set-level framework is not an instrument for boosting predictive accuracy, and—therefore—we can reject Hypothesis 2 for the used state-of-the-art gene set collections, data, and methods. However, set-level classifiers have a value per se, just as set-level units are useful in the standard differential analysis of gene expression data. In this light, it is important that with a suitable choice of techniques, set-level classifiers do achieve an accuracy competitive with conventional gene-level classifiers.

Results on the originally designed gene sets

We evaluated the performance of gene sets based on the structure of transcription-regulation networks and on the operon structure of bacterial genomes using machine learning and gene set aggregation. All the gene sets are new in the context of predictive classification. For classification using the all-features scheme, we conclude that using prior knowledge in the form of gene sets can significantly improve predictive accuracy. This finding is not in contradiction with the conclusion of the paper by Mramor et al. (2010) as the gene sets used here differed from the gene sets used by Mramor et al. When we tested the performance of the set-level method with the same type of gene sets used by Mramor et al., we obtained worse results than when no gene sets were used which is in agreement with not only Mramor et al. but also other recent works (Staiger et al., 2012, 2013). For the feature selection case, we cannot assert significantly the dominance of the novel gene sets, or the combined versions where gene sets and individual genes were used together over the baseline due to a very high variance of accuracy typical for this type of experiment and a very low power of available statistical tests for the case when we have multiple statistically dependent points on every plot of sum of ranks. However the hybrid gene sets—despite their higher susceptibility to overfitting—perform better than individual genes for most of the points on the plots of sum of ranks.

The main conclusion is that methods based on an aggregation of gene sets are able to improve predictive accuracy when provided with suitable gene sets. When inappropriate gene sets are used, e.g., when one uses GO terms or KEGG pathways, then the accuracy may actually drop significantly. Therefore, we concede Hypothesis 2 when suitable gene set collection is chosen.

8.1.3 Software implementation of the set-level methods

XGENE

XGENE.ORG is a web tool for the analysis of gene expression data collected from heterogeneous (multi-platform) microarray platforms under the presence of genomic prior knowledge. The integration of multi-platform data is conducted automatically by using the available genomic prior knowledge to define candidate working units general enough to be quantified in any sample regardless of the platform on which it was measured. The heterogeneous data are transformed into a single-tabular representation which summarizes the activity of the working units for all the collected samples. Such a unified representation lends itself to various types of analysis provided by XGENE.ORG based on statistical or machine learning methods. The contribution of this tool is at least twofold. First, microarray experiments are costly, often resulting in numbers of samples insufficient for reliable modeling. The possibility of systematically integrating the experimenter's data with numerous public expression samples coming from heterogeneous platforms, would obviously help the experimenter. Second, such integrated analysis provides the principal means to discover biological markers shared by different-genome species.

miXGENE

The web tool miXGENE offers automated learning from heterogeneous genomic measurements that makes use of prior knowledge. The resulting models and markers match the actual measurements as well as the relationships among biological entities recorded in curated biological databases. The main contribution of this tool is the following. First, it provides the principal means for the user-friendly discovery of dedicated models in particular domains. Second, it is the platform for assembly, development, comparison and eventual dissemination of the methods for joint analysis of "omics" data. When compared with the traditional learning and statistical tools such as WEKA, RapidMiner, Orange or R/Bioconductor, it offers web interface with the possibility to easily fetch NCBI data and implements specific learning methods, currently SubAgg and SVDAgg proposed in Kléma et al. (2014). When compared with the bioinformatics WMSs such as Galaxy, it is focused on the specific task of learning from heterogeneous expression data. In particular, it facilitates the access both to the expression data and prior knowledge on their interaction, it provides specific learning methods and suggests sample workflows relevant to the given task.

8.2 Main observations

Several important observations follow from the previous section. Firstly, the success of the highly correlated gene set collections (gene sets based on the transcriptional regulatory network and the fully coupled fluxes in prokaryote and eukaryote organisms data, respectively). By the same token, experimental papers by Staiger et al. (2012, 2013) clearly demonstrate that there is no difference in performance of methods using the prior knowledge in the form of genuine and randomly generated protein-protein interaction networks, which is also theoretically supported by the recently discovered low correlation between mRNA and protein level (Vogel and Marcotte, 2012). Secondly, the consensus of our results with other studies addressing the set-level analysis based on the state-of-the-art gene set collections and machine learning performed on an appropriate number of different datasets (Abraham et al., 2010; Mramor et al., 2010; Staiger et al., 2012, 2013, or in Chapter 5). Our observation also supports conclusions that studies performed on a small number of datasets cannot provide scientifically sound results (Ioannidis, 2005; Jelizarow et al., 2010). Lastly, the importance of comparison of methods based on the predictive accuracy criterion[1] without any biological knowledge, which gives a truly unbiased performance estimate

[1] Or other appropriate metrics (e.g., the area under ROC curve, the Matthew's correlation coefficient) when needed.

criterion; however, general knowledge of the biological phenomenon assessed in the data is essential.

Despite the number of experiments performed during this study, only a fraction of the available methods (for feature selection, data aggregation, data integration, and classification) and their parameters were considered. Complete evaluation of all major approaches for different data domains exploiting various prior knowledge remains a challenge due to its combinatorial nature which leads to time-consuming and human-labor-demanding experiments. A solution to this challenge could lie in a specially designed experimental environment which could synergically focus the efforts of many researchers in one place. Important attempts for such an environment are the *Amsterdam Classification and Evaluation Suite* (ACES) by Staiger et al. (2013) and our tool miXGENE (Section 7.2) which combines experimental environment and a visual language based on workflow editing and is currently under development.

8.3 Future work

The integration of gene expression data and knowledge in the form of gene set collections can be extended in a qualitative or quantitative way. The first type of extension consists of using more precise data obtained, e.g., by RNA-seq technology. The latter type consists of additional data and knowledge source integration (e.g., microRNA, DNA methylation data, and the complete available transcriptional regulatory knowledge) which allows more precise modeling on the pathway level.

Another direction, in which this work can be extended, consists of a combination of the current set-level aggregation and feature selection methods in order to take into consideration the hierarchy of structured gene sets (e.g., the GO structure or organization of transcription units in operons) in a similar way as it is done in some statistics based set-level methods (e.g., Goeman and Mansmann, 2008). The ultimate integration method should take into account the generalization abilities of the set level (or functional) prior knowledge, and last but not least, include gene level features during the data modeling process.

The last remark regards the recent strong effort of the scientific community to find effective ways to merge different modalities of gene expression measurements as well as related background knowledge. A vast amount of papers comparing different aspects of the integration have appeared recently but a convincing organized methodology is still lacking. It remains a challenge to provide suitable tools (e.g., in the form of language or testing environment) which would connect the different relevant data sources.

We note here that the tool miXGENE (Section 7.2) reflects our response to the first and last remarks about possible future work.

Bibliography

Gad Abraham, Adam Kowalczyk, Sherene Loi, Izhak Haviv, and Justin Zobel. Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context. *BMC Bioinformatics*, 11:277, January 2010.

Marit Ackermann and Korbinian Strimmer. A general modular framework for gene set enrichment analysis. *BMC bioinformatics*, 10:47, January 2009.

Andrey Alexeyenko, Woojoo Lee, Maria Pernemalm, Justin Guegan, Philippe Dessen, Vladimir Lazar, Janne Lehtiö, and Yudi Pawitan. Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. *BMC Bioinformatics*, 13:226, September 2012.

David B Allison, Xiangqin Cui, Grier P Page, and Mahyar Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*, 7(1):55–65, January 2006.

Scott A Armstrong, Jane E Staunton, Lewis B Silverman, Rob Pieters, Monique L den Boer, Mark D Minden, Stephen E Sallan, Eric S Lander, Todd R Golub, and Stanley J Korsmeyer. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, 30(1):41–47, January 2002.

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, Midori A Harris, David P Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzana Lewis, John C Matese, Joel E Richardson, Martin Ringwald, Gerald M Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000.

Francisco Azuaje, Yvan Devaux, and Daniel R Wagner. Integrative pathway-centric modeling of ventricular dysfunction after myocardial infarction. *PLoS ONE*, 5(3): e9661, March 2010.

Adam Barker and Jano Van Hemert. Scientific workflow: a survey and research directions. In Roman Wyrzykowski, Jack Dongarra, Konrad Karczewski, and Jerzy Wasniewski, editors, *Parallel Processing and Applied Mathematics*, volume 4967 of *Lecture Notes in Computer Science*, pages 746–753. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Research*, 41 (Database issue):D991–995, January 2013.

William T Barry, Andrew B Nobel, and Fred A Wright. Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics (Oxford, England)*, 21(9):1943–1949, May 2005.

David G Beer, Sharon L R Kardia, Chiang-Ching Huang, Thomas J Giordano, Albert M Levin, David E Misek, Lin Lin, Guoan Chen, Tarek G Gharib, Dafydd G Thomas, Michelle L Lizyness, Rork Kuick, Satoru Hayasaka, Jeremy M G Taylor, Mark D Iannettoni, Mark B Orringer, and Samir Hanash. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine*, 8(8): 816–824, August 2002.

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(1):289–300, 1995.

Carolyn J M Best, John W Gillespie, Yajun Yi, Gadisetti V R Chandramouli, Mark A Perlmutter, Yvonne Gathright, Heidi S Erickson, Lauren Georgevich, Michael A Tangrea, Paul H Duray, Sergio González, Alfredo Velasco, W Marston Linehan, Robert J Matusik, Douglas K Price, William D Figg, Michael R Emmert-Buck, and Rodrigo F Chuaqui. Molecular alterations in primary prostate cancer after androgen ablation therapy. *Clinical Cancer Research*, 11(19 Pt 1):6823–6834, October 2005.

Vitoantonio Bevilacqua, Paolo Pannarale, Mirko Abbrescia, Claudia Cava, Angelo Paradiso, and Stefania Tommasi. Comparison of data-merging methods with SVM attribute selection and classification in breast cancer gene expression. *BMC Bioinformatics*, 13(Suppl 7):S9, January 2012.

Arindam Bhattacharjee, William G Richards, Jane Staunton, Cheng Li, Stefano Monti, Priya Vasa, Christine Ladd, Javad Beheshti, Raphael Bueno, Michael Gillette, Massimo Loda, Griffin Weber, Eugene J Mark, Eric S Lander, Wing Wong, Bruce E Johnson, Todd R Golub, David J Sugarbaker, and Matthew Meyerson. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences of the United States of America*, 98(24):13790–13795, November 2001.

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(7074):353–357, January 2006.

Benjamin M Bolstad. *Low-level analysis of high-density oligonucleotide array data: background, normalization and summarization.* PhD thesis, University of California, Berkeley, 2004.

Benjamin M Bolstad, Rafael A Irizarry, Magnus Astrand, and Terry P Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England)*, 19(2):185–193, January 2003.

Thomas Breslin, Morten Krogh, Carsten Peterson, and Carl Troein. Signal transduction pathway profiling of individual tumor samples. *BMC Bioinformatics*, 6(1): 163, January 2005.

Michael E Burczynski, Ron L Peterson, Natalie C Twine, Krystyna A Zuberek, Brendan J Brodeur, Lori Casciotti, Vasu Maganti, Padma S Reddy, Andrew Strahs, Fred Immermann, Walter Spinelli, Ulrich Schwertschlag, Anna M Slager, Monette M Cotreau, and Andrew J Dorner. Molecular classification of Crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. *Journal of Molecular Diagnostics*, 8(1):51–61, February 2006.

Roger D Canales, Yuling Luo, James C Willey, Bradley Austermiller, Catalin C Barbacioru, Cecilie Boysen, Kathryn Hunkapiller, Roderick V Jensen, Charles R Knight, Kathleen Y Lee, Yunqing Ma, Botoul Maqsodi, Adam Papallo, Elizabeth Herness Peters, Karen Poulter, Patricia L Ruppel, Raymond R Samaha, Leming Shi, Wen Yang, Lu Zhang, and Federico M Goodsaid. Evaluation of DNA microarray results with quantitative gene expression platforms. *Nature Biotechnology*, 24(9):1115–1122, September 2006.

Brendan J Carolan, Adriana Heguy, Ben-Gary Harvey, Philip L Leopold, Barbara Ferris, and Ronald G Crystal. Up-regulation of expression of the ubiquitin carboxyl-terminal hydrolase L1 gene in human airway epithelium of cigarette smokers. *Cancer Research*, 66(22):10729–10740, November 2006.

Xi Chen, Lily Wang, Jonathan D Smith, and Bing Zhang. Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes. *Bioinformatics (Oxford, England)*, 24(21):2474–2481, November 2008.

Han-Yu Chuang, Eunjung Lee, Yu-Tsueng Liu, Doheon Lee, and Trey Ideker. Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 3(140):140, January 2007.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995.

David Croft, Antonio Fabregat Mundo, Robin Haw, Marija Milacic, Joel Weiser, Guanming Wu, Michael Caudy, Phani Garapati, Marc Gillespie, Maulik R Kamdar, Bijay Jassal, Steven Jupe, Lisa Matthews, Bruce May, Stanislav Palatnik, Karen Rothfels, Veronica Shamovsky, Heeyeon Song, Mark Williams, Ewan Birney, Henning Hermjakob, Lincoln Stein, and Peter D'Eustachio. The Reactome pathway knowledgebase. *Nucleic Acids Research*, 42(Database issue):D472–477, January 2014.

Colleen Cutcliffe, Donna Kersey, Chiang-Ching Huang, Yong Zeng, David Walterhouse, and Elizabeth J Perlman. Clear cell sarcoma of the kidney: up-regulation of neural markers with activation of the sonic hedgehog and Akt pathways. *Clinical Cancer Research*, 11(22):7986–7994, November 2005.

Patricia L M Dahia, Ken N Ross, Matthew E Wright, César Y Hayashida, Sandro Santagata, Marta Barontini, Andrew L Kung, Gabriela Sanso, James F Powers, Arthur S Tischler, Richard Hodin, Shannon Heitritter, Francis Moore, Robert Dluhy, Julie Ann Sosa, I Tolgay Ocal, Diana E Benn, Deborah J Marsh, Bruce G Robinson, Katherine Schneider, Judy Garber, Seth M Arum, Márta Korbonits, Ashley Grossman, Pascal Pigny, Sérgio P A Toledo, Vania Nosé, Cheng Li, and Charles D Stiles. A HIF1alpha regulatory loop links hypoxia and mitochondrial signals in pheochromocytomas. *PLoS Genetics*, 1(1):72–80, July 2005.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, January 2006.

Irina Dinu, John D Potter, Thomas Mueller, Qi Liu, Adeniyi J Adewale, Gian S Jhangri, Gunilla Einecke, Konrad S Famulski, Philip Halloran, and Yutaka Yasui. Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics*, 8:242, January 2007.

Irina Dinu, Qi Liu, John D Potter, Adeniyi J Adewale, Gian S Jhangri, Gunilla Einecke, Konrad Famulsky, Philip Halloran, and Yutaka Yasui. A biological evaluation of six gene set analysis methods for identification of differentially expressed pathways in microarray data. *Cancer Informatics*, 1(780):357–368, June 2008.

Harsh Dweep, Carsten Sticht, Priyanka Pandey, and Norbert Gretz. miRWalk–database: prediction of possible miRNA binding sites by "walking" the genes of three genomes. *Journal of biomedical informatics*, 44(5):839–847, October 2011.

Elena Edelman, Alessandro Porrello, Justin Guinney, Bala Balakumaran, Andrea Bild, Phillip G Febbo, and Sayan Mukherjee. Analysis of sample set enrichment scores: assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles. *Bioinformatics (Oxford, England)*, 22(14):e108–116, July 2006.

Ron Edgar, Michael Domrachev, and Alex E Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, January 2002.

Sol Efroni, Carl F Schaefer, and Kenneth H Buetow. Identification of key processes underlying cancer phenotypes using biologic pathway analysis. *PLoS One*, 2(5): e425, January 2007.

Liat Ein-Dor, Itai Kela, Gad Getz, David Givol, and Eytan Domany. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics (Oxford, England)*, 21(2):171–178, January 2005.

Michael B Eisen, Paul T Spellman, Patrick O Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25):14863–14868, December 1998.

Jeff Elhai, Arnaud Taton, J P Massar, John K Myers, Mike Travers, Johnny Casey, Mark Slupesky, and Jeff Shrager. BioBIKE: a web-based, programmable, integrated biological knowledge base. *Nucleic acids research*, 37(Web Server issue):W28–32, July 2009.

Pierre Farmer, Herve Bonnefoi, Veronique Becette, Michele Tubiana-Hulin, Pierre Fumoleau, Denis Larsimont, Gaetan Macgrogan, Jonas Bergh, David Cameron, Darlene Goldstein, Stephan Duss, Anne-Laure Nicoulaz, Cathrin Brisken, Maryse Fiche, Mauro Delorenzi, and Richard Iggo. Identification of molecular apocrine breast tumours by microarray analysis. *Oncogene*, 24(29):4660–4671, July 2005.

William A Freije, F Edmundo Castro-Vargas, Zixing Fang, Steve Horvath, Timothy Cloughesy, Linda M Liau, Paul S Mischel, and Stanley F Nelson. Gene expression profiling of gliomas strongly predicts survival. *Cancer Research*, 64(18):6503–6510, September 2004.

Dragan Gamberger, Nada Lavrac, Filip Zelezný, and Jakub Tolar. Induction of comprehensible models for gene expression datasets by subgroup discovery methodology. *Journal of Biomedical Informatics*, 37(4):269–284, August 2004.

Isabella Gashaw, Ruth Grümmer, Ludger Klein-Hitpass, Oliver Dushaj, Martin Bergmann, Ralph Brehm, Rainer Grobholz, Sabine Kliesch, Tanja P Neuvians, Kurt W Schmid, Chrstine von Ostau, and Elke Winterhager. Gene signatures of testicular seminoma with emphasis on expression of ets variant gene 4. *Cellular and Molecular Life Sciences*, 62(19-20):2359–2368, October 2005.

Ludwig Geistlinger, Gergely Csaba, Robert Küffner, Nicola Mulder, and Ralf Zimmer. From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems. *Bioinformatics (Oxford, England)*, 27(13):i366–373, July 2011.

Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H Yang, and Jianhua

Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, January 2004.

Jeremy Goecks, Anton Nekrutenko, and James Taylor. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology*, 11(8):R86, January 2010.

Jelle J Goeman and Peter Bühlmann. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics (Oxford, England)*, 23(8):980–987, April 2007.

Jelle J Goeman and Ulrich Mansmann. Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics (Oxford, England)*, 24(4):537–544, March 2008.

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 (Oxford, England)*, 20(1):93–99, December 2004.

Valentin Gologuzov. mixgene: Development of a web tool for integrative analysis of genomic data. Master's thesis, Czech Technical University, Czech Republic, 2014.

Todd R Golub, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P Mesirov, Hilary Coller, Mignon L Loh, James R Downing, Michael A Caligiuri, Clara D Bloomfield, and Eric S Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, October 1999.

Gavin J Gordon. Transcriptional profiling of mesothelioma using microarrays. *Lung Cancer*, 49(Suppl 1):S99–103, July 2005.

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

Rosa María Gutiérrez-Ríos, David A Rosenblueth, José Antonio Loza, Araceli M Huerta, Jeremy D Glasner, Fred R Blattner, and Julio Collado-Vides. Regulatory network of Escherichia coli: consistency between literature knowledge and microarray profiles. *Genome Research*, 13(11):2435–2443, November 2003.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3): 389–422, March 2002.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software: an update. *SIGKDD Explorations Newsletter*, 11(1):10–18, November 2009.

Blaise Hanczar, Jianping Hua, and Edward R Dougherty. Decorrelation of the true and estimated classifier errors in high-dimensional settings. *EURASIP Journal on Bioinformatics and Systems Biology*, page 38473, January 2007.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2nd edition, 2001.

Yoshitaka Hippo, Hirokazu Taniguchi, Shuichi Tsutsumi, Naoko Machida, Ja-Mun Chong, Masashi Fukayama, Tatsuhiko Kodama, and Hiroyuki Aburatani. Global gene expression analysis of gastric cancer by oligonucleotide microarrays. *Cancer Research*, 62(1):233–240, 2002.

Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:832–844, August 1998.

Matěj Holec, Jiří Kléma, Filip Železný, Jiří Bělohradský, and Jakub Tolar. Cross-genome knowledge-based expression data fusion. In *International Conference on Bioinformatics, Computational Biology, Genomics, and Chemoinformatics*, 2009a.

Matěj Holec, Filip Železný, Jiří Kléma, and Jakub Tolar. Integrating multiple-platform expression data through gene set features. In Ion Mandoiu, Giri Narasimhan, and Yanqing Zhang, editors, *Bioinformatics Research and Applications*, volume 5542 of *Lecture Notes in Computer Science*, pages 5–17. Springer Berlin Heidelberg, 2009b.

Neal S Holter, Madhusmita Mitra, Amos Maritan, Marek Cieplak, Jayanth R Banavar, and Nina V Fedoroff. Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proceedings of the National Academy of Sciences of the United States of America*, 97(15):8409–8414, July 2000.

Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, 4 (1):44–57, January 2009.

Jianping Huang, Hong Fang, and Xiaohui Fan. Decision forest for classification of gene expression data. *Computers in Biology and Medicine*, 40(8):698–704, August 2010.

Seungwoo Hwang. Comparison and evaluation of pathway-level aggregation methods of gene expression data. *BMC genomics*, 13(Suppl 7):S26, January 2012.

John P A Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8), 2005.

Monika Jelizarow, Vincent Guillemot, Arthur Tenenhaus, Korbinian Strimmer, and Anne-Laure Boulesteix. Over-optimism in bioinformatics: an illustration. *Bioinformatics (Oxford, England)*, 26(16):1990–1998, August 2010.

Minoru Kanehisa, Susumu Goto, Shuichi Kawashima, Yasushi Okuno, and Masahiro Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Research*, 32(Database issue):D277–280, January 2004.

Purvesh Khatri, Marina Sirota, and Atul J Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, 8(2): e1002375, January 2012.

Jihoon Kim, Kiltesh Patel, Hyunchul Jung, Winston P Kuo, and Lucila Ohno-Machado. AnyExpress: integrated toolkit for analysis of cross-platform gene expression data using a fast interval matching algorithm. *BMC Bioinformatics*, 12: 75, January 2011.

Jiří Kléma, Arnaud Soulet, Bruno Cremilleux, Sylvain Blachon, and Olivier Gandrillon. Mining plausible patterns from genomic data. In *CBMS 2006 19th IEEE International Symposium on Computational Medical Systems*, pages 183–190, 2006.

Jiří Kléma, Jan Záhalka, Michael Anděl, and Zdeněk Krejčík. Knowledge-based subtractive integration of mRNA and miRNA expression profiles to differentiate myelodysplastic syndrome. In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms*, 2014.

Ana Kozomara and Sam Griffiths-Jones. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic acids research*, 39(Database issue):D152–157, January 2011.

Miloš Krejník and Jiří Kléma. Empirical evidence of the applicability of functional clustering through gene expression classification. *IEEE/ACM Transactions on Computational Biology and Bioinformacs*, 9(3):788–798, 2012.

Alexandre Kuhn, Ruth Luthi-Carter, and Mauro Delorenzi. Cross-species and cross-platform gene expression studies with the Bioconductor-compliant R package 'annotationTools'. *BMC Bioinformatics*, 9:26, January 2008.

Amar Kumar, Marne A Higgins, John N Calley, Scott M McAhren, Bartley W Halstead, Ernst R Dow, Timothy P Ryan, Adam West, Hong Gao, and George H Searfoss. Abstracting genes to Gene Ontology terms allows comparison across multiple species. In *18th International Conference on System Engineering*, pages 320–325. IEEE, 2005.

Moni A Kuriakose, Wan T Chen, Zara M He, Andrew G Sikora, Ping Zhang, Zhi Y Zhang, W L Qiu, D Frank Hsu, Cameron McMunn-Coffran, Stuart M Brown, E Murugaian Elango, Mark D Delacure, and Fangan A Chen. Selection and validation of differentially expressed genes in head and neck cancer. *Cellular and Molecular Life Sciences*, 61(11):1372–1383, June 2004.

Marcello La Rosa, Petia Wohed, Jan Mendling, Arthur H. M. ter Hofstede, Hajo A. Reijers, and Wil M. P. van der Aalst. Managing process model complexity via abstract syntax modifications. *IEEE Transactions on Industrial Informatics*, 7(4): 614–629, November 2011.

Ronilda Lacson, Erik Pitzer, Jihoon Kim, Pedro Galante, Christian Hinske, and Lucila Ohno-Machado. DSGeo: software tools for cross-platform analysis of gene expression data in GEO. *Journal of Biomedical Informatics*, 43(5):709–715, October 2010.

Päivi Laiho, Antti Kokko, Sakari Vanharanta, Reijo Salovaara, Heli Sammalkorpi, Heikki Järvinen, Jukka-Pekka Mecklin, Tuomo J Karttunen, Karoliina Tuppurainen, Veronica Davalos, Simo Schwartz, Diego Arango, Markus J Mäkinen, and Lauri A Aaltonen. Serrated carcinomas form a subclass of colorectal cancer with distinct molecular basis. *Oncogene*, 26(2):312–320, January 2007.

Giovanni Lanza, Manuela Ferracin, Roberta Gafà, Angelo Veronese, Riccardo Spizzo, Flavia Pichiorri, Chang-gong Liu, George A Calin, Carlo M Croce, and Massimo Negrini. mRNA/microRNA gene expression profile in microsatellite unstable colorectal cancer. *Molecular cancer*, 6:54, January 2007.

Eunjung Lee, Han-Yu Chuang, Jong-Won Kim, Trey Ideker, and Doheon Lee. Inferring pathway activity toward precise disease classification. *PLoS Comput. Biol.*, 4 (11):e1000217, December 2008.

David M Levine, David R Haynor, John C Castle, Sergey B Stepaniants, Matteo Pellegrini, Mao Mao, and Jason M Johnson. Pathway and gene-set activation measurement from mRNA expression data: the tissue distribution of human pathways. *Genome Biology*, 7(10):R93, January 2006.

Johan Leyritz, Stéphane Schicklin, Sylvain Blachon, Céline Keime, Céline Robardet, Jean-Francois Boulicaut, Jérémy Besson, Ruggero Pensa, and Olivier Gandrillon. SQUAT: a web tool to mine human, murine, and avian SAGE data. *BMC Bioinformatics*, 9:378, September 2008.

Helena Líbalová, Miroslav Dostál, Pavel Rossner, Jan Topinka, and Radim J Šrám. Gene expression profiling in blood of asthmatic children living in polluted region of the Czech Republic (Project AIRGEN). In *10th International Conference on Environmental Mutagenesis*, 2010.

Helena Líbalová, Kateřina Uhlířová, Jiří Kléma, Miroslav Machala, Radim J Šrám, Miroslav Cigánek, and Jan Topinka. Global gene expression changes in human embryonic lung fibroblasts induced by organic extracts from respirable air particles. *Part Fibre Toxicol*, 9:1, January 2012.

Robert J Lipshutz, Stephen P Fodor, Thomas R Gingeras, and David J Lockhart. High density synthetic oligonucleotide arrays. *Nature Genetics*, 21(Suppl 1):20–24, January 1999.

Huan Liu and Hiroshi Motoda. *Feature selection for knowledge discovery and data mining.* Kluwer, 1998.

Jiajun Liu, Jacqueline M Hughes-Oliver, and J Alan Menius. Domain-enhanced analysis of microarray data using GO annotations. *Bioinformatics (Oxford, England)*, 23(10):1225–1234, May 2007a.

Qi Liu, Irina Dinu, Adeniyi J Adewale, John D Potter, and Yutaka Yasui. Comparative evaluation of gene-set analysis methods. *BMC Bioinformatics*, 8(1):431, January 2007b.

Yong Lu, Peter Huggins, and Ziv Bar-Joseph. Cross species analysis of microarray expression data. *Bioinformatics (Oxford, England)*, 25(12):1476–1483, June 2009.

Werner K Maas. Studies on the mechanism of repression of argninine biosynthesis in Escherichia coli. II. Dominance of repressibility in diploids. *Journal of Molecular Biology*, 8:365–370, March 1964.

Ben D MacArthur, Avi Ma'ayan, and Ihor R Lemischka. Systems biology of stem cell fate and cellular reprogramming. *Nature Reviews Molecular Cell Biology*, 10(10): 672–681, October 2009.

Tobias Maier, Alexander Schmidt, Marc Güell, Sebastian Kühner, Anne-Claude Gavin, Ruedi Aebersold, and Luis Serrano. Quantification of mRNA and protein and integration with protein turnover in a bacterium. *Molecular Systems Biology*, 7:511, January 2011.

Theodora Manoli, Norbert Gretz, Hermann-Josef Gröne, Marc Kenzelmann, Roland Eils, and Benedikt Brors. Group testing for pathway analysis improves comparability of different microarray datasets. *Bioinformatics (Oxford, England)*, 22(20): 2500–2506, October 2006.

Joëlle Michaud, Ken M Simpson, Robert Escher, Karine Buchet-Poyau, Tim Beissbarth, Catherine Carmichael, Matthew E Ritchie, Frédéric Schütz, Ping Cannon, Marjorie Liu, Xiaofeng Shen, Yoshiaki Ito, Wendy H Raskind, Marshall S Horwitz, Motomi Osato, David R Turner, Terence P Speed, Maria Kavallaris, Gordon K Smyth, and Hamish S Scott. Integrative analysis of RUNX1 downstream pathways and target genes. *BMC Genomics*, 9(1):363, January 2008.

Stefan Michiels, Serge Koscielny, and Catherine Hill. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, 365(9458):488–492, 2005.

Tom Mitchell. *Machine learning.* McGraw-Hill Education (ISE Editions), 1st edition, May 1997.

Sushmita Mitra and Sampreeti Ghosh. Feature Selection and Clustering of Gene Expression Profiles Using Biological Knowledge. *IEEE Transactions on Systems, Man,* *and Cybernetics, Part C (Applications and Reviews)*, 42(6):1590–1599, November 2012.

Vamsi K Mootha, Cecilia M Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstrå le, Esa Laurila, Nicholas Houstis, Mark J Daly, Nick Patterson, Jill P Mesirov, Todd R Golub, Pablo Tamayo, Bruce Spiegelman, Eric S Lander, Joel N Hirschhorn, David Altshuler, and Leif C Groop. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, 34(3):267–273, July 2003.

Minca Mramor, Marko Toplak, Gregor Leban, Tomaz Curk, and Blaz Zupan. On utility of gene set signatures in gene expression-based cancer class prediction. *Journal of Machine Learning Research – Proceedings Track*, 8:55–64, 2010.

Richard A Notebaart, Bas Teusink, Roland J Siezen, and Balázs Papp. Co-regulation of metabolic genes is better explained by flux coupling than by network distance. *PLoS Comput. Biol.*, 4(1):e26, January 2008.

Mary Nugent. MicroRNA function and dysregulation in bone tumors: the evidence to date. *Cancer management and research*, 6:15–25, January 2014.

Kristian Ovaska, Marko Laakso, Saija Haapa-Paananen, Riku Louhimo, Ping Chen, Viljami Aittomäki, Erkka Valo, Javier Núñez Fontarnau, Ville Rantanen, Sirkku Karinen, Kari Nousiainen, Anna-Maria Lahesmaa-Korpinen, Minna Miettinen, Lilli Saarinen, Pekka Kohonen, Jianmin Wu, Jukka Westermarck, and Sampsa Hautaniemi. Large-scale data integration framework provides a comprehensive view on glioblastoma multiforme. *Genome medicine*, 2(9):65, January 2010.

Xinxia Peng, Yu Li, Kathie-Anne Walters, Elizabeth R Rosenzweig, Sharon L Lederer, Lauri D Aicher, Sean Proll, and Michael G Katze. Computational identification of hepatitis C virus associated microRNA-mRNA regulatory modules in human livers. *BMC genomics*, 10:373, January 2009.

Ernesto Perez-Rueda and Julio Collado-Vides. The repertoire of DNA-binding transcriptional regulators in Escherichia coli K-12. *Nucleic Acids Research*, 28(8):1838–1847, April 2000.

R Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2013.

Jörg Rahnenführer, Francisco S Domingues, Jochen Maydt, and Thomas Lengauer. Calculating the statistical significance of changes in pathway activity from gene expression data. *Statistical applications in genetics and molecular biology*, 3, June 2004.

Franck Rapaport, Andrei Zinovyev, Marie Dutreix, Emmanuel Barillot, and Jean-Philippe Vert. Classification of microarray data using gene networks. *BMC Bioinformatics*, 8(1):35, January 2007.

Jason Rudy and Faramarz Valafar. Empirical comparison of cross-platform normalization methods for gene expression data. *BMC Bioinformatics*, 12(1):467, December 2011.

Heladia Salgado, Martin Peralta-Gil, Socorro Gama-Castro, Alberto Santos-Zavaleta, Luis Muñiz Rascado, Jair S García-Sotelo, Verena Weiss, Hilda Solano-Lira, Irma Martínez-Flores, Alejandra Medina-Rivera, Gerardo Salgado-Osorio, Shirley Alquicira-Hernández, Kevin Alquicira-Hernández, Alejandra López-Fuentes, Liliana Porrón-Sotelo, Araceli M Huerta, César Bonavides-Martínez, Yalbi I Balderas-Martínez, Lucia Pannier, Maricela Olvera, Aurora Labastida, Verónica Jiménez-Jacinto, Leticia Vega-Alvarado, Victor Del Moral-Chávez, Alfredo Hernández-Alvarez, Enrique Morett, and Julio Collado-Vides. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Research*, 41(Database issue):D203–213, January 2013.

Clemens R Scherzer, Aron C Eklund, Lee J Morse, Zhixiang Liao, Joseph J Locascio, Daniel Fefer, Michael A Schwarzschild, Michael G Schlossmacher, Michael A Hauser, Jeffery M Vance, Lewis R Sudarsky, David G Standaert, John H Growdon, Roderick V Jensen, and Steven R Gullans. Molecular markers of early Parkinson's disease based on gene expression in blood. *Proceedings of the National Academy of Sciences of the United States of America*, 104(3):955–960, January 2007.

Andrey A Shabalin, Hå kon Tjelmeland, Cheng Fan, Charles M Perou, and Andrew B Nobel. Merging two gene-expression studies via cross-platform normalization. *Bioinformatics (Oxford, England)*, 24(9):1154–1160, May 2008.

Andrey S Shaw and Erin L Filbert. Scaffold proteins and immune-cell signalling. *Nature Reviews Immunology*, 9(1):47–56, January 2009.

Sarah Song and Michael A Black. *pcot2: Principal Coordinates and Hotelling's T-Square method*, 2006.

Sarah Song and Michael A Black. Microarray-based gene set analysis: a comparison of current methods. *BMC bioinformatics*, 9:502, January 2008.

Avrum Spira, Jennifer E Beane, Vishal Shah, Katrina Steiling, Gang Liu, Frank Schembri, Sean Gilman, Yves-Martine Dumas, Paul Calner, Paola Sebastiani, Sriram Sridhar, John Beamis, Carla Lamb, Timothy Anderson, Norman Gerry, Joseph Keane, Marc E Lenburg, and Jerome S Brody. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nature Medicine*, 13(3):361–366, March 2007.

Christine Staiger, Sidney Cadot, Raul Kooter, Marcus Dittrich, Tobias Müller, Gunnar W Klau, and Lodewyk F A Wessels. A critical evaluation of network and pathway-based classifiers for outcome prediction in breast cancer. *PLoS One*, 7(4): e34796, January 2012.

Christine Staiger, Sidney Cadot, Balázs Győrffy, Lodewyk F. A. Wessels, and Gunnar W. Klau. Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis. *Frontiers in Genetenetics*, 4(289), December 2013.

Maria A Stalteri and Andrew P Harrison. Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips. *BMC Bioinformatics*, 8:13, January 2007.

Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102 (43):15545–15550, October 2005.

Lixin Sun, Ai-Min Hui, Qin Su, Alexander Vortmeyer, Yuri Kotliarov, Sandra Pastorino, Antonino Passaniti, Jayant Menon, Jennifer Walling, Rolando Bailey, Marc Rosenblum, Tom Mikkelsen, and Howard A Fine. Neuronal and glioma-derived stem cell factor induces angiogenesis within the brain. *Cancer Cell*, 9(4):287–300, April 2006.

Yuting Sun and Jie Chen. mTOR signaling: PLD takes center stage. *Cell Cycle*, 7 (20):3118–3123, October 2008.

Dmitri Talantov, Abhijit Mazumder, Jack X Yu, Thomas Briggs, Yuqiu Jiang, John Backus, David Atkins, and Yixin Wang. Novel genes associated with malignant melanoma but not benign melanocytic lesions. *Clinical Cancer Research*, 11(20): 7234–7242, October 2005.

Adi L Tarca, Sorin Draghici, Purvesh Khatri, Sonia S Hassan, Pooja Mittal, Jung-Sun Kim, Chong Jai Kim, Juan Pedro Kusanovic, and Roberto Romero. A novel signaling pathway impact analysis. *Bioinformatics (Oxford, England)*, 25(1):75–82, January 2009.

Adi L Tarca, Sorin Draghici, Gaurav Bhatti, and Roberto Romero. Down-weighting overlapping genes improves gene set analysis. *BMC Bioinformatics*, 13(1):136, June 2012.

Adi L Tarca, Gaurav Bhatti, and Roberto Romero. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PloS one*, 8(11): e79217, January 2013.

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. *Nature Biotechnology*, 27(2):199–204, March 2009.

Lu Tian, Steven A Greenberg, Sek Won Kong, Josiah Altschuler, Isaac S Kohane, Peter J Park, and Sek Won. Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America*, 102(38):13544–13549, September 2005.

Nathan L Tintle, Alexandra Sitarik, Benjamin Boerema, Kylie Young, Aaron A Best, and Matthew Dejongh. Evaluating the consistency of gene sets used in the analysis of bacterial gene expression data. *BMC Bioinformatics*, 13(1):193, January 2012.

John Tomfohr, Jun Lu, and Thomas B Kepler. Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, 6:225, January 2005.

Virginia G Tusher, Robert Tibshirani, and Gil Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9):5116–5121, April 2001.

Christine Vogel and Edward M Marcotte. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics*, 13(4):227–232, April 2012.

Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, January 2009.

Patrick Warnat, Roland Eils, and Benedikt Brors. Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics*, 6(1):265, January 2005.

Thomas Weichhart and Marcus D Säemann. The PI3K/Akt/mTOR pathway in innate immune cells: emerging therapeutic applications. *Ann. Rheum. Dis.*, 67 (Suppl 3):iii70–74, December 2008.

Ian H Witten and Frank Eibe. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2005.

Katherine Wolstencroft, Robert Haines, Donal Fellows, Alan Williams, David Withers, Stuart Owen, Stian Soiland-Reyes, Ian Dunlop, Aleksandra Nenadic, Paul Fisher, Jiten Bhagat, Khalid Belhajjame, Finn Bacall, Alex Hardisty, Abraham Nieva de la Hidalga, Maria P Balcazar Vargas, Shoaib Sufi, and Carole Goble. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. *Nucleic acids research*, 41(Web Server issue): W557–561, July 2013.

David J Wong, Helen Liu, Todd W Ridky, David Cassarino, Eran Segal, and Howard Y Chang. Module map of stem cell genes guides creation of epithelial cancer stem cells. *Cell Stem Cell*, 2(4):333–344, April 2008.

Guanghua Xiao, Betsy Martinez-Vaz, Wei Pan, and Arkady B Khodursky. Operon information improves gene expression estimation for cDNA microarrays. *BMC Genomics*, 7:87, January 2006.

Sam S Yoon, Neil H Segal, Peter J Park, Kara Y Detwiller, Namali T Fernando, Sandra W Ryeom, Murray F Brennan, and Samuel Singer. Angiogenic profile of soft tissue sarcomas based on analysis of circulating factors and microarray gene expression. *Journal of Surgical Research*, 135(2):282–290, October 2006.

Aidong Zhang. *Advanced Analysis of Gene Expression Microarray Data*. World Scientific, 2006.

Xionghui Zhou, Juan Liu, Xinghuo Ye, Wei Wang, and Jianghui Xiong. Ensemble classifier based on context specific miRNA regulation modules: a new method for cancer outcome prediction. *BMC bioinformatics*, 14(Suppl 1):S6, January 2013.

Elias Zintzaras and Axel Kowald. Forest classification trees and forest support vector machines algorithms: Demonstration using microarray data. *Computers in Biology and Medicine*, 40(5):519–524, May 2010.

Information about the author and thesis

Thesis related publications

Impacted journal papers

- Matěj Holec, Jiří Kléma, Filip Železný, and Jakub Tolar. Comparative evaluation of set-level techniques in predictive classification of gene expression samples. *BMC Bioinformatics*, 13(Suppl 10):S15, January 2012. **35%**, 2 times cited

Web of Science-indexed papers

- Jiří Bělohradský, David Monge, Filip Železný, Matěj Holec, and Carlos Garcia Garino. Template-based semi-automatic workflow construction for gene expression data analysis (poster paper). In *Proceedings of the 24th International Symposium on Computer-Based Medical Systems*, 2011. **5%**

- Ondřej Kuželka, Andrea Szabóová, Matěj Holec, and Filip Železný. Gaussian logic for predictive classification. In *ECML/PKDD 2011: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2011. **10%**

- Jiří Kléma, Matěj Holec, Filip Železný, and Jakub Tolar. Comparative evaluation of set-level techniques in microarray classification. In Jianer Chen, Jianxin Wang, and Alexander Zelikovsky, editors, *Bioinformatics Research and Applications*, volume 6674 of *Lecture Notes in Computer Science*, pages 274–285. Springer Berlin Heidelberg, 2011. **30%**, 2 times cited

- Matěj Holec, Filip Železný, Jiří Kléma, and Jakub Tolar. Integrating multiple-platform expression data through gene set features. In Ion Mandoiu, Giri Narasimhan, and Yanqing Zhang, editors, *Bioinformatics Research and Applications*, volume 5542 of *Lecture Notes in Computer Science*, pages 5–17. Springer Berlin Heidelberg, 2009. **40%**, 5 times cited

The participation percentage were obtained from the VVVS system.

Other publications

- Matěj Holec, Valentin Gologuzov, and Jiří Kléma. miXGENE tool for learning from heterogeneous gene expression data using prior knowledge. In *Proceedings of the 27th International Symposium on Computer-Based Medical Systems*, pages 247–250, 2014. **40%**
- Ondřej Kuželka, Andrea Szabóová, Matěj Holec, and Filip Železný. Gaussian logic for proteomics and genomics. In *MLSB 2011: the 5th International Workshop on Machine Learning in Systems Biology*, 2011. **10%**
- Matěj Holec, Filip Železný, Jiří Kléma, and Jakub Tolar. A comparative evaluation of gene set analysis techniques in predictive classification of expression samples. In *International Conference on Bioinformatics, Computational Biology, Genomics and Chemoinformatics (BCBGC-2010)*, 2010. **50%**
- Matěj Holec, Jiří Kléma, Filip Železný, Jiří Bělohradský, and Jakub Tolar. Cross-genome knowledge-based expression data fusion. In *International Conference on Bioinformatics, Computational Biology, Genomics, and Chemoinformatics*, 2009. **30%**
- Matěj Holec, Filip Zelezny, Jiří Kléma, Jiří Svoboda, and Jakub Tolar. Using Bio-Pathways in Relational Learning. In Filip Železný and Nada Lavarač, editors, *Late Break. Pap. 18th Int. Conf. Inductive Log. Program.*, 2008. **30%**

Unrelated publications

Other publications

- Ondřej Kuželka, Matěj Holec, Jiří Matas, and Filip Železný. Systém pro výpočet generálního a hlavního klíče. Technical report, České vysoké učení technické, Fakulta elektrotechnická, Katedra kybernetiky – Centrum strojového vnímání, Praha, 2013. **25%**
- Jiří Matas, Filip Železný, Ondřej Kuželka, Matěj Holec, and Radomír Černoch. Key/lock system computation: progress report. Technical report, České vysoké učení technické, Fakulta elektrotechnická, Katedra kybernetiky – Centrum strojového vnímání, Praha, 2013. **20%**