FROM CORRELATION TO CAUSALITY: DOES NETWORK INFORMATION IMPROVE CANCER OUTCOME PREDICTION?

Dissertation

zur Erlangung des akademischen Grades
Doctor rerum naturalium (Dr. rer. nat)

vorgelegt an der
Technischen Universität Dresden
Fakultät Informatik

von

DIPL. BIOINF. JANINE ROY
geboren am 24. August 1982 in Herzberg/Elster

Betreuender Hochschullehrer:	Prof. Dr. Michael Schroeder Technische Universität Dresden
Gutachter:	Prof. Dr. Tim Beißbarth Universität Göttingen
Tag der Verteidigung:	16. 04. 2014

Dresden, im Juni 2014

Janine Roy: *From Correlation to Causality: Does network information improve cancer outcome prediction?*, PhD Thesis, © 2014, February.

WEBSITE:
`http://biotec.tu-dresden.de/Schroeder`

E-MAIL:
`janine.roy@biotec.tu-dresden.de`

Enjoy reading!

SUMMARY

MOTIVATION Disease progression in cancer can vary substantially between patients. Yet, patients often receive the same treatment. Recently, there has been much work on predicting disease progression and patient outcome variables from gene expression in order to personalize treatment options. A widely used approach is high-throughput experiments that aim to explore predictive signature genes which would provide identification of clinical outcome of diseases. Microarray data analysis helps to reveal underlying biological mechanisms of tumor progression, metastasis, and drug-resistance in cancer studies. Despite first diagnostic kits in the market, there are open problems such as the choice of random gene signatures or noisy expression data. The experimental or computational noise in data and limited tissue samples collected from patients might furthermore reduce the predictive power and biological interpretability of such signature genes. Nevertheless, signature genes predicted by different studies generally represent poor similarity; even for the same type of cancer.

Integration of network information with gene expression data could provide more efficient signatures for outcome prediction in cancer studies. One approach to deal with these problems employs gene-gene relationships and ranks genes using the random surfer model of Google's PageRank algorithm. Unfortunately, the majority of published network-based approaches solely tested their methods on a small amount of datasets, questioning the general applicability of network-based methods for outcome prediction.

METHODS In this thesis, I provide a comprehensive and systematically evaluation of a network-based outcome prediction approach – NetRank - a PageRank derivative – applied on several types of gene expression cancer data and four different types of networks. The algorithm identifies a signature gene set for a specific cancer type by incorporating gene network information with given expression data. To assess the performance of NetRank, I created a benchmark dataset collection comprising 25 cancer outcome prediction datasets from literature and one in-house dataset.

RESULTS NetRank performs significantly better than classical methods such as foldchange or *t*-test as it improves the prediction performance in average for 7%. Besides, we are approaching the accuracy level of the authors' signatures by applying a relatively unbiased but fully automated process for biomarker discovery. Despite an order of magnitude difference in network size, a regulatory, a protein-protein interaction and two predicted networks perform equally well.

Signatures as published by the authors and the signatures generated with classical methods do not overlap – not even for the same cancer type – whereas the network-based signatures strongly overlap. I analyze and discuss these overlapping genes in terms of the Hallmarks of cancer and in particular single out six transcription factors and seven proteins and discuss their specific role in cancer progression. Furthermore several tests are conducted for the identification of a Universal Cancer Signature. No Universal Cancer Signature could be identified so far, but a cancer-specific combination of general master regulators with specific cancer genes could be discovered that achieves the best results for all cancer types.

As NetRank offers a great value for cancer outcome prediction, first steps for a secure usage of NetRank in a public cloud are described.

CONCLUSION Experimental evaluation of network-based methods on a gene expression benchmark dataset suggests that these methods are especially suited for outcome prediction as they overcome the problems of random gene signatures and noisy expression data. Through the combination of network information with gene expression data, network-based methods identify highly similar signatures over all cancer types, in contrast to classical methods that fail to identify highly common gene sets across the same cancer types.

In general allows the integration of additional information in gene expression analysis the identification of more reliable, accurate and reproducible biomarkers and provides a deeper understanding of processes occurring in cancer development and progression.

PUBLICATIONS AND CONTRIBUTIONS

The research done during my thesis led to the following publications:

1. **J Roy**, C Winter, Z Isik, M Schroeder.
   Network information improves cancer outcome prediction. In Briefings in Bioinformatics. 2012

   Contribution: This starting paper is described in chapter 4. The design of the study, performing the experiments as well as the main analysis of the data was led by me.

2. C Winter, G Kristiansen, S Kersting, **J Roy**, D Aust, T Knösel, P Rümmele, B Jahnke, V Hentrich, F Rückert, M Niedergethmann, W Weichert, M Bahra, HJ Schlitt, U Settmacher, H Friess, M Büchler, HD Saeger, M Schroeder, C Pilarsky, R Grützmann.
   Google goes cancer: Improving outcome prediction for cancer patients by network-based ranking of marker genes. In PLOS Computational Biology. 2012.

   Contribution: My contribution to this publication is shown in chapter 4. I mainly supported the analysis of the data, regarding the performance of the proposed algorithm.

3. **J Roy**, Z Isik, C Pilarsky and M Schroeder.
   Can specific transcriptional regulators assemble a Universal Cancer Signature? International Symposium on Computational Models in Life Science, November 2013, Sydney, published in AIP Conference Proceedings, 2013

   This publication won the best paper prize at the conference.

   Contribution: Results from this publication are shown in chapter 5. My contributions are the design of the study, the execution of the experiments as well as the analysis of the data.

4. M Beck, VJ Haupt, **J Roy**, J Moennich, R Jäkel, M Schroeder, and Z Isik.
   GeneCloud: Secure Cloud Computing for Biomedical Research, In Lecture Notes in Computer Science, 2014

   Contribution: Analysis of the requirements for calculation in a public cloud as well as implementation of the proposed secure algorithm. The results from this analysis are shown in chapter 6.

LIST OF FIGURES

Figure 1 NetRank general workflow 22

Figure 2 Workflow of RRFE 23

Figure 3 Workflow from netSAM 24

Figure 4 The general idea of NetRank 27

Figure 5 Detailed workflow of NetRank 29

Figure 6 Cancer types in the benchmark dataset 38

Figure 7 Detailed analysis of benchmark dataset 39

Figure 8 Comparison of NetRank with a direct neighbor algorithm 42

Figure 9 Distribution of expression levels and correlation with survival in the pancreas cancer dataset 43

Figure 10 NetRank accuracy vs. damping factor 45

Figure 11 NetRank accuracy vs. outcome task 58

Figure 12 Accuracy vs. cohort and signature size 59

Figure 13 Accuracy heatmap of transferred signatures with different machine learning methods 61

Figure 14 Signature similarity of original signatures 68

Figure 15 Signature similarity of greedy techniques 70

Figure 16 Signature similarity for HPRD-based NetRank genes 71

Figure 17 Signature similarity Transfac-based NetRank genes 71

Figure 18 Heatmap of the $-\log(p - value)$ of the Jaccard Index of Transfac and HPRD-based NetRank genes 72

Figure 19 Top selected Transfac genes are strongly interconnected 78

Figure 20 Top selected HPRD genes are in-between connected 81

Figure 21 Prediction accuracies of individual NetRank signatures against all datasets 91

Figure 22 GeneCloud project description 98

Figure 23 Randomization of vectors 100

LIST OF TABLES

Table 1 Investigated single gene tumor markers 8
Table 2 Tests for breast cancer outcome prediction 14
Table 3 State-of-the-art network-based methods 19
Table 4 Summary of benchmark dataset 38
Table 5 Cancer specific damping factor 46
Table 6 NetRank results on HPRD and Transfac 48
Table 7 NetRank results on STRING and hPrint 51
Table 8 Comparison of NetRank with other network-based methods 53
Table 9 The prediction accuracies of the classical methods by using only Transfac genes 56
Table 10 Examples of running time for one cross validation 63
Table 11 Signatures for each dataset using NetRank on Transfac 74
Table 12 Signatures for each dataset using NetRank on HPRD 75
Table 13 Literature analysis of top NetRank genes for Transfac 76
Table 14 Literature analysis of top NetRank genes for HPRD 79
Table 15 Hallmarks of cancer covered by signature genes based on the Transfac network 83
Table 16 Hallmarks of cancer covered by signature genes based on the HPRD network 84
Table 17 Prediction accuracies achieved by two possible Universal Cancer Signatures 87
Table 18 Description of Transfac and HPRD genes tested as a possibel Universal Cancer Signature 89
Table 19 Prediction results of a combination of Transfac and HPRD genes 90
Table 20 Influence of general and specific genes on the prediction results 92
Table 21 Overlap of anti-profile signature with NetRank genes 93

CONTENTS

1 OPEN PROBLEMS 1

2 INTRODUCTION 5
- 2.1 Problems in cancer outcome prediction 5
- 2.2 Network-based cancer outcome prediction 17
- 2.3 Universal Cancer Signature 25

3 METHODS 27
- 3.1 NetRank algorithm 27
- 3.2 Preprocessing and filtering of the microarray data 31
- 3.3 Accuracy 32
- 3.4 Signature similarity 32
- 3.5 Classical approaches 32
- 3.6 Random signatures 34
- 3.7 Networks 34
- 3.8 Direct neighbor method 35
- 3.9 Dataset extraction 35

4 PERFORMANCE OF NETRANK 37
- 4.1 Benchmark dataset for evaluation 37
- 4.2 The influence of NetRank parameters 40
- 4.3 Evaluation of NetRank 47
- 4.4 General findings 54
- 4.5 Computational complexity of NetRank 62
- 4.6 Discussion 64

5 UNIVERSAL CANCER SIGNATURE 67
- 5.1 Signature overlap – a sign for Universal Cancer Signature 67
- 5.2 NetRank genes are highly connected and confirmed in literature 73
- 5.3 Hallmarks of Cancer 82
- 5.4 Testing possible Universal Cancer Signatures 85
- 5.5 Conclusion 94

6 CLOUD-BASED BIOMARKER DISCOVERY 97
- 6.1 Introduction to secure Cloud computing 97
- 6.2 Cancer outcome prediction 98
- 6.3 Security analysis 99
- 6.4 Conclusion 101

7 CONTRIBUTIONS AND CONCLUSION 103

BIBLIOGRAPHY 111

1 OPEN PROBLEMS

OPEN PROBLEM 1: DOES NETWORK INFORMATION IMPROVE CANCER OUTCOME PREDICTION?

MOTIVATION In the last decade microarrays developed to be a powerful tool to predict the outcome of patients in several diseases. In contrast to standard DNA-based assays that mainly focuses on single genes with rare conditions, microarray based methods are perfect for the investigation of diseases underlying complex genetic causes. Many recent microarray studies have been accepted by the Food and Drug Administration (FDA) and are nowadays included in clinical routines. These studies cover a wide range of malignant and chronic diseases and improve the live span and quality of many patients. However, there is an even bigger number of studies that did not found acceptance as they did not found their way into clinical usage.

Predictive signatures are often not consistent

A huge topic in outcome prediction based on large scale microarray experiments is the transferability of signatures in different datasets, the applicability in other patients than in the patients of the study. Generally, predictive gene sets derived from different studies for the same disease tend to have zero overlap, questioning their biological relevance.

Low statistical power of microarrays

The problem lies in the microarray experiment itself. Due to limited financial support, many studies are based on a limited amount of patients, which of course cannot reflect all possible disease variations. In addition differs the gene expression between patients but the differences are often subtle and the technical noise introduced during experiments often extends the biological signal.

Integration of network information improves cancer outcome prediction

Network information can help to improve outcome prediction and reduce noise in microarray experiment. Many recent studies applied network based methods and successfully improved outcome prediction of already published microarray data [40, 42, 45, 61, 129, 172, 181, 196].

To our knowledge, all published network-based approaches test their methods solely on a small amount of datasets to indicate their functionality. In this thesis I will test the general applicability of network-based outcome prediction on a benchmark dataset covering a wide range of cancer types. For this purpose I evaluate a network-based algorithm – NetRank – on 26 gene expression datasets.

OPEN PROBLEMS The following questions are answered in this PhD thesis:

- Is NetRank able to successfully predict the outcome of several cancer types?
- How does the NetRank accuracies compare to:
    1. random signatures,
    2. signatures obtained from classical outcome prediction methods and
    3. the original accuracies as published in the benchmark datasets?
- Does network outcome prediction provide consistent improvements across different cancer types?
- Does the size of the signature and patient cohort influence the results?
- Which types of interactions – physical or regulatory – are best suited? Are current networks sufficiently large?
- Which parameters influence the prediction results of NetRank?

OPEN PROBLEM 2: IS THERE A UNIVERSAL CANCER SIGNATURE?

MOTIVATION Cancer outcome prediction aims to forecast disease progression from gene expression data. In the last years, many studies published predictive signatures that allow refined outcome prediction and thus leading to better treatment options. It has been shown, that single gene markers are not able to reliably predict cancer outcome. For that reason a combination of several genes is nowadays used as predictive signatures for outcome prediction. One success in this area is the diagnostic assay MammaPrint®, which predicts metastasis formation in breast cancer based on the expression of 70 genes. The original publication claims that these genes predict the metastasis probability of a patient with an accuracy of 83% [191], which was later corrected to 69% [120].

No signature overlap for same type of cancer

Ein-Dor and co-workers compared the breast cancer progression signature of MammaPrint® with another study and found that ″the overlap between these gene sets is almost zero″ and that ″many equally predictive lists could have been produced″ [55]. Venet and co-workers continued these lines and showed in a systematic analysis that 60% of predictive signatures are not better than random and that 90% of all random signatures with more than 100 genes would be a comparably good predictor [188].

Integration of background knowledge improves cancer outcome prediction

Network information efficiently helps to improve outcome prediction and reduce noise in microarray experiments. In chapter 4 I evaluate a network-based approach on several cancer datasets and the resulting signatures showed overlapping key genes. This strengthen the results of another study that found a high correlation between the predictive genes of two breast cancer datasets and claimed that predictive signatures would reveal underlying mechanisms of certain diseases [54].

In the last decade, the several stages of cancer development have been a focus in biomedical research. Hanahan and Weinberg have summarized the biological capabilities acquired during the multistage development of human cancers in the Hallmarks of cancer [79].

OPEN PROBLEMS Several questions arise based on the similarity of predictive signatures:

- How similar are:
  1. signatures published by other authors regarding the same and different biological questions?
  2. NetRank signatures intra and inter cancer types?
- Are overlapping genes in NetRank signatures biological meaningful?

- In which relation are these genes to the Hallmarks of cancer?
- Is it possible to construct a Universal Cancer Signature that is able to detect any kind of cancer by using the overlapping genes?
- Do we need individual signatures for cancer outcome prediction?
- But then, how do we improve the overlap between signatures obtained from different datasets?

2 INTRODUCTION

2.1 PROBLEMS IN CANCER OUTCOME PREDICTION

Cancer remains the second leading cause of death in the United States, behind heart disease, with an estimated 1.6 million new cases and 580.000 deaths in 2013 alone [165]. It is a highly complex disease, which can encompass multiple genomic alterations, including gene amplifications, epigenetic modifications, point mutations, translocations, aberrant splicing, deletions, and altered gene expression. These changes may be somatically acquired or inherited during progression from a normal to a cancerous cell. In the past decade, it has been discovered how these genomic perturbations drive cancer cell survival by altering the mechanism for apoptosis, cell cycle control, DNA repair, differentiation, and metabolism. By improving the understanding of these molecular mechanisms, scientists have gained greater insight into the initiation of cancer, its progression, and its sensitivity to therapeutics.

2.1.1 Goal of cancer outcome prediction

Outcome prediction tries to define the future state of a patient based on its current disease state. The main goal of cancer outcome prediction is to improve the diagnosis and the treatment of cancer through more accurate disease classification and patient stratification. Outcome-based cancer research spans from discovery research to validation and into clinical utility including:

1. Identification of cancer biomarkers and therapeutic targets,
2. Elucidation of the mechanisms of cancer pathways,
3. Validation of therapeutic targets and cancer biomarkers,
4. Clinical classification and stratification.

2.1.2 Cancer staging – the standard way of outcome prediction

Staging describes the severity of a person's cancer based on the size and/or extent of the primary tumor and whether the cancer has spread in the body. After initial surgery, the state of a tumor is microscopically examined by a pathologist and predictions are made using different parameters.

To determine the degree to which a cancer has developed is possible due to the tumor-node-metastasis (TNM) staging system, which is a measure of the extent to which the cancer has spread and the grading of a tumor that is a measure of cell anaplasia in the sampled tumor and is based on the resemblance of the tumor to the tissue of origin.

The tumor grading system, described by the American Joint Commission on Cancer guidelines uses numerals I, II, III, and IV to characterize the progression of cancer.

The TNM staging system is the standard way to access the prognosis of a patient

The TNM staging system includes several parameters:

1. primary tumor pT: describes the size of the original tumor and whether it has invaded nearby tissue,

2. regional lymph nodes pN: describes whether nearby (regional) lymph nodes are involved,

3. distant metastasis M: describes whether distant metastasis occur,

4. histological grade G: reflects the grade of the cancer cells,

5. residual tumor R: describes the tumor status following treatment and reflects the effects of treatment.

These measures aid medical staff in categorizing the tumor. This categorization is helpful in treatment planning and for assessing the prognosis of a patient. They furthermore assist in the evaluation of the results of treatment and help health care providers and researchers to exchange information about patients.

TNM staging has different drawbacks

The usability of this system for outcome prediction has been questioned [28] and it has been shown for several cancer types that the system is not fine grained enough for effective outcome prediction [93, 140]. A new system is needed to reliable predict the outcome of a patient. Gene expression profiling adds a great value to the existing TNM grading system.

2.1.3 Single gene tumor marker

Besides the TNM grading system, many gene markers are nowadays in use to guide clinical doctors in making treatment decisions. The following sections describe some biological markers used in daily routine as well as the advantages and disadvantages of single gene tumor marker.

A single tumor marker can be evaluated easily through simple, noninvasive and cost-effective laboratory tests, yet facilitates earlier diagnoses and improved treatment outcomes. A single tumor marker consists of any product of either the tumor itself or the host in reaction to the tumor's presence, which distinguishes malignant tissues from healthy and is measurable in body fluids or tissues.

Usually, tumor markers cover one of the following 5 different types of proteins and can be either:

1. Oncofetal proteins, such as carcinoembryonic antigen (CEA) and alpha-fetoprotein (AFP).

2. Hormones, such as calcitonin and human chorionic gonadotropin (HCG).

3. Organ-specific antigens, such as prostate-specific antigen (PSA).

4. Monoclonal antibody-defined antigens, such as tumor associated glycoproteins CA 125, CA 19-9, CA 27-29, and CA 15-3.

5. Enzymes, such as prostatic acid phosphatase.

The expression/amount of a protein or single gene tumor marker can be measured in several ways. As proteins are molecules with a 3D-structure, direct measurement is not easily possible. Proteins are encoded in the DNA that is transcribed into mRNA, which is later on translated into the protein. Through measuring the quantity of mRNA, the expression of a protein can be estimated. A protein is called up-regulated, when more mRNA is measured than in a comparable state and down-regulated vice versa.

Single gene markers in practice

Nowadays, there are many single gene markers in clinical use. Table 1 lists the most used single tumor marker tests nowadays available to medical doctors and their clearance by the FDA. The following section describes some of the single gene marker frequently used in daily clinical routine.

ALPHA-FETOPROTEIN Hepatocellular carcinoma (HCC) is one of the most often occurring cancers worldwide. It develops usually in a liver suffering already chronic damages as cirrhosis or a viral hepatitis infection. Alpha-fetoprotein (AFP) is a major mammalian embryo-specific and tumor-associated antigen, which progressive elevation helps to diagnose hepatocellular carcinoma in patients with liver cirrhosis [175]. Expression levels of AFP are also influenced by other diseases like hepatitis C or cirrhosis [75, 143], therefore it cannot not be used to specifically diagnose HCC. For that reason and because of having poor sensitivity and specificity the usability of AFP for early-stage diagnosis is limited. Nevertheless, AFP is a useful marker for post-treatment control like tumor recurrence [205].

Table 1: Investigated tumor markers. The last column indicates whether the marker is approved by the FDA. Information taken from `www.cancer.gov` and `www.cancer.org`

Marker	Tumor	FDA
Alpha-fetoprotein	Liver cancer, germ cell tumours, ovarian cancer	✓
BCR-ABL	Chronic myeloid carcinoma	
Beta-2-microglobulin	Multiple myeloma, chronic lymphocytic leukemia, and some lymphomas	
Bladder tumor antigen	Bladder cancer	✓
BRAF	Cutaneous melanoma, colorectal cancer, thyroid cancer	
CA 15-3	Breast cancer	
CA 19-9	Pancreatic cancer, gallbladder cancer, bile duct cancer, and gastric cancer	
CA 27-29	Breast cancer	
CA 125	Epithelial ovarian cancer	✓
Calcitocin	Medullary thyroid carcinoma	
Carcinoembryonic antigen (CEA)	Colorectal cancer, lung cancer, breast cancer	✓
Chomogranin A	Neuocrine tumors (carcinoid tumor, neuroblastoma, small cell lung cancer)	
Estrogen receptor (ER)	Breast cancer	✓
EGFR/HER1	Breast cancer, non-small cell lung cancer, head and neck cacner, colon, pancreas	
HER2/ERBB2/EGFR2	Breast cancer, gastric cancer, and esophageal cancer	
Human chorionic gonadotropin (HCG)	Ovarian cancer, germ cell tumor, choriocarcinoma	
IgA/IgG/IgD/IgM	Bone marrow cancer	
K-Ras	Colorectal cancer, lung cancer	
Nuclear Mitotic Apparatus protein (NMP22)	Bladder cancer	✓
PSA	Prostate cancer	✓
Prostatic acid phosphatase (PAP)	Prostate cancer	
S-100	Melanoma	
Thyroglobulin	Thyroid cancer	

CARCINOEMBRYONIC ANTIGEN The Carcinoembryonic antigen (CEA) is a protein involved in cell adhesion. It is only present at very low levels in the blood of healthy adults, but higher levels occur in some types of cancer. It is specially used in monitoring colorectal cancer treatment, to identify recurrence as well as to localize cancer spreading [2]. Elevated CEA levels have been also found in other cancerous conditions as gastric carcinoma, pancreatic carcinoma, lung carcinoma, breast carcinoma, as well as some non-neoplastic conditions like ulcerative colitis, pancreatitis, cirrhosis [117].

CANCER ANTIGEN 125 The carcinoma antigen 125 (CA-125) also known as mucin 16 is a member of the mucin family glycoprotein. It is frequently used as tumor marker in ovarian cancer detection. CA-125 is a powerful tool for detecting ovarian cancer as around 90% of women with ovarian cancer show elevated levels of CA-125 in their blood [74]. It is furthermore useful to measure CA-125 levels to access the effectiveness of a certain cancer treatment as the disease-free survival correlates with the decrease of CA-125 in blood [16]. Moreover, Göcze and co-workers showed, that high levels during cancer treatment are associated with poor survival outcome [70]. Using CA-125 as a tool for early cancer detection has been controversially discussed [15, 66]. The major problem is the low sensitivity in early cancer stages as well as the lack of specificity in premenopausal women [131]. These constraints lead to false positives, introducing patients to unnecessary screening procedures. CA-125 plays a crucial role in advancing tumorgenesis and tumor proliferation. It helps the tumor to evade the immune system by suppressing the response to natural killer cells [142]. Furthermore it allows the formations of metastasis by promoting cell motility [176] and it reduces the sensitivity of the tumor cell to cancer treatment [23].

PROSTATE-SPECIFIC ANTIGEN Prostate-specific antigen (PSA) is a glycoprotein secreted by the epithelias cells of the prostate glands. It is needed to liquefy the semen allowing sperm to swim freely [12]. The PSA test is approved in the U.S. for annual screening of prostate cancer for men older than 50 years. It is normally present in the blood at very low levels, while increased levels of PSA indicate the presence of prostate cancer. However, prostate cancer can occur with no elevation of PSA [177], whereas obesity, prostatitis or benign prostatic hyperplasia (non-cancerous growth of the prostate) can lead to elevated levels of PSA [13, 126, 187]. This may lead to many false negatives/positives. For that reason, the usefulness of the PSA test is controversially discussed [1] .

[1] U.S. Preventive Services Task Force http://www.uspreventiveservicestaskforce.org/prostatecancerscreening/prostatefinalrs.htm

Advantages of single gene markers

Single gene markers help to monitor success of therapy

Expression levels of single gene markers are easy to access

The measurement of tumor associated protein levels in numerous body fluids or tissues can aid in the monitoring of certain cancers. Monitoring is defined here as estimating the progression of tumor growth or as assessing the response of cancer cells to therapy. This includes the sequential measurement of patients with confirmed diagnoses who are undergoing cancer therapy. Increasing tumor marker concentrations are often indicative for a progressive disease, whereas decreasing concentrations indicate response to therapy. Constant serum tumor marker levels are widely associated with a stable disease. Monitoring can also mean serial measurements used to detect recurrent or residual disease in patients, who are following primary curative treatment. Continuous elevations in marker concentrations are suggestive of residual cancer, while increasing concentrations are characteristic of recurring disease. Single gene markers are also easy to access via non-costly laboratory test as blood tests, real-time polymerase chain reaction (RT-PCR) and immunohistochemistry.

Drawbacks of single gene markers

Single gene marker show often high levels in non-cancerous states

Single gene markers come with a main drawback. Most tumor markers are expressed in normal as well as cancer cells and often noncancerous diseases lead to higher levels of certain tumor markers as well. Furthermore, not all cancer types express the tumor marker in a stable amount. It has been found that some patients show no expression, whereas elevated levels have been measured in healthy persons. Due to these reasons, using single gene markers seldom leads to high sensitivity and specificity in cancer diagnosis.

Heterogenity of expression in-between but also within tissues

Another problem for early-cancer diagnosis is the detection limit of proteins in blood. Many proteins are expressed in low abundance in blood, therefore the measurement is difficult. Large amounts of cancer cells present in the blood are needed to overcome the detection limits. It is furthermore well accepted that gene expression differs in-between tissues. But it has been recently shown, that also within tissues, even in the same organ or tumor, the gene expression differs and that there might be a lot of heterogeneity. Rodriguez-Gonzalez and co-workers investigated the heterogeneity in the case of EpCAM, which is a cell surface molecule extensively used as marker for the enrichment of epithelial cancer cells in blood. They show that the expression of EpCAM in breast cancer can be heterogeneous spread over the same tumor [152].

2.1.4 Gene expression profiling with high-density oligonucleotide arrays - the new way for outcome prediction

To overcome the problems of single gene biomarkers, a combination of several genes can be used. The following section describes a technique – the high-density oligonucleotide arrays or DNA-Microarrays – that helps to measure the expression of thousands of genes simultaneously.

The assays can measure gene expression from any biological sample - from tissue as well as body liquids. As a result, the researcher obtains a gene expression profile of such a sample at a particular time, in a single experiment. The comparison of such profiles created from healthy and cancerous tissue, helps researchers to identify proteins involved in cancer formation and progression as wells as to identify how cells react to a particular treatment. The usual goal of such a study is the identification of a set of genes that change their expression. It is possible with such a set of genes to predict the outcome of a patient or to understand the behavior of a cell regarding external stimuli. A gene is called down-regulated if it has a decreased expression compared to another tissue or disease state whereas it is called up-regulated with an increased expression.

First experiments

Numerous studies have used gene expression profiling to describe tumor samples from different tissues and cancer subtypes since its initial description by Brown's group. He and his co-workers made for the first time use of DNA molecules spotted onto glass slides to measure the expression levels of thousands of mRNAs simultaneously [157].

The first complete eukaryotic genome (*Saccharomyces cerevisiae*) on a microarray chip was published in 1997 [101]. Over the past decade, global gene expression profiling has become a standard tool for pathway and biomarker discovery. In the past 15 years, almost 20.000 peer-reviewed publications attest to the value of the high-density oligonucleotide array technology in the area of life sciences.

The structure of microarrays

A typical chip for gene expression profiling of the human genome is the Affymetrix HG-U133 Plus 2.0 Array, which measures most of the currently 22.000 known human genes. This microarray chip is a small (a few centimeters on each side) glass or other solid surface on which millions of immobilized oligonucleotide probes have been synthesized or robotically deposited in a predetermined array, resulting in a density of approximately 1 million probes per $1.3cm^2$. A probe is a particular 25-nucleotide long single-strand DNA, and 11 to 20 probes typically correspond to one probeset. To achieve the most reliability and avoid errors, one gene or expressed sequence tag (EST) is represented by one to several probe sets.

The workflow of a microarray experiment

The basic principle behind microarrays is the hybridization between two DNA strands that have the property of nucleic acid sequences to specifically pair with each other by forming hydrogen bonds between complementary nucleotide base pairs. When performing a microarray experiment, the mRNA from the sample has to be extracted, fragmented and copied into complementary DNA (cDNA) or complementary RNA (cRNA). Afterwards each fragment is fluorescently labeled and incubated to the chip. During incubation, the cDNA or cRNA fragments will hybridize to their matching counterpart probes on the microarray surface. After washing, fluorescent molecules are excited by using a laser and the measurement takes place by counting of the emitted photons. The entire chip will be scanned and the result is a digital image, where signal intensities for each probe are measured and a raw intensity for each probe set is assigned. The results of the experiment are raw measurements of fluorescence intensities.

To ensure that differences in intensities are truly due to differential expression, not printing, hybridization, or scanning artifacts, preprocessing has to take place. Preprocessing consists of different steps of background correction, normalization and summarization, which lead to a final expression value for each probe set.

Several methods for preprocessing have been published in the last decade. The most popular are MAS5 [3], Model-based Expression Index (MBEI) [109], Robust Multi-array Average (RMA) [87] and GC Robust Multi-array Average (GCRMA) [199]. In numerous comparisons none of the methods turned out to be generally superior to any other method [21, 76, 190, 198].

2.1.5 The role of microarrays in cancer outcome prediction

Cancer is a broad group of diseases in which genetic changes drive cells to divide and grow uncontrolled [193]. One method in cancer outcome prediction is the analysis of gene expression with the help of high-density oligonucleotide arrays. Due to the application of the microarray technology in medical research, numerous studies were conducted to shed light on gene expression alterations in cancer cells. Four different applications can be investigated though comparison of RNA expression levels:

1. Diagnosis of a Disease - Comparing cancer tissue with healthy tissue,

2. Prediction of success/failure of treatment - Comparing cancer cells treated with different reagents,

3. Subtyping different types of cancer - Comparing tissue samples of the same type of cancer,

4. Prediction of tumor progression - Comparing gene expression in a patient tissue sample with a clinical parameter.

During diagnosis, genes are identified whose changes in expression correlate with certain cancer phenotypes. In contrast, for prediction of treatment response, genes and proteins that are involved in the mode of action of that reagent or cause therapy related side effects such as resistance, have to be identified. Treatment response includes a prolonged metastasis-free survival or faster response to therapy. Subtyping copes with finding genes that represent different subtypes of cancer. This is important as various subtypes may differ in the aggressiveness of the cancer along with different spreading behavior and may end up in different treatment needs. Finally, in prediction of cancer progression, genes have to be identified whose expression correlates with a clinical parameter of the patient such as metastasis.

Once a signature of genes is found for one of these applications, the medical doctor would execute a lab test measuring expression of the signature genes for a patient, and use this information for his decision on how to start/proceed with a therapy.

Gene expression-based applications in clinical use

The following section describes different clinical application used (or in application to be used) for cancer outcome prediction in a medical environment. The majority of these tests has been proven successfully and is nowadays used in the daily clinical routine. All assays are based on the measurement of gene expression.

Table 2: Several tests assesseing the outcome of breast cancer are nowadays available. Table is adapted from [7].

Test	Reference	No. Genes	Output
MammaPrint®	van't Veer *et al.* [185]	70	2 Categories of tumors with different risk to develop metastasis at 10 years.
Oncotype DX^{TM}	Paik *et al.* [138]	21	Recurrence score: risk of 10-year distant recurrence in ER-positive, lymph node negative patients.
Theros-Breast Cancer Gene Expression Ratio Assay®	Ma *et al.* [116]	3	HOXB13:IL17R ratio stratifies ER-positive breast cancer into low or high risk for recurrence and is predictive of benefit from endocrine therapy.
PAM50/Breast BioClasifier^{TM}	Parker *et al.* [141]	55	Continuous risk of recurrence.
MapQuant^{TM}	Sotiriou *et al.* [167]	8	Genomic Grade Index Divides histologically defined G2 tumors.
Mammostrat®	Ring *et al.* [151]	5	Mammostrat risk score: high, moderate, or low risk of recurrence after tamoxifen treatment.

In the field of breast cancer alone, several tests are available. Table 2 shows different breast cancer outcome prediction tests nowadays available in clinical routine. The test of MammaPrint® and Oncotype DX^{TM}, as well as other non-breast cancer tests are explained in more detail in the following paragraph.

MAMMAPRINT® MammaPrint® is a microarray-based test that employs 70-genes developed by van't Veer *et al.* [185] to assess the risk of metastasis caused from breast cancer. The signature was published 2002 in Nature from the Netherlands Cancer Institute. In this publication, patients are classified by calculating the correlation coefficient between a patient's expression levels of the 70 genes and an average low risk (good prognosis) expression profile. If the correlation coefficient exceeds 0.4, the patient is classified as having a low risk, if less, the patient is classified as having a high risk of cancer recurrence. Two independent validation studies were conducted to assess the performance of the 70-gene signature. The validation by van de Vijver *et al.* with 151 patients showed a 10 year survival of 95% for low-risk, and of 55% for high-risk patients [191]. In an independent European validation with 302 patients [30], the accuracy for the metastasis-free survival at 10 years was 88% for low-risk patients, and 71% for high risk patients. The test was approved by the FDA in 2007 for use in the United States. The Mindact clinical trial is an international prospective phase 3 study investigating the clinical utility of the MammaPrint 70-gene expression signature [155]. Until 2011, more than 6600 patients have joined the clinical trial. One of the main goals of the study is to prove that low risk patients can safely spare chemotherapy.

ONCOTYPE DX^{TM} Oncotype DX^{TM} is a diagnostic test that quantifies the likelihood of disease recurrence in women with early-stage breast cancer and assesses the likely benefit from certain types of chemotherapy. Oncotype DX uses the expression levels of 21 genes within a tumor to determine a recurrence score, which lies between 0 and 100. The score represents the likelihood of breast cancer recurrence within 10 years after initial diagnosis. To retrieve the score formalin-fixed, paraffin-embedded tissue samples have to be sent to Genomic Health, where they are analyzed via RT-PCR. Five of the 21 genes are reference genes used to normalize the expression of the cancer genes. The remaining 16 genes are associated with cell proliferation, cellular invasion, HER2 and estrogen activity. Since Oncotype DX became available in 2004, it has been used by over 19,000 physicians to help to guide treatment for over 350,000 patients in 70+ countries[2]. The MammaPrint and the Oncotype DX test have only one gene in common.

ROCHE AMPLICHIP®CYP450 TEST The AmpliChip CYP450 Test, developed by Roche, performs genotyping of two Cytochrome P450 (CYP2D6 and CYP2C19) genes, which play a major role in the metabolism of an estimated 25% of all prescript drugs. It is intended to be an aid to clinicians in determining therapeutic strategies and treatment doses for therapeutics metabolized by the CYP2D6 or CYP2C19 gene product. The test uses purified genomic DNA extracted from whole blood. The assay distinguishes 29 known polymorphisms in the CYP2D6 gene, including gene duplication and gene deletion, as well as two major polymorphisms in the CYP2C19 gene. In 2004, the test was approved by the FDA.

[2] Information taken from `http://www.oncotypedx.com/en-US/Breast/PatientCaregiver/OncoOverview?sc_lang=en-GB`

PATHWORK® DIAGNOSTICS' TISSUE OF ORIGIN TEST The Pathwork® Tissue of Origin, from Affymetrix, is a microarray assay, that measures the expression pattern of more than 1,500 genes and compares them to the expression pattern of a panel of 15 malignant tumors. The data is compared to the following malignant expression patterns: bladder, breast, colorectal, gastric, hepatocellular, kidney, non-small cell lung, ovarian, pancreatic, prostate, and thyroid carcinomas, melanoma, testicular germ cell tumor, non-Hodgkins lymphoma, and soft tissue sarcoma. The company's Pathwork®Tissue of Origin Test is the only FDA-cleared, molecular diagnostic test to define the origin of a tissue. It helps pathologists and oncologists in the diagnosis of challenging cancer cases such as those that are metastatic or that have a complex clinical history.

AMLPROFILER^{TM} MOLECULAR DIAGNOSTIC ASSAY The AMLprofiler, from Skyline Diagnostic, simplifies the challenging task of recognizing acute myeloid leukemia subtypes and making individualized therapy decisions. It is a diagnostic microarray assay that detects mutations in NPM1 (nucleophosmin) and CEBPA (CCAAT/enhancer binding protein) and determines expression levels of 2 genes with prognostic significance in AML (EVI1 and BAALC). The clinical trial to pursue FDA pre-market approval for the AMLprofiler was unsuccessful due to a lack of participants [3].

Problems of gene expression-based outcome prediction

The successes in the area of gene expression-based identification of cancer biomarkers are the diagnostic kits MammaPrint® and Oncotype DX^{TM}.

MammaPrint predicts metastasis formation in breast cancer from the expression of 70 genes. It is in the market since 2007 and has been tested on more than 6600 patients in the meantime. Oncotype DX – another multigene signature – is based on RT-PCR of 16 cancer-related genes and 5 reference genes. It is commercially available since 2004 for a similar use; predicting the recurrence of tamoxifen-treated, node-negative breast cancer [138].

Accuracy of MammaPrint not reproducible

Despite the success of MammaPrint®, the approach has been criticized. The original paper underlying MammaPrint argued that up to 80% of patients receive unnecessary chemotherapy [185] and showed a prediction accuracy of 83%, but this was later on corrected to 69% [120].

Signature overlap between breast cancer studies not significant

Ein-Dor *et al.* compared the 70 genes signature published by van't Veer and a different signature of 76 genes by Wang *et al.* [192]. They concluded that 'the overlap between these gene sets is almost zero' and 'many equally predictive lists could have been produced' [55].

[3] `http://clinicaltrials.gov/show/NCT01463410`

Recently, Venet *et al.* continued along these lines and showed in a systematic analysis that 60% of the analyzed signatures were not better than random and that more than 90% of random signatures with over 100 genes were significant outcome predictors [188].

Random signature with more than 100 genes as good as any constructed signature

Besides the problem of random signatures, the use of microarrays has been the subject of discussion: Michiels and co-workers showed in seven large cancer prognosis studies that signatures derived by microarray data analysis do not achieve prediction accuracies better than random [120]. Furthermore, DNA microarray technology has basic limitations which include: (1) cross-hybridization between genes of similar sequence [133, 154]; (2) not all genes are reliably detectable, especially those with a low expression level [94]; (3) lack of information about the exact length and sequence of the mRNAs being analyzed; and (4) the inability to detect novel transcripts. Moreover, the comparison of expression levels across different experiments is often difficult and may require complicated normalization methods.

DNA microarray technology exhibit biological limitations

On the other hand, a large study of over 50 groups came to the conclusion that technical noise in microarrays is low [161]. Applied to cancer, Dobbin *et al.* concluded that biological variation between tumors exceeds technical variation [51].

Despite the discussion about the reliability of microarray, many other studies continue using this technology. After the landmark publication of van't Veer *et al.*, other publications reported signatures that also predict the relapse risk of breast tumors [19, 83, 97, 167, 192]. Having such a huge amount of predictive signatures, clinicians need guidance on usage and interpretation. Galina and co-workers published a framework for oncologists to understand and evaluate these biomarkers in clinical application [67].

2.2 NETWORK-BASED CANCER OUTCOME PREDICTION

The last sections introduced microarrays and discussed problems occurring during standard gene expression analysis. Such problems include only a small similarity between predictive signature and noise in the gene expression data. The following section gives an overview over different approaches employing various types of network information to overcome the mentioned problems.

2.2.1 The idea of network-based outcome prediction

Both problems, i.e. random signatures and noisy expression data, can be addressed by data integration [84]. Assuming that the three independent statements 'A is up-regulated', 'A regulates B' and 'B is up-regulated' have each an error probability of 10%, then the joint statements have only an error probability of 0.1%. Hence, the error can be reduced by integrating consistent pieces of knowledge. Ideker and co-workers summarized this general principle and also applied it to the above breast cancer data combining gene expression with protein interaction data [42]. This lead to an improvement of 8% compared with the corrected accuracy of van't Veer [185].

Network-based outcome prediction improves prediction accuracy

Other studies followed these lines and show, that network information efficiently helps to improve outcome prediction and reduce noise in microarray experiments [62]. For this purpose, network information has been integrated with microarray data in various studies in the last decade. The basic principle of all network-based analysis is the usage of functional association networks. Edges in these networks are based on integration of different sources such as high-throughput experiments (like yeast-2-hybrid), physical binding extracted from literature or co-expression networks built from microarray as well as sequencing experiments. Furthermore, gene annotations from Gene Ontology [8] or MeSH [153] can be used to build associations between genes. Although such a network covers only a limited number of real interactomes, integration of microarray and network data has been shown to dramatically improve the outcome prediction of diseases.

Network-based outcome prediction creates consistent signatures

In addition, the integration of such network information reveals high overlap between predictive signatures. Dutkowski and co-workers found a high correlation between the predictive genes of two breast cancer datasets and claimed that predictive signatures would reveal underlying mechanisms of certain diseases [54].

Network-based analysis is nowadays not only used in cancer outcome prediction, but also in different biomedical applications such as drug repositioning, [200], neurodegenerative disorders [71] and for the interpretation of genomic variation data [78].

Based on these observations, we hypothesize that network-based methods are able to substantially improve cancer outcome prediction through data integration. The results of such methods are overlapping signatures, which reveal more insights into the working mechanism of cancer.

2.2.2 State-of-the-art network-based approaches

Integration of gene expression and network information has been proven to identify enhanced gene signatures providing better outcome prediction [62]. For this purpose network information has been integrated with microarray data in various studies in the last decade.

Table 3 gives an overview of different network-based methods. In the last years many studies adopted the idea of network-based outcome analysis, but moved from gene expression data to sequencing data [4, 5, 48, 68, 69, 103, 105, 184]. Integrating protein-protein interaction (e.g. HPRD) seems to be more successful than pathway information (e.g. KEGG and GO).

Table 3: The table gives an overview over different network-based prediction methods: if they use PageRank (PR) or greedy methods for signature identification; if they use protein-protein interaction (PPI) or Pathway information; if they want to obtain predictive subnetworks or single genes and what type of data they use.

			Greedy	**Type of**	
Author	**Net**	**PR**	**Search**	**Marker**	**Data**
Akula *et al.* [5]	PPI	✓	✗	subnetwork	sequencing
Davis *et al.* [48]	Pathway	✓	✗	single gene	sequencing
Hai *et al.* [77]	PPI	✓	✗	single gene	expression
Johannes *et al.* [89]	PPI	✓	✗	single gene	expression
Lee *et al.* [103]	Pathway	✓	✗	single gene	sequencing
Nibbe *et al.* [129]	PPI	✓	✗	single gene	expression
Osmani *et al.* [134]	own	✓	✗	single gene	expression
Tarca *et al.* [174]	Pathway	✓	✗	subnetwork	expression
Winter *et al.* [196]	PPI	✓	✗	single gene	expression
Yang *et al.* [202]	Pathway	✓	✗	single gene	expression
Akavia *et al.* [4]	Pathway	✗	✗	single gene	expression, sequencing
Garcia-Alonso *et al.* [68]	PPI	✗	✗	subnetwork	sequencing
Chen *et al.* [37]	PPI	✗	✗	single gene	expression
Chowdhury *et al.* [40]	PPI	✗	✗	single gene	expression
Chuang *et al.* [42]	PPI	✗	✓	subnetwork	expression
Dao *et al.* [45]	PPI	✗	✓	subnetwork	expression
Dao *et al.* [46]	PPI	✗	✗	subnetwork	expression
Fortney *et al.* [61]	Pathway	✗	✓	subnetwork	expression
Guo *et al.* [73]	PPI	✗	✗	subnetwork	expression
Gilman *et al.* [69]	Pathway	✗	✓	subnetwork	sequencing
Leiserson *et al.* [105]	Pathway	✗	✓	subnetwork	sequencing
Su *et al.* [172]	PPI	✗	✓	subnetwork	expression
Ulitsky *et al.* [181]	PPI	✗	✗	subnetwork	expression
Vandin *et al.* [184]	PPI	✗	✓	subnetwork	sequencing

Recent work by the Ideker lab defines subnetwork activity as the aggregate expression of genes in a given subnetwork [42]. The score of each subnetwork is derived from the mutual information between the subnetwork activity and the outcome variable. The greedy search algorithm is used to identify subnetworks with discriminative activities. An extension of this approach aims to infer activity levels of pathways for disease classification by overlaying gene expression on pathways and searching discriminative gene sets related to the disease phenotype [102].

Another approach developed by Chowdhury and Koyutürk considers binary representation of gene expression data to retrieve subnetwork markers. The algorithm identifies subnetwork markers that are composed of genes deregulated in the same direction (either up- or down-regulated) [40]. This group recently introduced an extension of their previous algorithm by creating networks associated with known genetic marker instead of subnetwork associated with random genes [41]. Both methods were validated on colon cancer and the authors were able to predict colon cancer metastasis with high confidence.

A different approach of Ulitsky *et al.* tries to detect modules that capture genes which not only show highly similar expression behavior but that also exhibit significant correlation with different clinical parameters [181]. The extension of this work aims to explore deregulated pathways enriched with genes representing common regulations for some outcome classes of a disease [85].

Su *et al.* identifies pathways containing many differentially expressed and co-expressed genes from protein-protein interaction networks and greedily combines these paths to obtain subnetwork markers for using disease phenotype classification [172].

In order to classify disease phenotypes, Dao and coworkers develop a density-constrained biclustering approach taking into account pathways being dysregulated in many, but not necessarily all samples [45].

Fortney *et al.* incorporate topological modularity into the expression for subnetwork score by combining genes with high correlation to the outcome variable and with a high modularity [61].

By an edge-based simulated annealing algorithm Guo *et al.* identify responsive subnetworks for prostate cancer [73].

2.2.3 PageRank-based methods

Page and Brin revolutionized the world of search engines by introducing the PageRank algorithm [137], which is a way of measuring the importance of website pages. The basic idea behind PageRank is that a document should be important if it is highly referenced by other documents. Moreover, citations from important documents should have more weight than citations from unimportant documents. As employed by the Google Internet search engine, the PageRank algorithm uses hyperlink information between documents in the world wide web to assign a numerical weighting (termed the PageRank) to each document with the purpose of measuring its relative importance within the set of all web documents. The PageRank algorithm further introduces a constant d representing the probability of randomly jumping to a document instead of following a hyperlink pointing to it. The PageRank for a webpage is calculated as the following:

$$r_j^n = (1-d) + d \sum_{i=1}^{N} \frac{w_{ij} r_i^{n-1}}{deg_i} \qquad 1 \leqslant j \leqslant N \tag{1}$$

A link from page i to page j is regarded as a "vote of confidence" for page j from page i. The algorithm views the web as a directed graph $G(V, E)$, with V being the web pages and E the edges, which represent the links between pages. This information can be stored in an adjacency matrix, $W \in R^{N \times N}$, where $w_{ij} = 1$ if there is a link from page i to page j and $w_{ij} = 0$ otherwise. We define deg_i to be the degree of the i^{th} page. Suppose we have assigned an initial ranking $r^0 \in R^N$. The PageRank algorithm proceeds iteratively, updating the ranking for the j^{th} page from 1 to N according to the formula. Here r_j^n denotes the ranking of page j at the n^{th} iteration and $d \in (0, 1)$ is a fixed parameter. The damping factor d regulates the influence of the network on the results. The higher the value of d, the lower the probability to randomly jumping to a document. Google uses a damping factor of 0.85

Many studies use PageRank for cancer classification

Several recent studies have applied the idea of using the PageRank algorithm for both, classification of cancer tumor data [77, 89, 124, 129, 134, 174, 196, 202] and genome wide association studies [5, 48, 103].

Nibbe *et al.* apply the PageRank algorithm to identify predictive subnetworks by identifying cross-talking proteins between seed genes obtained from other cancer related studies [129]. Hai *et al.* [77] and Osmani *et al.* [134] use the PageRank algorithm to identify characteristic genes specific for a disease by selecting highest ranked genes on a regulatory network which is constructed based on pairwise expression correlations. Another approach aims to explore the effects of gene expressions on the pathways as they calculate an impact score of each pathway for the specific disease conditions [174].

GeneRank, a Gene Ontology based PageRank

Morrison *et al.* recently introduced a modified version of PageRank, which they called GeneRank [124]. The GeneRank algorithm is an adaptation of PageRank which uses networks where genes are connected if they share a Gene Ontology annotation. The authors adapted the algorithm to work with biological data: Gene Ontology annotation and gene expression data. Here, the connections in the network are no longer hyperlinks but genes are connected if they share a Gene Ontology annotation. Thus, GeneRank combines the gene expression data with topological information thus boosting highly connected genes having low gene expression.

The following sections describe approaches using the idea of GeneRank in detail.

NetRank

NetRank, a PPI network-based PageRank

The NetRank algorithm is a network-based prediction approach for identifying marker genes benefiting from gene expression and network data. Instead of Gene Ontology annotations as applied in GeneRank, NetRank employs known transcription factor-target relationships, protein-protein interaction, and gene co-expression to define three different gene-gene networks. For that purpose, NetRank first assigns a score for each gene and then the network is used to spread

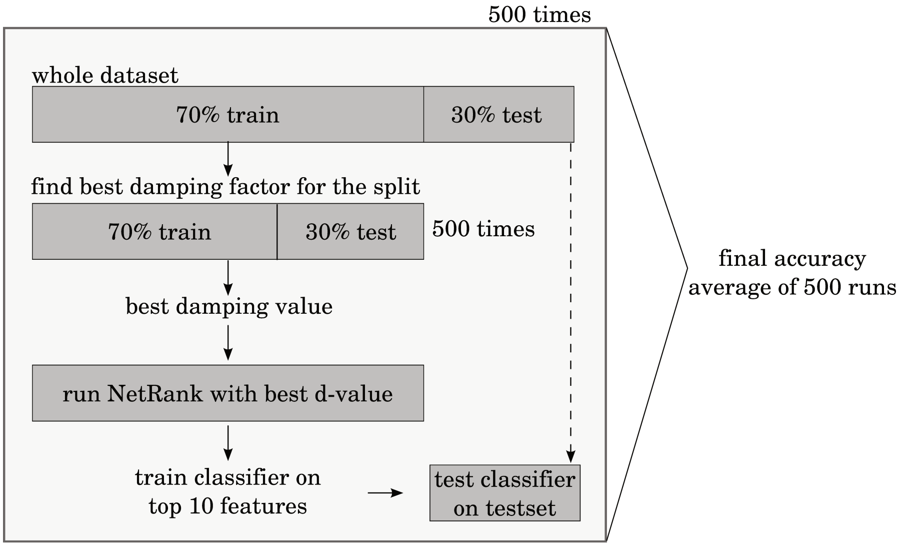


Figure 1: The general workflow of NetRank. First, the full dataset is randomly split into the training and test set. The test set is put aside and later used for validation. Then an additional inner cross-validation loop is added to define the best damping factor for the training set. Afterwards, NetRank is run on the training set using the best damping factor and the most different 10 genes are selected. These genes are used to train a classifier on the training set, which is used to predict the outcome of the remaining test set patients. The whole process is repeated 500 times and a final accuracy is obtained by averaging over the 500 runs.

this correlation to its neighbors and beyond. The genes with the highest NetRank score are then selected as signature genes. The general workflow of NetRank is pictured in Figure 1. The algorithm is applied in a Monte Carlo cross-validation, which is a relatively unbiased evaluation strategy. The novelty of the NetRank algorithm is the dynamic setting of the damping factor d as part of the Monte Carlo cross-validation workflow. Different values of d ranging from 0 to 1 in steps of 0.1 are used and no information of the test set is integrated for the choice of an optimal d value. More details on the NetRank algorithm can be found in chapter 4.

The authors applied the algorithm to predict survival in 30 patients with pancreatic cancer. They found 7 transcription factors with a predictive power 0.68 (AUC) in a leave-one-out cross-validation [196]. Furthermore, they validated the prognostic value of the candidate markers by using immunohistochemistry on an independent set of 412 pancreatic cancer samples.

Reweighted Recursive Feature Elimination (RRFE)

An approach similar to NetRank has been developed by Johannes *et al.*, which identifies the significant genes from an interaction network to predict breast cancer risk of patients [89]. The algorithm – called Reweighted Recursive Feature Elimination (RRFE) – identifies significant genes by stepwise elimination of features [89]. For a stable accuracy computation 10-fold cross-validation is used. Figure 2 shows the general workflow of the algorithm. First, a PPI network is created by using data from the Human Protein Reference Database (HPRD). As a next step, the GeneRank is calculated for each node in

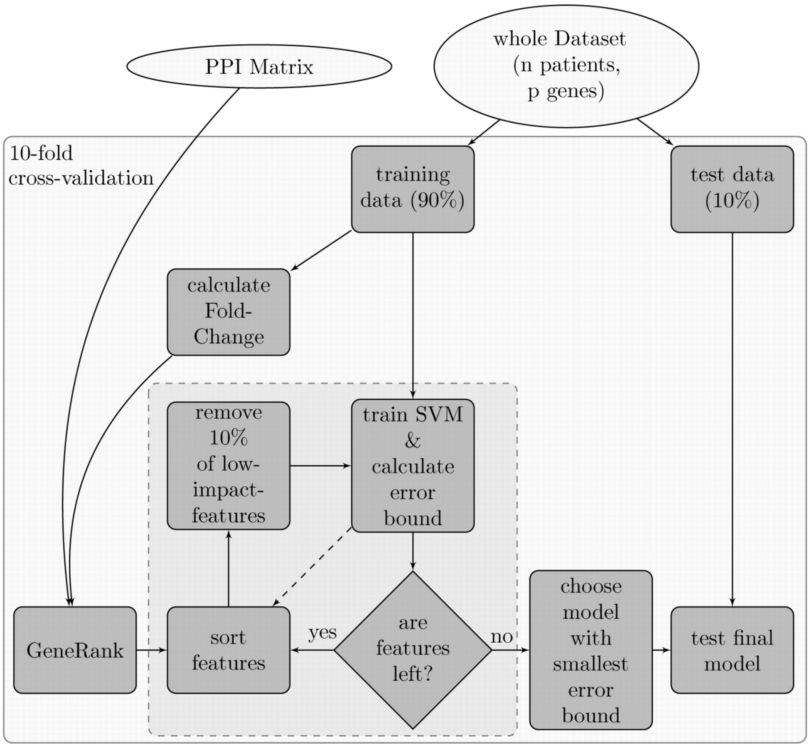


Figure 2: The workflow of the Reweighted Recursive Feature Elimination (RRFE) approach [89]. First, a PPI matrix is created using HPRD data. The gene-expression dataset is split, such that 90% is used as training set. The remaining 10% is later used for evaluation. Then, foldchange is used as input for GeneRank to rank the features. In the next step, a SVM is trained on all features and the lowest 10% of the features is discarded. This procedure of eliminating genes is repeated until the set of surviving features is empty. Afterwards, the SVM with the optimal number of features is selected and tested on the test set. The whole process is repeated 10 times.

the network and a SVM with linear kernel is trained on all features/-genes. For GeneRank, a damping factor *d* of 0.5 is used, as suggested in the original GeneRank publication. After the training, a RFE-score is calculated for each node based on the GeneRank and the importance of that feature/gene for the SVM. This RFE-score takes into account both the impact of a particular feature on the weight vector of the hyperplane and the connectivity of that feature in the underlying biological network. After ordering the list of features, 10% of features with the smallest RFE-score are removed. This procedure of calculating SVM, re-scoring and feature elimination is repeated until the set of surviving features is empty. After finishing the elimination process, the best feature combination has to be selected based on the leave-one-out error.

Johannes *et al.* tested their approach on real data from breast cancer patients. They predicted relapse events based on gene expression with an AUC of 0.67, which was superior to all methods they compared with.

They furthermore found a high overlap of resulting features in the 10-fold cross-validation process.

netSAM

Yang and co-workers used the GeneRank algorithm to prioritize network hubs [202]. Their framework can be found in Figure 3. The authors start with extracting differentially expressed genes with a foldchange larger than 2 and a $p-value$ less than 0.01. Out of these differentially expressed genes they infer two types of networks, a 'case'

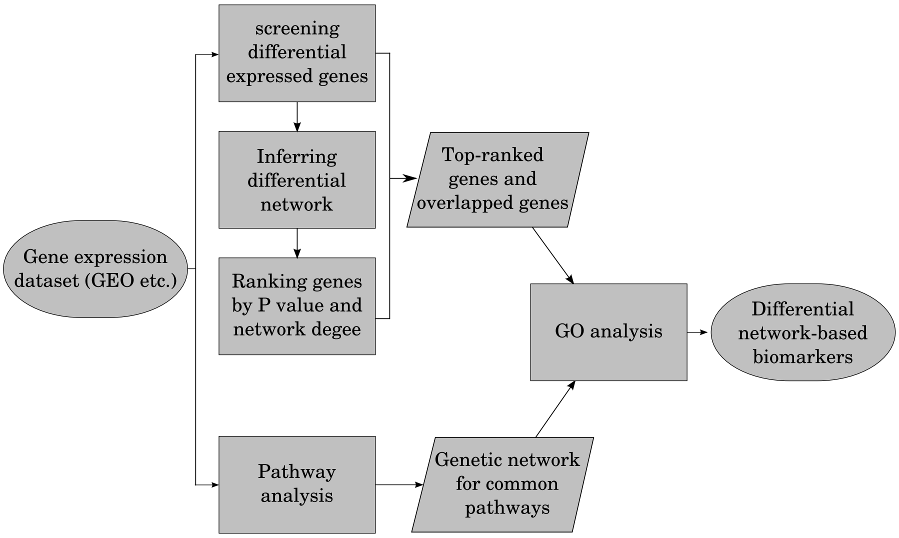

screening differential expressed genes
Inferring differential network
Ranking genes by P value and network degee
Gene expression dataset (GEO etc.)
Top-ranked genes and overlapped genes
GO analysis
Differential network-based biomarkers
Pathway analysis
Genetic network for common pathways

Figure 3: The workflow of netSAM. First, differentially expressed genes are extracted from GEO. A differential network is constructed through comparison of the difference of the interactions and the subtraction of a 'case' from a 'control' network. Then, differential expressed network hubs are prioritized through GeneRank. Besides, associations between differential genetic interactions and known pathways are investigated. GO Analysis of Candidate Genes from GeneRank and pathway analyses leads to the final network-based biomarker. The figure is adapted from [202].

and a 'control' network, via boosting regression. Afterwards, they identify differential network hubs via subtraction of the 'case' from the 'control' network. These hubs get prioritized via the GeneRank algorithm. As a final step, a Gene Ontology analysis is conducted. In their analysis the damping factor d is set to 0.5, as suggested in the original GeneRank publication.

The algorithm was validated on synthetic datasets and achieved an AUC of 0.80. After applying netSAM to two real breast cancer datasets from Wang *et al.* [192] and van de Vijver *et al.* [191], they showed that genes selected by netSAM are higher enriched in cancer related terms than genes found by standard methods (t-test and lasso).

2.2.4 Related Work

In a recent work Cun and Fröhlich applied fourteen published gene selection methods of which eight are using network information on six public available breast cancer datasets [44]. They compared the results with respect to prediction accuracy, signature stability and biological interpretability. Regarding the prediction accuracy they did not find a general advantage of network-based methods over standard analyses methods (e.g. PAM [178] or SAM [180]). Nevertheless, some of the network-based methods revealed significantly higher gene selection stability. Biomarker signatures developed by network-based approaches revealed a high enrichment of disease related genes, KEGG pathways and known drug targets.

Altogether, they showed that some network-based approaches have significantly higher signature stability, but do not perform better than classical methods. The question arises if these results supports the conclusions made in this thesis (see subsection 4.3.6).

2.3 UNIVERSAL CANCER SIGNATURE

Elevated levels of single gene markers as shown in Table 1 are often not only found in a single type of cancer. Perkins and co-workers showed, that e.g. the carbohydrate antigen CA 19-9 does not only show elevated levels in pancreas cancer but also in other cancer types as colon, esophageal and hepatic cancers [144]. They also found the oncofetal protein alpha-fetoprotein differentially expressed in hepatocellular carcinoma as well as in gastric, billiary and pancreatic cancers. This points towards that several cancers share the same mechanism of survival, tumor growth and invasion.

In the last decade the different stages of cancer development have been a focus in biomedical research. Hanahan and Weinberg have summarized the biological capabilities acquired during the multistage development of human cancers in the Hallmarks of cancer [79]. These Hallmarks of cancer represents distinctive biological processes that are responsible for e.g. tumor growth, invasion and metastatic dissemination. Imagine a signature that is able to predict general cancer outcome is existing, this signature should cover most of the Hallmarks.

The usage of networks for outcome prediction reveals high overlap between predictive signatures. Shi *et al.* developed a network-based signature for colorectal cancer recurrence by integrating several colorectal cancer signatures and interaction networks. They highlighted the dysregulated biological processes in colorectal cancer recurrence [162]. In addition, Dutkowski and co-workers found a high correlation between the predictive genes of two breast cancer datasets and claimed that predictive signatures would reveal underlying mechanisms of certain diseases [54].

In contrast, a study compared two studies focusing on breast cancer progression and compared their signatures [55]. They found, that the overlap between these gene sets is almost zero and that many equally predictive lists could have been produced, indicating that no general cancer signature is existing.

Although these breast cancer signatures are not overlapping, common effected biological processes are captured, proving the analogous predictive power of the signatures [58].

2.3.1 Network-based analysis revealed overlap of signatures between different types of cancer

Integration of gene expression and network information revealed more efficient gene signatures providing better outcome prediction [40, 42, 45, 61, 129, 172, 181, 196].

A network-based outcome prediction approach – NetRank – is in this study applied on several types of cancer gene expression data. NetRank is a PageRank derivative, which assigns a rank score to a gene, being a combination of its score and the scores of genes linked to it. Basically, if a gene is linked to other highly ranked genes, it will also get a high rank due to the boosting effect of neighboring genes. Hence, the algorithm provides an integration of the topological information (i.e. connectivity and random walk) and microarray data (i.e. node score) to explore crucial signature genes for the outcome prediction. While systematically evaluating the network-based outcome prediction on 25 literature derived cancer datasets and one in-house dataset, we show that fully automated integration of network information and gene expression data leads to more accurate outcome predictions compared to the signatures of previous studies and classical methods such as *t*-test and foldchange (see chapter 4). The predicted signatures for a specific cancer type are more interpretable and reproducible. Furthermore, signatures predicted by NetRank show a high similarity even for different types of cancers. Hence, the biological mission and prognostic ability of such similar signatures will be investigated in more detail (chapter 5).

Cancer biomarkers can be classified into three main classes: prognosis, predictive and pharmacodynamics [156]. Prognosis biomarkers aim to distinguish patients having good or poor prognosis. Predictive biomarkers determine which drugs would be effective for treatment of a specific tumor. Pharmacodynamics biomarkers define the optimal dosage of a drug for a patient. The signatures predicted in this study are more adequate for prognosis, since they help to discriminate tumor behavior (e.g. good, poor, aggressive) and to make decision for treatment.

2.3.2 Is there already a Universal Cancer Signature?

An anti-profile signature was developed in another study, where the genes show hyper-variability across tumor samples and are selected as biomarker to discriminate healthy and cancer patients [27]. This anti-profile identifies colon cancer patients using peripheral blood samples with an AUC of 0.89. The universal cancer anti-profile can also distinguish healthy and cancer patients (with different tissues) with more than AUC of 0.92.

There are two aspects that have to be considered: What are the key roles of the single genes in the predictive signatures in the cancer progression? Does a Universal Cancer Signature exists that could possibly predict prognosis or progression of different cancer types?

3 METHODS

The following section describes methods used for the evaluation of network-based outcome prediction approaches. The general idea as well as the algorithmic details of NetRank is explained. The classical approaches are introduced as well as the formula to calculate accuracy and signature similarity. Finally, the detailed workflow for the benchmark dataset creation is explained.

3.1 NETRANK ALGORITHM

GENERAL IDEA OF NETRANK A general overview of the idea of NetRank is depicted in Figure 4. First a value representing either the expression change between the different outcome groups or the correlation of the expression with the outcome variable is overlaid with the network. Often these values do not represent the biological truth. The central and highly connected node is only represented by a relatively small value of 0.4 compared to other less connected nodes. This node represents an important hub that either physically interacts with or regulates many different nodes. The small value compared to other nodes might be due to biological or technical noise happening during the experiment. After applying NetRank the values are representing the high biological relevance and highly connected genes tend to have high scores, when neighboring nodes are highly expressed.

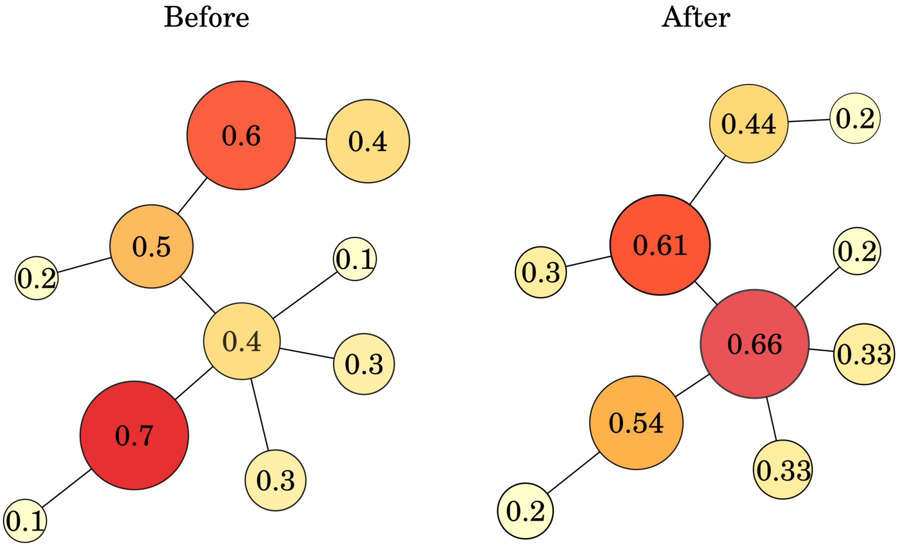


Figure 4: Before NetRank, the values are not representing the biological truth as the central node is represented by a relatively small value compared to other less connected nodes. This might be due to biological or technical noise happening during the experiment. After applying NetRank the values are representing the biological relevance and highly connected genes tend to have high scores.

ALGORITHM For ranking of genes, NetRank combines the expression level of a gene with the outcome variable of the patient by using a network of known gene-gene relationships. The ranking might be computed by eigenvalue decomposition or iteratively. Here, we follow the notation and implementation of Morrison *et al.* [124], that defines the ranking of gene j at the n^{th} iteration as follows:

$$r_j^n = (1-d)c_j + d\sum_{i=1}^{N} \frac{w_{ij} r_i^{n-1}}{deg_i} \qquad 1 \leqslant j \leqslant N \tag{2}$$

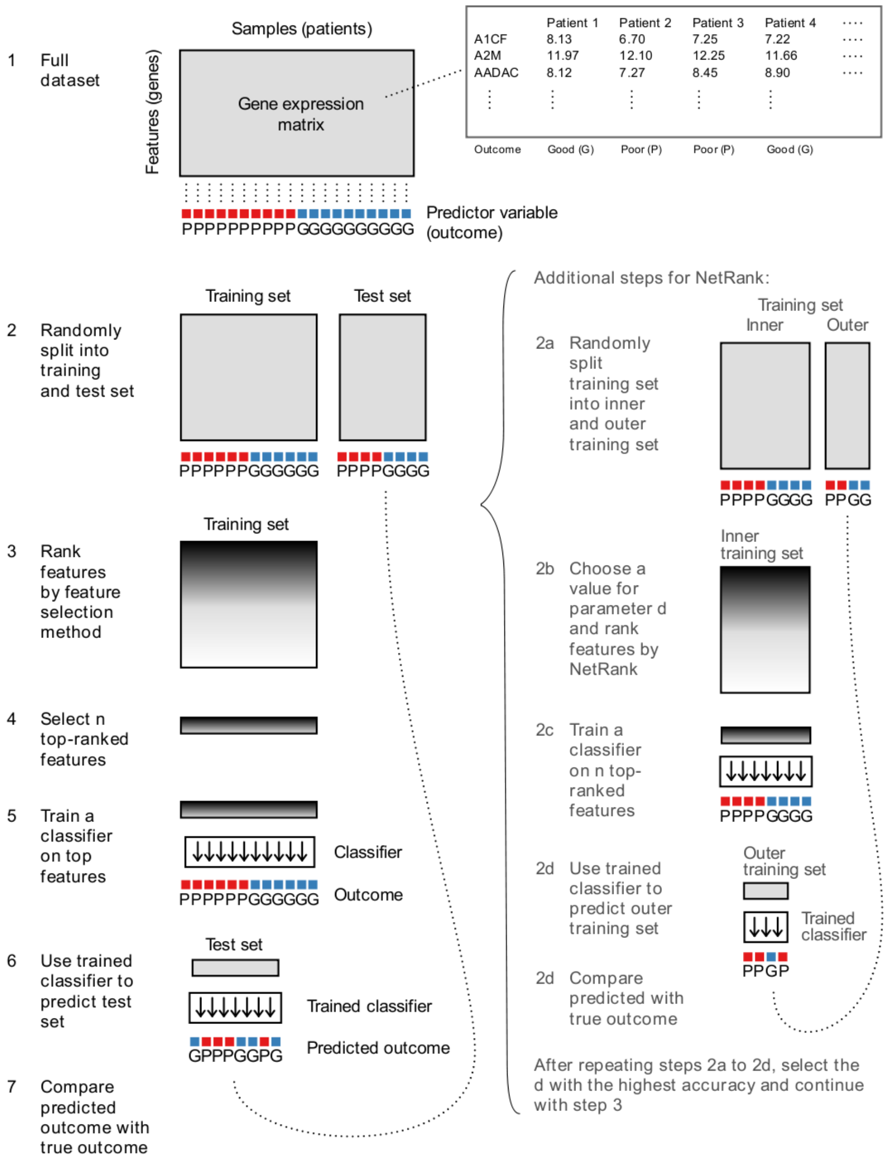
where $W \in \mathbb{R}^{N \times N}$ is a symmetric adjacency matrix representing a gene network, so $w_{ij} = w_{ji} = 1$ if genes i and j are connected, and otherwise $w_{ij} = w_{ji} = 0$. $\vec{c}$ is a vector of coefficients representing the gene expression values with the patients outcome and it is calculated by foldchange, Student's t-statistic or correlation (Pearson, Spearman or Kendall).

Damping factor regulates the influence of the network

The damping factor $d \in (0, 1)$ is a parameter describing the influence of the network on the rank of a gene. A damping factor of $d = 0$ corresponds to no influence of the network and full influence of the gene expression data, whereas setting $d = 1$ corresponds to full influence of the network and no influence of the gene expression data on the rank value of that gene. The value $d = 0.85$ appears to be used by Google [137]. The rank of a gene depends on the rank of all genes that connect to it. Scaling by $1/deg_i$ in the summation ensures that each gene has equal influence in the voting procedure. Each gene gets a rank of $1-d$ automatically and also gets d times the votes given by other genes.

The damping factor is set dynamically for each dataset

The parameter d is here set as part of the Monte Carlo cross-validation workflow. For NetRank, we added an additional inner cross-validation loop. In this inner cross-validation, a part of the training set samples were excluded, and different values of d ranging from 0 to 1 in steps of 0.1 were used to run NetRank on the remaining training set samples. Accuracies of the top-ranked genes were later tested on the samples previously set aside. As a result of this inner cross-validation, one d value was chosen and used once for deriving a signature using the whole training set, and then evaluating its accuracy on the test set. It is important to note that no information of the test set is used for selecting d, so the choice of a value for d in the inner cross-validation does not rely on any prediction accuracy in data of the test set. Furthermore, the optimal d value for each cancer dataset is dynamically identified by applying the inner cross-validation step. The dynamic setting of d is a novel improvement of the NetRank algorithm.


1 Full dataset
Samples (patients)
Features (genes)
Gene expression matrix
Predictor variable (outcome)
PPPPPPPPPPGGGGGGGGGG
2 Randomly split into training and test set
Training set
Test set
PPPPPPGGGGGG
PPPPGGGG
3 Rank features by feature selection method
Training set
4 Select n top-ranked features
5 Train a classifier on top features
Classifier
Outcome
PPPPPPGGGGGG
6 Use trained classifier to predict test set
Test set
Trained classifier
Predicted outcome
GPPPGGPG
7 Compare predicted outcome with true outcome
Additional steps for NetRank:
2a Randomly split training set into inner and outer training set
Training set
Inner
Outer
PPPPGGGG
PPGG
2b Choose a value for parameter d and rank features by NetRank
Inner training set
2c Train a classifier on n top-ranked features
PPPPGGGG
2d Use trained classifier to predict outer training set
Outer training set
Trained classifier
PPGP
2d Compare predicted with true outcome
After repeating steps 2a to 2d, select the d with the highest accuracy and continue with step 3

	Patient 1	Patient 2	Patient 3	Patient 4	....
A1CF	8.13	6.70	7.25	7.22	....
A2M	11.97	12.10	12.25	11.66	....
AADAC	8.12	7.27	8.45	8.90	....
⋮	⋮	⋮	⋮	⋮	
Outcome	Good (G)	Poor (P)	Poor (P)	Good (G)	

Figure 5: The detailed workflow of NetRank. First the full dataset is randomly divided into the training and test set, and the test set is put aside and later used for validation (2). While applying NetRank on the initial training set, the genes are ranked by how different they are between patients with poor and good prognosis (3) and the most different genes are selected (4). These genes are used to train a classifier on the training set (5), which is used to predict the outcome of the remaining test set patients (6). The predicted outcome is compared to the real outcome and the number of correctly classified patients is noted (7). Steps 2-7 are repeated up to 500 times and a final accuracy is obtained by averaging over the 500 runs. The figure is adapted from [196].

DETAILED DESCRIPTION OF NETRANK A detailed overview of the pipeline can be found in Figure 5. First the full dataset – a gene expression matrix, with genes as rows and patients as columns (1) – is randomly divided into a training and test set, where the test set is exclusively used for validation (2). For a dynamical setting of the damping factor, the training set is again randomly splitted (2a - d) and the best damping factor is found via a cross-validation procedure for the initial split. While applying the best damping factor on the initial training set, the genes are ranked by how different they are expressed between patients with poor and good prognosis (3) and the genes with the highest variance are selected (4). These genes are used to train a classifier on the training set (5), which is used to predict the outcome of the remaining test set patients (6). The predicted outcome is compared to the real outcome and the number of correctly classified patients is noted (7). Steps 2-7 are repeated up to 500 times and a final accuracy is obtained by averaging over the 500 runs.

3.1.1 Monte Carlo cross-validation workflow

To get a robust estimate on the classification error rate, we adopted the multiple random validation strategy described by Michiels *et al.* [120]. Given a fixed signature size n and a feature selection method, the following steps were repeated 500 times:

1. The patients of the dataset are randomly split into training and test sets. The splitting is balanced such that the numbers of poor and good samples in the test set are either equal or differ by at most one. This is to ensure that there is no over-representation of one of the groups in the training set.

2. Using the training set data only, features are ranked according to a feature selection method.

3. The top-ranked n features are selected and these features become the signature.

4. The signature from the training set is used to train a classifier on the sample outcome, using the training set expression values of the signature genes as input.

5. The classifier is used to predict the outcome of the unseen test set patients.

6. The predicted outcome is compared with the true outcome. The fraction of correctly predicted patients defines the accuracy.

For the dynamical setting of the damping factor, additional steps are taken between step 2 and step 3. The overall classification accuracy is the average of all repeated workflow accuracies. In order to ensure maximally comparable results, the random splits into training and test sets were carried out once and the sets were recorded. Thus, the exact same training and test sets were used for each method applied.

3.1.2 Poor and good prognosis

Patients in the datasets were divided into two groups based on their clinical outcome. The patients were separated into groups regarding the median survival time, if no classification was already given in the publications. In that sense poor prognosis reflects a survival or response time less than a certain time threshold (e.g., 3 or 5 years), whereas patients who survive longer than the median survival time are classified into the good prognosis group.

3.1.3 Feature selection method

For a prognostic signature, genes with high variance between the classified groups/samples were selected. To evaluate the strength of difference following options were compared: (i) foldchange, as defined by the ratio of mean gene expression in one group over the other group, (ii) the Student's t-statistic, (iii) the Pearson as well as the Spearman rank correlation coefficient of gene expression with survival time of the patient.

3.1.4 Classification procedures

A prognostic model with 10 signature genes was developed by employing Support Vector Machine (SVM) based learning [158]. Support vector machines are powerful machine learning algorithms for classification problems [25, 158, 186]. A SVM was used to classify the tumors samples into poor or good prognosis groups based on the expression levels of selected genes. Here, the LIBSVM implementation is employed as provided in the R package e1071 (version 1.5-18, obtained July 2010). The expression of each gene was used as an independent feature to train the classifier and no kind of aggregation was applied. All feature selection and machine learning steps were subjected to Monte Carlo cross-validation, which is a recommended and relatively unbiased evaluation strategy [26, 120].

3.2 PREPROCESSING AND FILTERING OF THE MICROARRAY DATA

For some datasets in the benchmark dataset raw microarray data were available. Therefore, Affymetrix raw probe level intensity files were background - corrected, normalized, and summarized using RMA.

The goal was to find a small number of reliable markers, so we do not want any candidate markers fall into one the categories of (i) low variance between all samples and therefore not discriminative, (ii) de facto not expressed, or (iii) expressed at very low levels and thus not reliably measured in our microarray data. This is achieved by the filtering steps described in the following. Note that for NetRank, filtering out meant actually not removing, but setting initial values to zero in order to prevent loss of edges from the network due to node removal.

First, to remove noise from genes with low expression, probe sets with a mean expression below 6 on the scale were filtered out from the dataset. Second, genes whose expression shows little variation between patients are not informative as they cannot discriminate between patient groups. Therefore, probe sets with a standard deviation below 0.5 on the scale were filtered out. One gene or EST is normally represented by one to several probe sets. Thus, we decided to keep for each gene only the probe set with the highest mean expression over all patients. We generally found a high correlation between probe sets reporting for the same gene.

3.3 ACCURACY

Accuracy is the percentage of correctly classified samples and it is defined as the proportion of true predictions to all samples. If no prediction accuracy is stated in the original study, we calculated it by using the following equation:

accuracy = sensitivity * prevalence +
specificity * (1 - prevalence)

where the prevalence is determined as

prevalence = no. of ill persons in that dataset divided
by dataset size.

3.4 SIGNATURE SIMILARITY

Signatures A and B are compared by Jaccard index, i.e. the size of the intersection of A and B divided by the the size of their union.

$$JI = |\frac{A \bigcap B}{A \bigcup B}|. \tag{3}$$

The $p - value$ of each Jaccard index is calculated by the Fisher's exact test (fisher.test in stats R package). This test exploits the probability of an overlap between signatures by chance. In order to provide a better visualization of statistically significant Jaccard indices, the p-values are converted into $-log(p - value)$ values in the heatmap plots.

3.5 CLASSICAL APPROACHES

Classical approaches for gene expression analysis aim in finding differences between outcome groups without using any prior knowledge. Many such techniques exist for the identification of genes that are differentially expressed between conditions. The used method can greatly influence the set of genes identified. Despite the huge variety of possible methods, there exist two most often used techniques, the *t*-test and foldchange, presumably because of their simplicity and interpretability. It has been shown that foldchange results in more reproducible gene list than the *t*-test [160, 161]. Witten and co-workerss compared these two strategies on real and simulated data. They conclude, that 'there is no correct answer as to whether fold-change or the modified t-statistic should be used.' It depends on whether the interest lies in an absolute change of gene expression or in the change in gene expression relative to the underlying noise in the gene [197].

Both methods were applied as implemented in the genefilter R package. In order to select significant genes we did not apply either a cutoff threshold or a multiple correction test. After calculation of these measures, low expression and low variance genes from the potential signature set were filtered out. All signatures and accuracies for the classical methods were obtained by the same Monte Carlo cross-validation workflow as applied in the NetRank.

3.5.1 foldchange

The foldchange is a measure that describes the quantity of changes between different values. Genes that differ by more than an arbitrary threshold are then considered to be differentially expressed. In the classical microarray analysis these values are normally represented as the mean of the different outcome groups [180]. As the expression data is $\log_2$ transformed during preprocessing, the foldchange is defined as:

$$fc = \frac{\overline{x_t}}{\overline{x_c}} \tag{4}$$

Where $\overline{x_t}$ represents the mean of the gene expression of patients in the treatment group and $\overline{x_c}$ the mean gene expression of patients in the control group. Despite its simplicity, the foldchange method exhibits several disadvantages. First, the variability of variances is ignored as the foldchange only considers mean values. This results in genes with large variances passing the cutoff just because of noise. Second, to call a gene differentially expressed a certain threshold has to be chosen, without having an indicated level of confidence. Any chosen threshold therefore remains somewhat arbitrary, and biologically relevant changes in expression might be happening below this cutoff, leading to false negative results. To overcome these problems, the *t*-test can be applied.

3.5.2 *t*-test

The *t*-test is a statistical hypothesis test, that follows the Student's t distribution if the null hypothesis is supported. It assesses whether the means of two groups are statistically different. It does so by judging the different means relative to the spread of variability.

$$\begin{aligned} t &= \frac{\text{signal}}{\text{noise}} \qquad (5)\\ &= \frac{\text{difference between group means}}{\text{variability of groups}} \qquad (6)\\ &= \frac{\overline{x_t} - \overline{x_c}}{\sqrt{\frac{var_t}{n_t} + \frac{var_c}{n_c}}} \qquad (7) \end{aligned}$$

This equation does only work for parametric data. For data with unequal variance the so called Welch *t*-test has to be applied.

3.6 RANDOM SIGNATURES

We created 1000 different random signatures with the size of 10 genes for each dataset. These signatures were created by randomly selecting 10 genes out of all genes in each dataset. We then tested the outcome prediction accuracy of random signatures via the same Monte Carlo cross-validation as applied on NetRank on each dataset.

3.7 NETWORKS

For the exhaustive testing of NetRank four types of networks are employed. For the main analysis HPRD [146] and Transfac [119] are used. Later, the analysis is extended by using STRING [63] and hPrint [56].

Genes and nodes in networks are represented by Entrez gene identifiers, therefore probes with unknown Entrez identifiers were discarded.

TRANSFAC TRANSFAC is a manually curated database. It provides data on eukaryotic transcription factors, their genomic binding sites and regulated genes. The content is centered on the interaction between transcription factors and their DNA binding sites. It further contains information about the structural and functional features of the transcription factors. Transfac consist of 2.400 genes and 5.700 interactions.

HPRD HPRD, the Human Protein Reference Database, integrates information from domain architecture, post - translational modifications, interaction networks and disease association for each known human protein. The information is as well manually curated from literature and contains roughly 12.300 genes and 41.300 interactions.

hPrint

HPRINT contains predicted physical and functional interactions. It uses a sophisticated combination of random forest and Bayesian learning approaches in order to integrate various data sources like text mining, genetic relationships, evolutionary information as well as domain profiles. By using this ensemble of evidences hPrint predicts protein-protein interactions and integrates those predictions with known information. To keep only high confident interactions and to remove the majority of false positives, the network was filtered with a confidence score of 0.7 leading to 12.800 nodes and 236.000 edges.

STRING

STRING is a database of known and predicted protein interactions. Interactions included are either direct (physical) or indirect (functional) associations and are derived from different sources: high-throughput and/or co-expression experiments, textmining as well as downloaded from existing databases. To have high confidence interactions, STRING was filtered and only interactions with a confidence score greater than 800 were used, which lead to 11.700 nodes and 342.000 interactions. The confidence score is based on an internal validation. Higher scores lead to sparse networks, whereas lower scores contain many false positives interactions.

3.8 DIRECT NEIGHBOR METHOD

The direct neighbor method averages a gene's correlation coefficient over its direct neighbors. Similar to NetRank, the direct neighbor method starts with an undirected gene network W (represented by a symmetric adjacency matrix with $w_{ij} = w_{ji} = 1$ if genes i and j are connected, and $w_{ij} = w_{ji} = 0$ otherwise) and a gene vector $\vec{c}$ with absolute Pearson correlation coefficients of gene expression values with the patient survival time. The rank r_i of a gene i is then determined by

$$\vec{r} = (W\vec{c}/deg\vec{g}_w) \qquad (8)$$

where $deg\vec{g}_w$ denotes the degree of the matrix W.

3.9 DATASET EXTRACTION

Through a comprehensive Medline search previously published mRNA expression datasets covering different cancer classification scenarios were obtained.

More than five hundred publications matching the following query were extracted on the 1.April.2011:

```
(cancer[tiab] OR neoplas*[tiab] OR tumor[tiab]) AND
humans[MESH] AND
"gene expression"[tiab] AND
2002:2011[pdat] AND
(marker[tiab] OR biomarker[tiab] OR signature[tiab]) AND
(predict*[tiab] OR diagnos*[tiab]) AND
(survival[tiab] OR outcome[tiab] OR progression[tiab] OR
response[tiab] OR metastasis[tiab] OR behavior[tiab])
NOT (networks[tiab] OR pathway[tiab] OR review[pt] OR
"Tissue Array Analysis"[MeSH Terms] OR
"Protein Array Analysis"[MeSH Terms])
```

To only retain high quality papers, we selected publications in journals with impact factor > 5. The remaining 239 articles are screened manually. For the majority of publications no data was available. The most widely used chips were HGU133plus2 and HGU133A. For the sake of comparability, only papers using these chips are retained. For a reliable accuracy calculation, a cross-validation step is included in the evaluation. Therefore, only datasets with more than 20 patients are included in the benchmark dataset. This stringent filtering resulted in 25 suitable datasets.

4 PERFORMANCE OF NETRANK

In the following chapter NetRank is evaluated on a benchmark dataset. First, the benchmark dataset is described in great detail. Second, certain parameters that might influence the outcome prediction accuracy are investigated. Then, NetRank is applied on the benchmark dataset and its performance is compared to the original studies and classical methods. Finally, some general findings are discussed as well as the complexity of NetRank evaluated.

4.1 BENCHMARK DATASET FOR EVALUATION

In the next section the benchmark dataset is analyzed regarding the cohort and signature size, the different cancer types covered and the different methods used. Finally we test, whether the datasets are representative for all microarray datasets.

Previously published mRNA expression datasets covering different cancer classification scenarios were obtained by a comprehensive Medline search. More than five hundred papers matching the query were extracted. After a stringent filtering process (see section 3.9), the benchmark dataset consists out of 25 suitable datasets summarized in Table 4. For comparison, also an in-house dataset was added [196].

Benchmark dataset reflects estimated cancer cases

BENCHMARK DATASET IS REPRESENTATIVE In 2013 there are 1.6 Mio. estimated new cancer cases in the United States [165]. Siegel and coworkers separated the cancer types into 16 different groups. Our benchmark dataset covers the biggest groups including cancer occurring in the genital, respiratory and digestive system. Figure 6 shows the frequency of cancer types in our benchmark dataset and the number of estimated new cancer cases in 2013. The remaining 6 cancer groups – regarding Siegel *et al.* - Bones & joints, Soft Tissue, Eye & Orbit, Endocrine system and Myeloma as well as unspecified cancers – cover only 8% of the overall new estimated cancer cases. Therefore, our benchmark dataset is representative as it covers the majority of all cancer types.

Signature and cohort size varies in benchmark dataset

Figure 7a shows the distribution of the signature and cohort size in the benchmark dataset. In cancer outcome prediction a small signature size is desirable as it can be easily tested in clinical settings via RT-PCR or immunohistochemistry. The optimal signature size has not been investigated yet, and it remains unclear if there is any best signature size. The size of the optimal signature might depend on the

Table 4: The resulting 26 datasets are covering 13 cancers, four outcome tasks, and different cohort sizes.

Author	Year	Cancer	Var.	#Pat.	Sig.size	Acc†
Jones *et al.* [90]	2005	Renal cell	S	69∗	30	100
Spira *et al.* [168]	2007	Lung	D	129	80	82
Lee *et al.* [104] ⋄	2010	Bladder	P	165	1	79
Steidl *et al.* [170]	2010	Lymphoma	T	130	86	77
Bhojwani *et al.* [20]	2008	Leukemia	T	82	24	74
			P	59	41	73
Bogunovic *et al.* [22]	2009	Melanoma	P	38	50	74
Raponi *et al.* [149] ⋄	2006	Lung	P	129	50	64
van de Vijver *et al.* [191] ⋄	2002	Breast	P	295	70	62
Wang *et al.* [192]	2005	Breast	P	276∗	76	62
Korkola *et al.* [96]	2009	Germ cell	P	82∗	140	60
O'Donnell *et al.* [132]	2005	Oral cavity	P	27∗	116	100⋆
Friedman *et al.* [65]⋄	2009	Leukemia	P	68	180	65⋆
Landemaine *et al.* [100]⋄	2008	Breast	P	23	6	-
Nanni *et al.* [127]⋄	2006	Prostate	D	30	20	-
			T	20	30	-
Iqbal *et al.* [86]	2010	Lymphoma	D	80∗	52	-
Fernandez *et al.* [59]	2010	Lymphoma	S	22∗	13	-
Frank *et al.* [64]	2006	Leukemia	T	41∗	128	-
Smith *et al.* [166]	2010	Colon	P	55	34	-
Lenz *et al.* [106]	2008	Lymphoma	T	414	357	-
Mok *et al.* [122]	2009	Ovarian	P	53	11	-
Dressman *et al.* [52]	2006	Breast	S	37	22	-
Murat *et al.* [125]	2008	Glioblastoma	T	70∗	20	-
Zhu *et al.* [206]	2010	Lung	P	133	15	-
Winter *et al.* [196]	2012	Pancreas	P	30	7	-

† Accuracy of the original studies.
∗ A subset of the original dataset was used for NetRank validation.
⋄ These dataset did not provide raw data.
⋆ There is no separation between training and test patients.
clinical variables: D – Diagnosis, P – Prognosis/Progression, S – Subtyping, T – Treatment Response/Outcome.

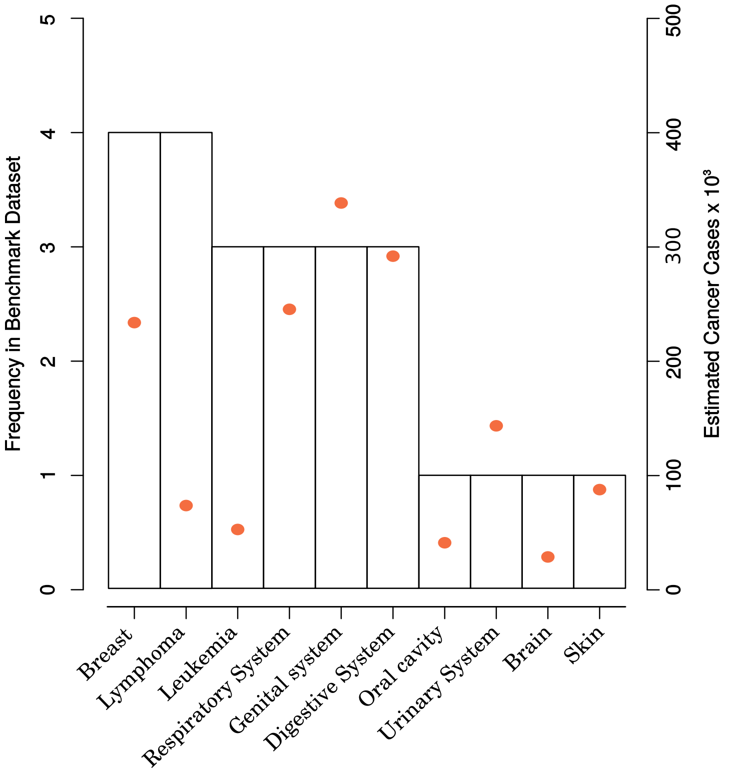
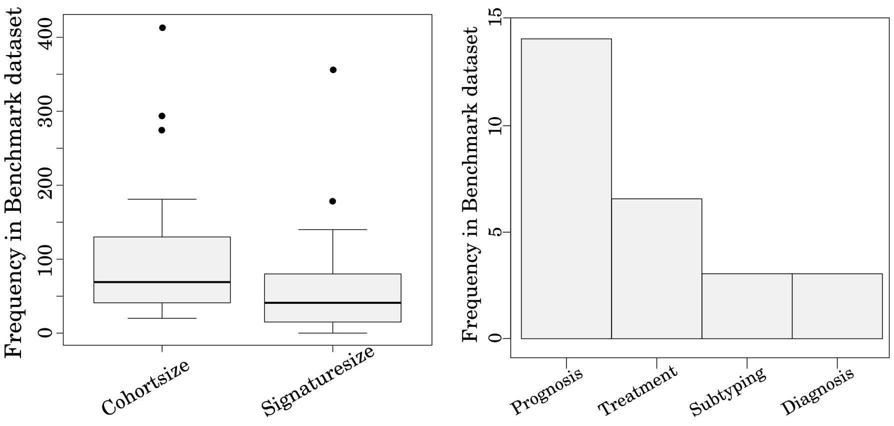


Figure 6: Frequency of cancer types in our benchmark dataset and the number of estimated new cancer cases in 2013. Siegel and co-workers separated the cancer types into 16 different groups [165]. Our benchmark dataset covers the biggest groups including cancer occurring in the genital, respiratory and digestive system.



(a) Cohort and signature size

(b) Outcome variable

Figure 7: Patient and signature size (a) as well as frequency of outcome variables (b) in the benchmark dataset. (a) The cohort and signature size in the benchmark dataset is in average 103 and 65, respectively. (b) The majority of studies in the benchmark dataset are predicting the prognosis of a patient.

biological question underlying a certain study as well as the desired prediction accuracy. The datasets in the benchmark dataset discovered signatures containing from one to 357 genes, having 65 genes in average.

The benchmarks datasets contain in average 103 patients, reaching from 20 up to 414. There have been several studies correlating the size of the cohort with the predictive power of the resulting signatures [31, 55]. They show that studies analyzing more than 50 patients result in better predictions. Having less than 50 patients a systematic improvement is possible using bootstrapping [31]. The question arises, whether these correlations can be also found in our benchmark dataset and if the cohort size also correlates with the prediction accuracy.

Four different outcome variables covered

Four different outcome prediction categories exist in cancer outcome prediction – Prognosis/Progression, Treatment, Subtyping and Diagnosis. Figure 7b shows the distribution of the outcome variables in the benchmark dataset. The datasets differ in the outcome variable. Disease progression – either death or relapse – is the outcome in 14 studies and therapeutic response in 6 studies. Diagnosis and disease subtyping is the outcome in 3 studies, respectively. Having different outcome tasks, the question arises, if there is a correlation between outcome task and prediction accuracy.

Benchmark dataset covers variety of statistical methods

In the benchmark studies different computational and experimental methods were applied for the development of the biomarker signatures.

The studies used one or more of the following methods: Computational methodologies including Student's t-statistic (13 out of 26 cases) and its variances (Max-T, fishers exact test, random variance *t*-test) as well as cross-validation (mainly leave-one-out) and Cox proportional hazard models (15 out of 26). Less frequently applied methods are machine learning (ML), mainly Support Vector Machines with linear kernel, SAM [180] and prediction analysis of microarray (PAM) [178]. For analysis hardly used methods are chi-square test and principal component analysis (PCA) as well as pathway analysis. Hardly employed methods are regression like logistic or Bayesian regression along with partial least square and maximizing R square. Signatures developed by these studies have been validated by statistical methods like hazard ratio (HR), receiving operator curve (ROC) values as well as clustering (mainly hierarchical clustering). Immunohistochemistry or RT-PCR techniques were applied for experimental validation of the signatures.

Considering this broad spectrum of statistical methods is there a correlation between method and prediction accuracy? This question is difficult to investigate as the overlap of used methods in the original studies is quite low. To answer this question the same analysis should be applied on all datasets.

High range of prediction accuracies in benchmark dataset

Accuracies stated in the original studies range from 60% to 100%. Not all studies used an accuracy measure to prove their performance, but Hazard ratio. The highest accuracy achieved O'Donnell *et al.* [132] for predicting metastasis with 100% (validation based on clustering) as well as Jones *et al.* [90] for subtyping renal cell cancer (validation based on leave-one-out cross-validation). The lowest comparable prediction accuracy gained, 60%, Korkola *et al.* for predicting 5 year overall survival [96].

No network-based methods applied

None of these studies employed network-based approaches and only 13 studies reported prediction accuracies. However, two of these studies [65, 132] are not directly comparable to our accuracy calculation since they did not separate training and test data samples, thus, only 11 studies with reported accuracies are considered.

Benchmark dataset represents the average microarray dataset

To our knowledge it is the first time that such a benchmark dataset was used to test network-based prediction methods. We think that these dataset are good representatives for all microarray datasets, as they cover such a huge variety of diseases, patient and signature size as well as statistical methodologies.

4.2 THE INFLUENCE OF NETRANK PARAMETERS

NetRank depends on a number of parameters: the choice of the genes' initial values that spread through the network, the damping factor which influences the amount of spread, the choice of the network, and the role of noisy and uninformative genes, which are filtered out. Next, NetRank's dependence on these parameters is investigated. The parameters 'Choice of the initial values', the 'Influence of the direct neighbors' and 'The role of noisy and uninformative genes' were investigated on only one dataset of pancreas cancer [196].

4.2.1 Choice of genes' initial values

For the pancreas cancer dataset the best predictions accuracy was obtained by using the rank correlation coefficient of a gene's mRNA level with the survival time. The rank correlation values were then spread by NetRank through the network. Using these dataset specific values is the difference to the original PageRank algorithm, which starts with a uniform distribution of values over all nodes in the network. The question arises, what is happening if the value does not use this specific dataset-dependent fingerprint, but instead a constant value so that all genes are $apriori$ equally important or, from another perspective, if one only relies on network topology?

The initial values influence the prediction results

We tested this and found a prediction accuracy of 62-65% for several training set sizes (ranging from 15 to 28), which is considerably less than the 72% maximum accuracy of the original NetRank. This is interesting since it implies that although some improvement is already gained by focusing on network hubs independent of gene correlations, there is indeed more prognostic information that comes from the use of concrete expression values linked to clinical data.

4.2.2 Influence of direct neighbors

The PageRank can be viewed as an indication of the likely location of a random surfer who iteratively traverses the network. At each iteration the surfer makes a step from one node to one of its neighbors with probability d, while with probability $(1-d)$ he makes a jump to a random node in the network. In NetRank such a random node is selected with probability proportional to either the correlation of the corresponding gene expression with patient survival or the fold-change and *t*-test values of the different outcome groups (given by vector c in Equation 2).

Imagine three nodes A, B, C with edges (A, B) and (B, C). With a damping factor of e.g. $d = 0.3$, the probability of making two consecutive steps from A to C is $0.3 * 0.3 = 0.09$. In consequence, only 9% of information is moved from A to C over B. Thus, the final ranking of a node is obtained with information that comes for more than 90% from initial correlation values and direct neighbors only.

So, how does an algorithm perform that considers only direct neighbors instead of the whole network? In other words, is the global network structure needed to judge each gene, or is its local neighborhood sufficient? A variant of NetRank that spreads values only to direct neighbors was implemented. Each node is ranked according to the average of the initial node values of its direct neighbors.

To our surprise, as shown in Figure 8, this direct neighbor variant performed in the pancreas cancer dataset [196] almost identically to the ranking by Pearson correlation (without network information).

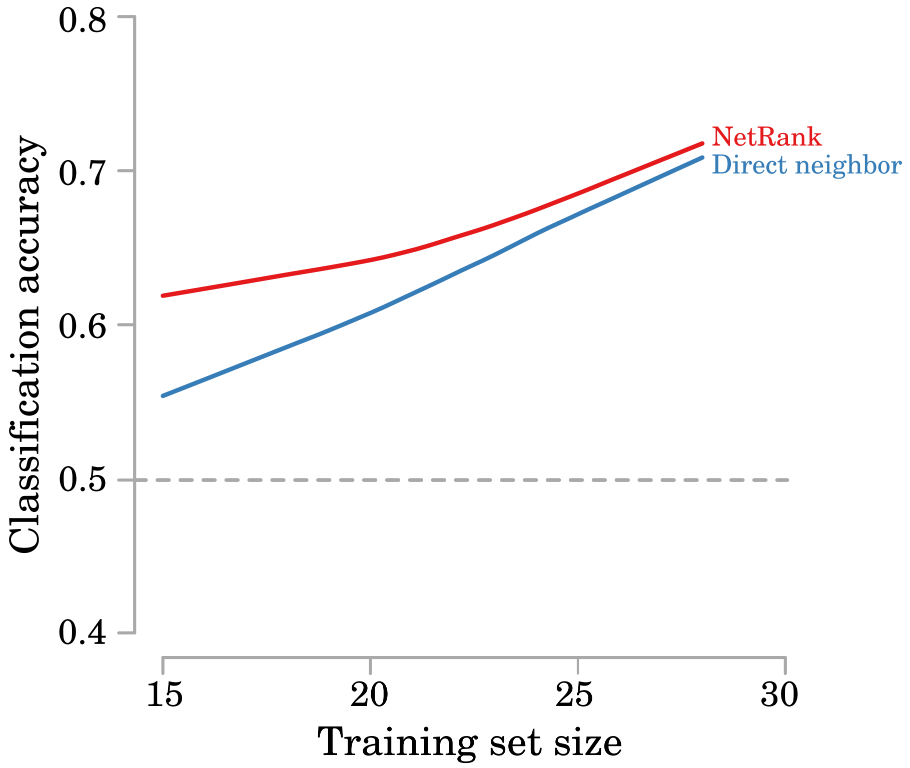


Figure 8: Comparison of NetRank with a direct neighbor algorithm. The plot shows the accuracy of a direct neighbor approach that only takes direct neighbors into account (as opposed to NetRank, which considers all nodes in the network) on the TRANSFAC network with different training set sizes. The direct neighbor approach performs almost identically to the Pearson correlation method (shown here for comparison) [196].

Hence, at least for this study, it is important to also consider distant neighbors. With a sufficiently large number of training samples, NetRank nearly always performs best, but the difference to Pearson correlation and the direct neighbor method becomes sufficiently small. With few training samples, NetRank and the network only (constant c value) approach compete for the best accuracy. It seems that the strength of NetRank is to rely on network topology when data are sparse and correlation can be misleading, and to shift to relying on correlation when sufficient data is available.

4.2.3 The role of noisy and uninformative genes

Low expression and low variance genes were initially filtered out from our microarray data. It is commonly agreed upon that this is a necessary first step when searching for discriminative genes in microarray data. But since NetRank is a network-based, integrative approach, removing genes that could otherwise provide information for their neighboring nodes is probably a suboptimal strategy. For NetRank, we therefore keep all genes, but assign an initial value of zero to those genes that do not pass the filter. This leads for example to down-ranking of a node that has many neighbors with a value of zero, but it also allows for up-ranking of a node with an initial value of zero that has many high value neighbors.

The question if filtering is necessary at all remains. One would expect NetRank to be robust against “noise” since it uses additional

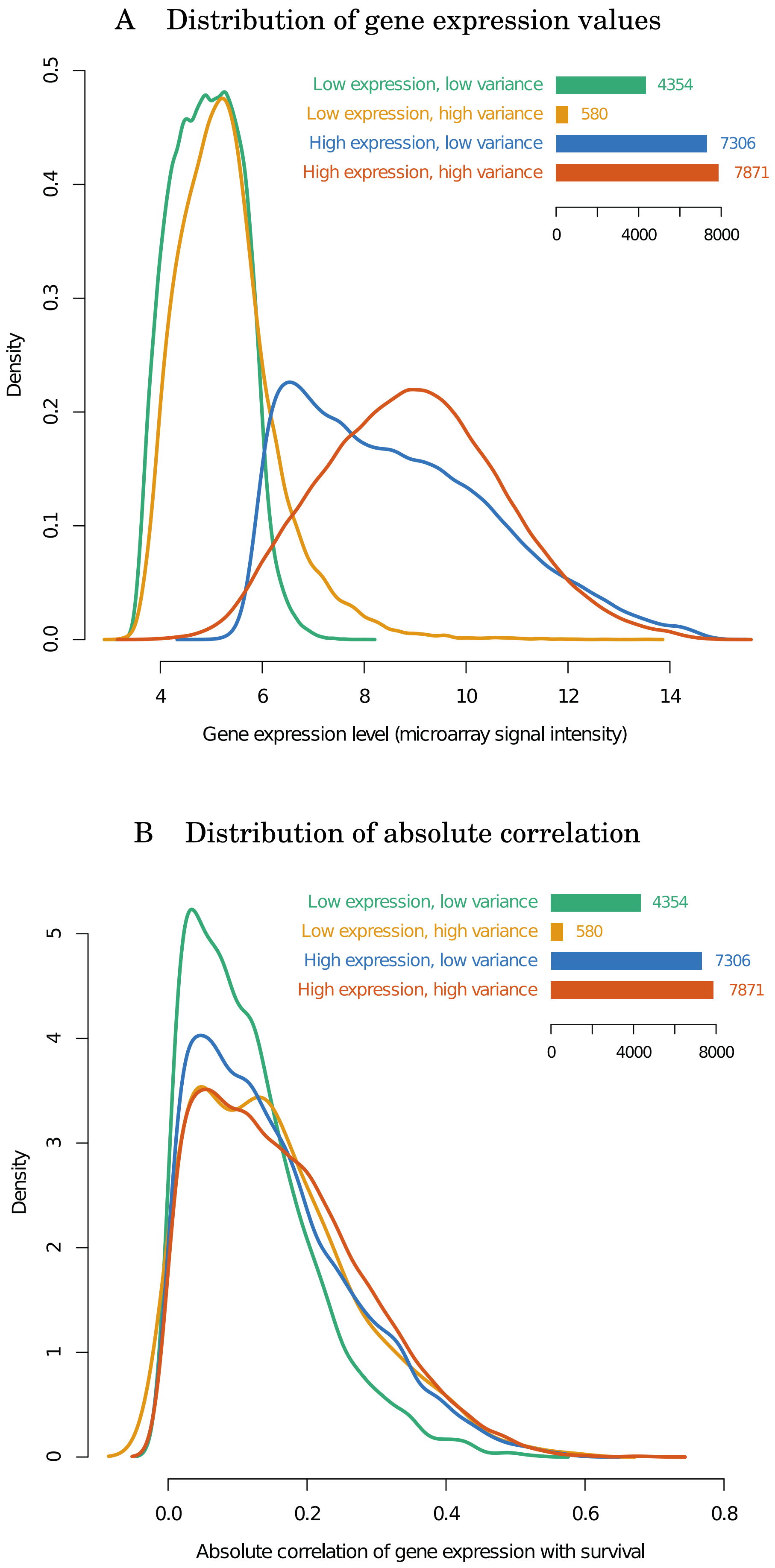


Figure 9: Distribution of expression levels and correlation with survival in four distinct subsets of the full screening dataset. (A) Histogram (density) of gene expression levels. Our filtering keeps only the high expression, high variance genes (red curve). Sizes of the four subsets are shown in the upper right. (B) Histogram (density) of absolute Pearson correlation coefficients of gene expression levels with patient survival. Since the red and the blue curve have very similar distribution, ranking by correlation (which is the starting point for our NetRank algorithm) will allow selection of uninformative, low variance genes (blue curve) that will impair prediction accuracy when included in a classifier. Hence, it is important to filter such genes out [196].

network information that can help to detect and ignore noise. To clarify, our initial filtering actually serves two purposes. It removes not only noise (genes with low expression that are most likely not expressed), but also removes uninformative genes (genes that have low variance and thus cannot be used to discriminate between any classes in the data). Note that uninformative genes can be highly expressed – in the benchmark dataset, one third of the genes have high expression, but low variance.

Noise does not affect the prediction result

To separately assess the effect of noisy (low expression) and uninformative (low variance) genes on the accuracy, further experiments with the pancreas cancer dataset [196] were executed. When we removed uninformative genes but kept noisy genes, the resulting accuracy was still 71% (compared to 72% in the original filtering). This suggests that NetRank is rather robust to noise. However, when we removed noisy genes but kept uninformative genes, the accuracy dropped to 56%, which suggests that NetRank is not robust at all with respect to the presence of uninformative genes. Inspection of the highly ranked genes in this case revealed that the majority of them were uninformative high expression, low variance genes. As the top 10 genes are used for classification in our SVM, it is not surprising for the accuracy to drop if more than half of these genes are uninformative and hence cannot help in classification. A comparison with the original filtered approach showed that the previous top-ranked nodes are still found, but that the uninformative nodes score even higher.

NetRank assigns high scores to uninformative genes

So why does NetRank assign high scores to uninformative genes? There is a simple explanation: the input for NetRank is not gene expression values, but the genes' correlation with survival, the fold-change or the t-test between two outcome groups. So besides the network, the basis for ranking a gene is not its expression value. For the pancreas cancer dataset, the distribution of the correlation coefficients in four different groups (low/high expression/variance) is plotted. As Figure 9B shows, highly expressed uninformative genes have very similar correlation coefficients compared to informative genes. It is virtually impossible for NetRank to detect if a gene will be uninformative for classification, as it can have an equally good correlation value than an informative gene. The fact that many highly expressed low variance genes are apparently correlated with survival suggests that either Pearson correlation is not an ideal measure here, or that the number of samples is too small to give a higher correlation signal in the informative genes. Since Spearman rank correlations show a similar pattern, we believe that the latter is true and that for this dataset size filtering is the best strategy.

Before running NetRank uninformative genes have to be filtered out

To summarize, while NetRank is rather robust against noisy low expression measurements, it is essential to filter out uninformative genes (i.e. set their initial values to zero) before running NetRank.

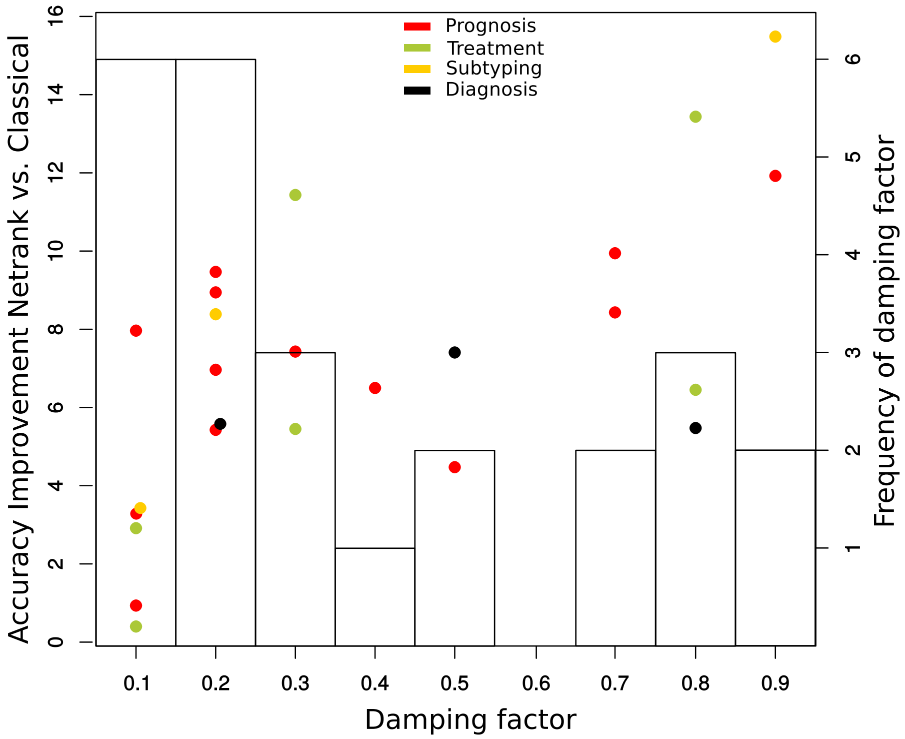


Figure 10: NetRank accuracy versus damping factor. The damping factor measures network influence. The factor is optimized for each dataset. Overall, stronger network influence leads to better accuracy improvements (correlation r = 0.62). Note: The average NetRank accuracies of the datasets were taken as the base in this plot.

4.2.4 Network influence - Influence of damping factor d on NetRank results

As discussed in the Introduction section, one objective is to assess the influence of the network on the outcome prediction accuracy. The PageRank algorithm applied by Goggle search engine uses a damping factor of 0.85. In contrast, in NetRank's random surfer model, the damping factor d is optimized for each individual prediction task, and thus allows judging the network influence. With a $d = 0$, the surfer never follows links and hence the network topology has no influence, whereas a $d = 1$ implies a surfer who only follows links without randomly restarting and thus increases the importance of the network. As the damping factor is not a pre-defined parameter as in original PageRank algorithm, the best damping factor for each dataset is dynamically set as a part of the Monte Carlo cross-validation workflow and ranges from 0 to 1 in steps of 0.1. Figure 10 shows the accuracy improvement of NetRank over the classical approaches plotted against the best damping factor for each literature derived dataset in the benchmark dataset. The figure shows a slight bimodal distribution as either a low $(0.1 - 0.2)$ or a high $(0.7 - 0.9)$ damping factor has been chosen as the best one. For 12 of 25 datasets, d is < 0.3 and for 5/25 datasets d is > 0.7. Therefore, no single damping factor value that yields the best results for all datasets was found. This implies that the underlying networks are not of equal importance in all tasks. One reason could simply be that current interaction networks are still sparse and cover only a small fraction of all interactions, hence important malignant genes might not be connected.

Damping factor correlates with accuracy improvement

Table 5: Damping factors sorted regarding cancer type. There is no representative damping factor for a certain cancer type.

Cancer type	Best d-values
Breast cancer	0.3 , 0.5, 0.7, 0.9
Lymphoma	0.1, 0.2, 0.2, 0.3
Lung cancer	0.2, 0.7, 0.8
Leukemia	0.1, 0.1, 0.8

HPRD's 40 000 interactions and Transfac's 5700 represent only a small fraction of the estimated 650 000 interactions of the human interactome [171]. If the coverage of the network is a key factor, then one would expect a correlation between the damping factor and the accuracy improvement. This is indeed the case with a correlation of $r = 0.62$, meaning by improving network coverage, outcome prediction results will be increased. Besides, Transfac and HPRD might include spurious interactions due to the errors in high-throughput experiments, however such errors are less important than false interactions predicted by text-mining methods [189, 195]. Under the assumption of independence of experimental errors in expression and interaction data, the outcome prediction is clearly improved by using network information. This further implies that if NetRank performs similar to the applied standard methods in selecting disease-relevant genes, current networks poorly cover disease interactions. Regarding the networks, selection of the damping factor does not show a remarkable difference between the transcription factor-target network (Transfac) and a protein-protein interaction network (HPRD).

No unique damping factor as for Google's PageRank

As there is no single damping factor value that yields the best results for all datasets, maybe there is a cancer-specific damping factor, which obtains the best prediction performance for a certain type of cancer? Table 5 summarizes the damping factors for 4 cancer types occurring more than three times in our benchmark dataset. Breast cancer and lymphoma as well as lung cancer and leukemia are investigated in 4 and 3 datasets, respectively. The lymphoma datasets [59, 86, 106, 170] obtain the best results with a relatively low damping factor between 0.1 and 0.3. Indicating, only little network information is needed to reliable predict the patients in the datasets. Nevertheless, no damping factor is in general representative for a certain cancer type.

A higher damping factor indicates strong network support

In datasets having a small damping factor the network information could not help to improve the prediction accuracy, whereas in datasets with a high damping factor the network information drastically improves the prediction results. The reason could be the quality and the strength of the signal in the dataset. In datasets with a strong signal as well as with low noise levels no network information is needed for a good outcome prediction performance. For example in the Jones *et al.* [90] dataset the damping factor is 0.1, indicating a small influence of the network on the results. Similar to the original results, predicts NetRank the different outcome classes with high reliability (100 and 97% accuracy, respectively). In contrast, in the van de Vijver *et al.* [191] dataset a damping factor of 0.7 shows, that quite some network information is needed to boost the prediction result from the original 62-70%.

As a result of the missing of a general damping factor, the dynamically setting of the damping factor has to be maintained for each new dataset.

4.3 EVALUATION OF NETRANK

After assessing the influence of several parameters on NetRank, we are now evaluating the performance of NetRank on the benchmark dataset. Therefore, first, the accuracies obtained with NetRank are compare to the accuracies of random predictors as well as to classical methods for outcome prediction (*t*-test and foldchange). Second, the difference of NetRank and the original results of the benchmark dataset are measured and finally, the performance of NetRank is compared to other network-based methods. A summary of the results can be found in Table 6.

This initial analysis is followed by a second analysis incorporating two larger networks. These networks are evaluated on a subset of the benchmark dataset.

Note that the analysis is based on compact signatures consisting of only 10 genes.

4.3.1 NetRank and random predictors

We created 1000 different random signatures for each dataset contained in the benchmark dataset. These signatures are created by randomly selecting 10 genes out of all genes in each dataset. The outcome prediction accuracy of the random predictors is assessed via a Monte Carlo cross-validation on each dataset.

NetRank outperforms random predictors

As expected, the NetRank algorithm performs better than the random predictor in all datasets, with one exception of the ovarian cancer dataset of Mok *et al.* [122]. Signatures obtained via network-based as well as signatures from classical approaches perform in this special dataset like random signatures. Signatures obtained via *t*-test achieve even a performance that is ten percent worse compared to random. In the remaining datasets, our network-based method is in average 13.5% better, ranging from 3 percent improvement up to 40 percent.

Table 6: Results for 26 datasets covering 13 cancers, four outcome tasks, and different cohort sizes. The comparison shows consistent accuracy improvement of NetRank over random (column *net\random*) signatures and classical methods (*net\no-net*).

Author	**Year**	**Cancer**	**Var.**	**#Pat.**	**Acc†**	**Non-network**		**Network-based**		**Random**	**Accuracy Improvement**			**Best**
						FC	t-test	HPRD	Transfac		net\no-net	net\orig	net\random	**d-value**
Jones *et al.* [90]	2005	Renal cell	S	69∗	100	96	96	97	97	79	1	-3	18	0.1
Spira *et al.* [168]	2007	Lung	D	129	82	66	67	**72**	69	65	4	-12	6	0.8
Lee *et al.* [104] ⋄	2010	Bladder	P	165	79	70	69	73	**76**	56	5	-5	19	0.1
Steidl *et al.* [170]	2010	Lymphoma	T	130	77	72	**75**	73	74	70	0	-4	4	0.1
Bhojwani *et al.* [20]	2008	Leukemia	T	82	74	64	58	63	**64**	51	3	-11	13	0.1
			P	59	73	67	64	70	**72**	54	6	-2	17	0.4
Bogunovic *et al.* [22]	2009	Melanoma	P	38	74	53	63	64	**70**	56	9	-7	11	0.9
Raponi *et al.* [149] ⋄	2006	Lung	P	129	64	59	55	**64**	59	54	5	-3	8	0.2
van de Vijver *et al.* [191] ⋄	2002	Breast	P	295	62	61	58	66	**68**	62	8	5	5	0.7
Wang *et al.* [192]	2005	Breast	P	276∗	62	66	65	**70**	67	66	3	7	3	0.5
Korkola *et al.* [96]	2009	Germ cell	P	82∗	60	67	64	70	**71**	67	5	11	5	0.2
O'Donnell *et al.* [132]	2005	Oral cavity	P	27∗	100⋆	66	78	81	**81**	75	9	-	6	0.2
Friedman *et al.* [65] ⋄	2009	Leukemia	P	68	65⋆	52	**60**	57	55	53	0	-	3	0.1
Landemaine *et al.* [100]⋄	2008	Breast	P	23	-	95	90	**100**	99	67	7	-	33	0.3
Nanni *et al.* [127]⋄	2006	Prostate	D	30	-	89	96	**100**	99	90	7	-	10	0.5
			T	20	-	55	62	**72**	70	31	13	-	40	0.8
Iqbal *et al.* [86]	2010	Lymphoma	D	80∗	-	76	77	81	**82**	67	5	-	15	0.2
Fernandez *et al.* [59]	2010	Lymphoma	S	22∗	-	78	67	80	**81**	44	8	-	37	0.2
Frank *et al.* [64]	2006	Leukemia	T	41∗	-	70	67	70	**75**	66	4	-	7	0.8
Smith *et al.* [166]	2010	Colon	P	55	-	68	57	**72**	71	62	9	-	10	0.2
Lenz *et al.* [106]	2008	Lymphoma	T	414	-	60	59	63	**65**	60	5	-	4	0.3
Mok *et al.* [122]	2009	Ovarian	P	53	-	60	54	64	**65**	65	8	-	0	0.1
Dressman *et al.* [52]	2006	Breast	S	37	-	51	46	55	**64**	43	11	-	17	0.9
Murat *et al.* [125]	2008	Glioblastoma	T	70∗	-	49	46	58	**59**	48	11	-	11	0.3
Zhu *et al.* [206]	2010	Lung	P	133	-	50	46	53	**58**	50	8	-	6	0.7
Winter *et al.* [196] ○	2012	Pancreas	P	30	-	39	65	68	**72**	52	20	-	30	0.3

Accuracy %: 30-34 35-40 41-45 45-50 51-55 56-60 61-65 66-70 71-75 76-80 81-85 86-90 91-95 96-100

† Accuracy of the original studies. If it was not provided by authors and if applicable, it is calculated by equation 3.3.
∗ A subset of the original dataset was used for NetRank validation.
⋄ These dataset did not provide raw data.
⋆ There is no separation between training and test patients.
○ In-house dataset
clinical variables: D – Diagnosis, P – Prognosis/Progression, S – Subtyping, T – Treatment Response/Outcome.

4.3.2 NetRank and classical approaches

We explored the Student's *t*-test and foldchange as classical methods. Both methods are applied as implemented in the genefilter R package. Foldchange finds the log ratio of the means of gene expression for two outcome classes. The t-statistic measure provided by the Student's *t*-test shows the difference between two class means by considering expression variance of the classes. More information about *t*-test and foldchange can be found in section 3.5. In order to select significant genes, we did not apply either a cutoff threshold or multiple correction tests. NetRank outperforms the classical methods in all datasets, except in the lymphoma dataset of Steidl *et al.* [170] and the Leukemia dataset of Friedman *et al.* [65], where NetRank performs as good as the classical methods in average. In the remaining datasets the network-based methods is in average 7.25 percent points better, ranging from one to twenty percent.

NetRank outperforms the classical methods

4.3.3 NetRank and original results

None of the authors of the 25 datasets employed network-based approaches and only 13 studies reported prediction accuracies. We considered only 11 studies with reported accuracies due to no separate training and test samples in two studies [65, 132]. How does NetRank compare to these studies? Caution must be taken here. First, the authors' own computation of accuracy may not be directly comparable to our definition. Second, most of the reported accuracies are substantially better than classical methods. In nine out of eleven studies the published prediction results are better than the classical methods with an average improvement of 9%. This indicates that most of the authors apply advanced analysis methods in the selection of signature genes. For example, Steidl *et al.* applied a sparse multinomial logistic regression to construct signatures in combination with clinical variables [170]. Bogunovic and co-workers derived a predictive signature by using principal component analysis [22]. Lee *et al.* assigned E2F1 as the predictor gene to identify the invasive progression of bladder tumors by using up-modulated gene expression and prior knowledge [104].

Original results are better than classical methods

Nonetheless, we compared NetRank to the author accuracies: For three out of eleven datasets NetRank is better, but on average NetRank is two percent worse. The difference ranges from -12 up to 11 percent accuracy points. This indicates that NetRank's improvement over classical methods is approaching the level of the authors' signatures. Taken together, a fully automated network-based approach – like NetRank – is able to obtain similar results as the hand-selected signatures of the authors.

The purely computational approach of NetRank approaches the level of the authors' signatures

In the renal cell cancer dataset of Jones *et al.* the signal seems to be really strong as it could be captured by the authors as good as by NetRank [90]. Both methods achieved close to 100 percent prediction accuracy.

4.3.4 NetRank and other networks

For further evaluation of the influence of different networks, we applied NetRank to two larger networks: STRING and hPrint.
Table 7 shows the results of 13 datasets on these two networks. The four networks deliver similar prediction performances as the accuracies lie between $71 - 75\%$. Transfac performs slightly better, as it obtains an average prediction accuracy of 75.7% in comparison to HPRD, hPrint and STRING, which result in an average accuracy of 74.2, 73.9 and 71.4 percent points, respectively. The superior performance of Transfac suggests that regulatory information is particularly suited for outcome prediction. Most cancer types arise due to mutations in regulatory elements, therefore the efficiency of Transfac in cancer outcome prediction is biologically reasonable.

Transcriptional information especially suited for outcome prediction

The accuracies between the different networks do not show a high variance, except for the dataset of Fernandez *et al.* [59] and Dressmann *et al.* [52]. In the lymphoma dataset of Fernandez *et al.* the STRING networks results in an accuracy 16% worse compared to the average of the remaining networks. The hPrint network outperforms the networks and lead to an accuracy of 87%. In the breast cancer dataset of Dressmann *et al.* the result is different. STRING, hPrint and HPRD result in an accuracy of around 54%, whereas Transfac obtains a prediction accuracy of 64%.

4.3.5 NetRank and other network-based methods

NetRank provided better outcome predictions compared to non-network-based methods, hence it is an interesting question whether other network-based methods also lead to improvements. To test this hypothesis, the prediction performance of NetRank is compared with other network-based methods [36, 42] on two breast cancer metastasis datasets [191, 192].

The dataset of Wang *et al.* [192] employs the expression of 22.000 transcripts from total RNA of frozen tumor samples. The datasets contains 286 lymph node- negative primary breast cancer samples that is composed of 77 estrogen-receptor negative (ER-) and 209 estrogen-receptor positive (ER+) samples. The gene expression profiles were analyzed with Affymetrix Human Genome U133A Array. The van de Vijver *et al.* [191] gene expression dataset consists of 295 samples, including 151 lymph node-negative disease and 144 lymph node-positive disease. There are approximately 25,000 human genes transcribed and labeled to self-made microarrays for each sample.

Table 7: Results for the 13 datasets for NetRank applied on 4 different networks (STRING, hPrint, HPRD and Transfac). None of the networks is superoir to the others as they all show comparable prediction accuracies.

Author	Cancer	Var.	hPrint	STRING	HPRD	Transfac
Jones *et al.* [90]	Renal cell	S	**98**	97	97	97
Spira *et al.* [168]	Lung	D	71	70	**72**	69
Bogunovic *et al.*[22]	Melanoma	P	66	63	64	**70**
Korkola *et al.* [96]	Germ cell	P	**72**	70	70	71
O'Donnell *et al.* [132]	Oral cavity	P	76	75	81	**81**
Landemaine *et al.* [100]⋄	Breast	P	96	92	**100**	99
Iqbal *et al.* [86]	Lymphoma	D	77	79	81	**82**
Fernandez *et al.* [59]	Lymphoma	S	**87**	66	80	81
Frank *et al.* [64]	Leukemia	T	70	72	70	**75**
Smith *et al.* [166]	Colon	P	72	71	**72**	71
Mok *et al.* [122]	Ovarian	P	61	60	64	**65**
Dressman *et al.* [52]	Breast	S	54	54	55	**64**
Murat *et al.* [125]	Glioblastoma	T	**61**	59	58	59

clinical variables: D – Diagnosis, P – Prognosis/Progression,
S – Subtyping, T – Treatment Response/Outcome.

Accuracy %:	30-34	35-40	41-45	45-50	51-55	56-60	61-65
	66-70	71-75	76-80	81-85	86-90	91-95	96-100

In this comparison, NetRank accuracies based on Transfac are used and the accuracies from the two network-based methods are taken from the respective publications [36, 42].

Both methods integrate gene expression and network data sets.

APPROACH OF CHEN *et al.* Chen and co-workers work with network-constrained support vector machines [36]. Their proposed method adapts the assumption that hub genes are usually expressed with low variability which requires the incorporation of network information to improve generalizability of the results. Therefore, the network information is explicitly formulated as a Laplacian matrix and embedded into the objective function of SVM for optimization. This leads to an improved classification as the contribution of hub genes to the SVM is greatly enhanced; even hub genes not significantly differently expressed between two phenotypes influence the resulting hyperplanes. Significant genes or subnetworks are detected through a significance test based on label permutation.

APPROACH OF CHUANG *et al.* In the approach of Chuang *et al.* [42], a candidate subnetwork is scored to assess its activity in each patient. The assessment of a subnetworks activity is based on averaging its normalized gene expression values. Afterwards, for each subnetwork the discriminative potential is measured based on the mutual information between its activity score and the metastatic/non-metastatic disease status over all patients. To identify significantly discriminative subnetworks, their discriminative potentials are compared to those of random networks. In their study, Chuang and co-workers apply their method on breast cancer patients in two data sets.

Results

All network-based methods perform comparable

The prediction results of the three network-based methods on two breast cancer datasets are shown in Table 8. All three methods perform comparably well in both breast cancer dataset in comparison to non-network-based approaches The approach of Chuang *et al.* is slightly superior with 2% and 5% percent points in the dataset of van de Vijver *et al.* and Wang *et al.*, respectively. Chuang and co-workers achieve a similar accuracy to NetRank, but outperform regarding sensitivity. They achieve in both breast cancer datasets a sensitivity of 90%. In comparison to Chen *et al.*, our approach has similar prediction accuracies but showing better sensitivity scores. These variances in the sensitivity and specificity measures of the three network-based methods might be caused by the usage of different machine learning methods and cross-validation procedures during the training of the classifiers.

In comparison to the classical methods, all three network-based methods improve the prediction result. The methods are in the dataset of van de Vijver *et al.*in average 7 and 10% better than foldchange

Table 8: Comparison of NetRank with other network-based methods. All three methods perform comparable in both breast cancer datasets. The variances in the sensitivity and specificity measures of the three network-based methods might be originated by the usage of different machine learning methods and crossvalidation procedures during the training of the classifiers.

Method	**Accuracy**	**TPR**	**TNR**	**Dataset**
NetRank	68	69	62	van de Vijver *et al.*
Chen *et al.* [36]	65	51	74	
Chuang *et al.* [42]	70	90	63	
Foldchange	61	-	-	
t-test	58	-	-	
NetRank	67	70	53	Wang *et al.*
Chen *et al.* [36]	71	42	81	
Chuang *et al.* [42]	72	90	62	
Foldchange	66	-	-	
t-test	65	-	-	

and *t*-test, respectively. In the dataset of Wang *et al.*the methods outperform the classical methods with 4 and 5%.

These results confirm that network information improves outcome prediction in comparison to the classical methods as well as to the original results.

4.3.6 Comparison to the related work of Cun and Fröhlich

Cun and Fröhlich applied fourteen different genes selection methods (eight incorporate network information) on six public available breast cancer datasets [44]. They could not identify one single of these fourteen approaches to perform best regarding prediction performance, which stands in contrast to the results found in this thesis.

Their statement is based on datasets of one kind of cancer – breast cancer – and one type of network – non-metabolic KEGG pathways. The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a manually curated database consisting of directed pathways representing experimental knowledge on metabolism and various other cell functions. The authors combined these information with data from the Pathway Commons database [34] resulting in a graph consisting of 13.840 nodes and 397.454 edges.

KEGG pathway information not useful for network-based cancer outcome prediction

In our analyses transcriptional information – as provided by Transfac – tend to be more suited for outcome prediction than the remaining network types. Directed pathway information is not included into our analyses, but following Cun and Fröhlich's results KEGG's pathways information does not seem to improve cancer outcome prediction.

The breast cancer dataset of Wang *et al.* [192] is included in our benchmark dataset. Unfortunately, the performance of the algorithm is stated as area under the curve (AUC) instead of prediction performance, thus results are not directly comparable. None of the eight network-based methods reached an AUC larger than 0.70, supporting the hypothesis that KEGG pathway information is not appropriate for the use in outcome prediction. According to Cun and Fröhlich findings, also our results indicate that for this dataset network information in general is not greatly enhancing the prediction accuracy. The application of network information through NetRank improves the predicting accuracy only for 3% compared to the classical methods. In addition only two out of the six non-network based approaches reached an average AUC of 0.70 and higher.

An algorithm using GeneRank is included into the ensemble of tested approaches. The algorithm – called Reweighted Recursive Feature Elimination (RRFE) – identifies significant genes by stepwise elimination of features [89]. The algorithm is explained in great detail in subsection 2.2.3. This approach reached an average AUC of 0.67 on the breast cancer dataset of Wang *et al.*. On all six breast cancer datasets the algorithm performs with an average AUC of 0.63, reaching from AUC 0.51 – 0.76, making it the network-based approach with the highest variance in AUC over all analyzed datasets.

The beneficial effect of background knowledge differs between datasets

For some datasets the use of background information is more beneficial than for others. In the case of the Wang *et al.* dataset the non-network based methods achieve an higher average AUC than the network-based methods, but the results look different for the breast cancer dataset of Desmedt *et al.* where the network-based methods are in average better than the applied classical non-network based methods.

Thus, their statement that in general pathway knowledge does not significantly improve cancer outcome prediction is conclusive, but following our results the picture looks different when applying different types of protein-protein interaction networks. In addition depends the beneficial effect of network information on the applied dataset.

Network-based signatures overlap

In concordance to our results they found the similarity and biological interpretability of resulting signatures significantly enhanced using network-based approaches (see chapter 5).

4.4 GENERAL FINDINGS

In the last section the performance of NetRank is compared with random predictors, classical methods and the results of the original studies as well as to other network-based methods. It has been shown that incorporation of network information is able to dramatically enhance the prediction accuracy for different outcome prediction tasks.

In the following section further general parameters are discussed that might influence the prediction results.

4.4.1 Physical interactions as good as regulation

For assessing the performance of NetRank, four different networks are applied: HPRD [146], Transfac [119], STRING [63] and hPrint [56]. The networks differ in type of interaction and size. HPRD covers physical interactions contains nearly 10.000 genes and 40.000 interactions. Similar in size of nodes are STRING and hPrint. HPrint uses predicted physical and functional interactions and contains roughly 13.000 nodes and 235.000 edges. STRING with its 11.000 nodes and 340.000 edges is a database of known and predicted protein interactions. The smallest of all four networks is Transfac, which is a network of transcription factors and targets and has only some 2.400 genes and 5.700 interactions.

As shown in Table 6 and Table 7, there is no substantial difference in prediction accuracy, while applying NetRank on these networks. Transfac is better than HPRD in 16 out of 25 cases with an average improvement of 1 percent point. Regarding STRING and hPrint, Transfac outperforms in 7 out of 13 datasets, with an average improvement of 4.3% and 1.8%, respectively.

Transcriptional information best suited for outcome prediction

Nevertheless, employing only the Transfac network achieves an equal performance using far less genes. Cancer often arises due to alterations in transcription factor expression, leading to unregulated cell growth and differentiation. Therefore, the efficiency of Transfac in cancer outcome prediction is biologically reasonable. Transcription factors have been already discussed as one of the main source of cancer development [128] as well as suggested as targets for cancer therapy [47]. This implicates that future regulatory networks lead to further improvement in outcome prediction.

4.4.2 Influence of cancer bias

Network information is generally extracted from literature studies. Therefore genes in such networks are often highly related to certain diseases, e.g. cancer and many biological interaction networks typically have small diameter due to the presence of 'hub' genes of high degree. There are reports that cancer-associated genes have more interaction partners than non-cancer genes [91, 110], and indeed genes found in NetRank signatures like SP1 have high degree in most interaction networks (e.g., the degree of SP1 in Transfac is 421).

In order to assess the impact of the well-studied cancer proteins in a network, the analysis was limited to these well-studied genes. For this purpose, genes mentioned in Transfac are solely provided as an input to the classical methods (foldchange, *t*-test) and their prediction accuracy is assessed via leave-one-out cross-validation.

As indicated in Table 9, the prediction accuracies of the classical methods obtained by this limited gene set are generally lower than

Table 9: The prediction accuracies of the classical methods (foldchange (FC), *t*-test (TT)) by using only Transfac genes. The classical methods select the signatures from Transfac genes rather than from the entire chip. The NetRank (NR) column shows the prediction accuracy obtained by Transfac. The results prove that the NetRank algorithm can efficiently use regulation information by following edge relations in Transfac, hence it provides more predictive signatures compared to the classical methods. The best result per dataset is marked in bold.

Author	Cancer	Var.	FC	TT	NR
Jones *et al.* [90]	Renal cell	S	73	77	**97**
Spira *et al.*	Lung	D	67	64	**69**
Lee *et al.* [104]	Bladder	P	59	60	**76**
Steidl *et al.* [170]	Lymphoma	T	70	69	**74**
Bhojwani *et al.* [20]	Leukemia	T	46	50	**64**
		P	56	47	**72**
Bogunovic *et al.* [31]	Melanoma	P	56	62	**70**
Raponi *et al.* [149]	Lung	P	53	52	**59**
van de Vijver *et al.* [191]	Breast	P	63	63	**68**
Wang *et al.* [192]	Breast	P	64	65	**67**
Korkola *et al.* [96]	Germ cell	P	66	66	**71**
O'Donnell *et al.* [132]	Oral cavity	P	75	75	**81**
Friedman *et al.* [65]	Leukemia	P	51	**57**	55
Landemaine *et al.* [100]	Breast	P	67	67	**99**
Nanni *et al.* [127]	Prostate	D	97	96	**99**
		T	25	40	**70**
Iqbal *et al.* [86]	Lymphoma	D	68	73	**82**
Fernandez *et al.* [59]	Lymphoma	S	38	39	**81**
Frank *et al.* [64]	Leukemia	T	66	67	**75**
Smith *et al.* [166]	Colon	P	62	62	**71**
Lenz *et al.* [106]	Lymphoma	T	60	63	**65**
Mok *et al.* [122]	Ovarian	P	35	35	**65**
Dressman *et al.* [52]	Breast	S	39	36	**64**
Murat *et al.* [125]	Glioblastoma	T	35	36	**59**
Zhu *et al.* [206]	Lung	P	48	47	**58**

Accuracy %:	30-34	35-40	41-45	45-50	51-55	56-60	61-65
	66-70	71-75	76-80	81-85	86-90	91-95	96-100

clinical variables: D – Diagnosis, P – Prognosis/Progression,
S – Subtyping, T – Treatment Response/Outcome.

NetRank results based on Transfac. Note, the non-limited results for all methods can be found in Table 6.

Limited search space leads to worse prediction results for classical methods

Limiting the dataset to only well-studied proteins achieves worse prediction accuracy in 20 out of 25 datasets for foldchange and *t*-test, respectively. For foldchange, the result got in average 11% worse for 20 datasets but for five datasets (O'Donnell, Spira, Bogunovic, van de Vijver, Nanni D) the prediction accuracy improved in average for 4.5 percent points compared to the original non-limited results. In the lymphoma dataset of Fernandez *et al.* the accuracy declined for 40 percent points in the limited dataset. It only reached a prediction accuracy of 38%. In contrast, the accuracy in the oral cavity cancer study of O'Donnell *et al.* and the prostate cancer dataset of Nanni *et al.* improved drastically for 9% and 8%. In the case of *t*-test a similar picture can be drawn. The limited results are in 20 out of 25 datasets worse than the non-limited results. In these 20 datasets the accuracies declined in average 9%. In the remaining datasets (van de Vijver, Landemaine, Smith, Lenz, Zhu) the results got in average 3.4% better. In the limited *t*-test results, he performance in the lymphoma dataset of Fernandez strongly declined for 28% and only reached an accuracy of 39 percent. In comparison, the breast cancer dataset of van de Vijver and co-workers gained 5% in prediction accuracy.

Cancer bias does not influence the prediction performance of NetRank

In general, roughly 75% of all datasets achieve worse prediction accuracies compared to the full dataset. In datasets already having a bad performance of around 50%, limiting the gene set to only the well-studies genes, decreases the accuracy even below the performance of a random predictor. The remaining datasets keep their prediction performance.

The results indicate that the information gain achieved by NetRank is not solely based on well-studied cancer proteins but also on the interaction between them.

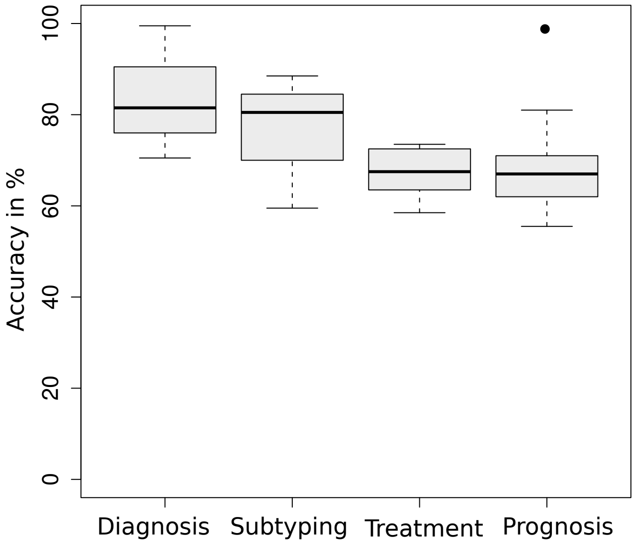


Figure 11: NetRank accuracy vs. outcome task. Diagnosis and subtyping are easier than treatment or prognosis. Note: The average prediction accuracies of Table 6 NetRank column were taken as the base in this plot.

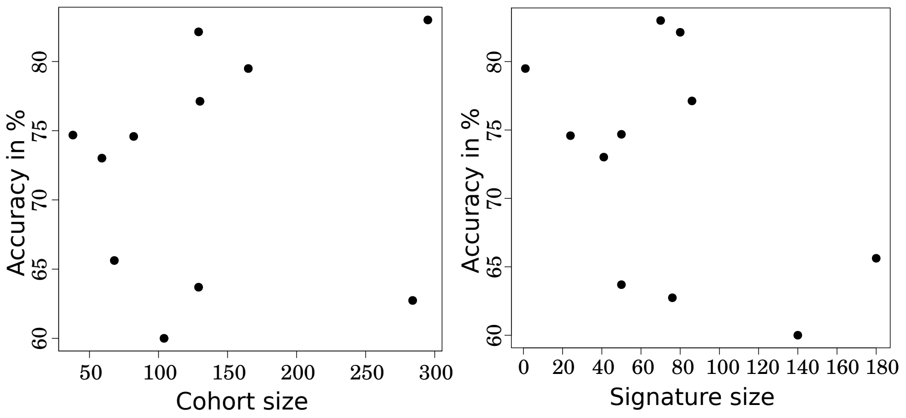
4.4.3 Prognosis and treatment is more difficult to predict than diagnosis and subtyping

Outcome prediction defines the future state of a patient based on its current disease state and focuses on the following topics: (1) Treatment response deals with the different response types of patients to treatments. Here, response means a prolonged metastasis-free survival or faster response to therapy. (2) Prognosis/Progression deals with cancer development after (3) Diagnosis. Finally, (4) Subtyping aims to distinguish different subtypes of cancer. Subtypes may alter in the aggressiveness of the cancer along with different spreading behavior and may end up in different treatment needs.

Figure 11 summarizes NetRank's average accuracies of Table 6 for the four different outcome prediction tasks. As the figure shows, diagnosis and subtyping can be predicted more reliable than response to treatment and prognosis. While datasets that try to diagnose cancer or predict subtypes have in average an accuracy of around 81%, the outcome variables treatment and prognosis only result in prediction accuracies around 66%.

Diagnosis is easier to predict than prognosis and treatment

One explanation for this result could be that treatment and prognosis depend more on external, non-molecular parameters such as age, sex or different living conditions. Another possible explanation is the difference in gene expression between outcome groups. When diagnosing a malignant disease, healthy and cancerous tissues are compared. These two types of tissue express highly different number of proteins [169], whereas in the case of progression and treatment response prediction, the difference in the gene expressions is very subtle. In addition, due to the combination of background noise in microarray experiments and low expression levels, these subtle differences are not reliably detectable [94], thus resulting in worse prediction performance.



(a) Accuracy vs. cohort size

(b) Accuracy vs. signature size

Figure 12: Accuracy vs. cohort size (a) and signature size (b). Prediction accuracy does not correlate with the number of patients (correlation $r = -0.14$). Nevertheless the signature size seems to influence the prediction results (correlation $r = -0.47$). The analysis is based on 11 accuracies, which were provided in the original publications.

4.4.4 Influence of signature and cohort size on the prediction accuracy

An interesting question that arises is whether the size of signature and patient cohort influence the prediction results. Also, a small cohort could be prone to more artifacts and hence having a better accuracy than a large and more representative study. Hence, we assessed whether signature size and prediction accuracy are correlated.

As our network-based method always produces signatures with a fixed length of 10 genes, these issues are explored by using the original accuracies as provided in 11 publications.

No influence of cohort size on prediction performance

While investigating the effects of these parameters on the prediction accuracies, no clear correlation could be observed, indicating any effect of the cohort size on the prediction accuracy. As can be observed in Figure 12a, prediction accuracies are not related to cohort size as the correlation between these two parameters is -0.14. Bogunovic and co-workers [22] achieve with 38 patients a similar good performance as Steidl *et al.* with 130 patients [170]. It has been shown, that an exact outcome prediction is possible with a small cohort of only 27 patients [132]. In this study the progression of oral cavity cancer is predicted with 100% accuracy.

The signature size influences prediction accuracy

Nevertheless, the signature size – at least in our data – seems to influence the prediction results. The signature size correlates with $r = -0.47$ to the prediction performance. Figure 12b plots the signatures size against the accuracy published by the authors. Smaller signatures tend to have higher accuracies and large signature does not ensure high prediction performance. Particularly, having the biggest signature of 140 genes, Korkola *et al.* performs worst of all benchmark datasets with an accuracy of 60%. Whereas, Lee *et al.* [104] obtains with the smallest signature of one gene an accuracy of 79%.

4.4.5 Influence of the machine learning method

During experiments, we were faced with the problem of prediction accuracies significantly lower than 50%. For instance, Support Vector Machines (SVMs) could not clearly be trained on the Mok *et al.* [122] and Nanni *et al.* [127] datasets. Notably, the few patient samples ($n = 20$) in the treatment dataset of Nanni *et al.* might be a reason of the poor performance of SVM. Therefore, we assumed that other machine learning approaches would be more suitable. To assess the prediction performance of different machine learning approaches, we trained them using the 10-gene signatures obtained by NetRank in one dataset and tested them on all other datasets (Figure 5: Steps 5–7). The prediction capability of the four different machine learning methods is determined and their performance for outcome prediction is assessed.

DECISION TREES Decision trees (DT) are a classifier, which uses a tree-like graph to model decisions and their possible consequences. In a decision tree an internal node represents a certain test on an attribute, each branch represents the outcome of that test and each leaf represents the class label. A path from root to leaf represents classification rules. The tests on the internal nodes can be any other machine learning method.

K-NEAREST NEIGHBOR The k-nearest neighbor approach (k-NN) is a non-parametric classification method based on the Euclidean distance between a test sample and the specified training samples. Due to a majority vote of its neighbors a test object is classified with the class being assigned to most of its k nearest neighbors.

LINEAR DISCRIMINANT ANALYSIS Linear discriminant analysis (LDA) is a commonly used technique for data classification and dimensionality reduction. LDA finds a linear combination of features to separate two or more classes and is able to handle unequal within-class frequencies. LDA explicitly models the difference between the classes of data by maximizing the ratio of between-class variance to the within-class variance in any particular data set.

RANDOM FOREST Random forest (RF) is an ensemble learning technique that operates by creating a multitude of decision trees by searching over random subsets of trees. To classify a new object, the input is classified by the trees in the forest and each tree decides a classification for the input. The resulting class is defined via a majority vote of all trees.

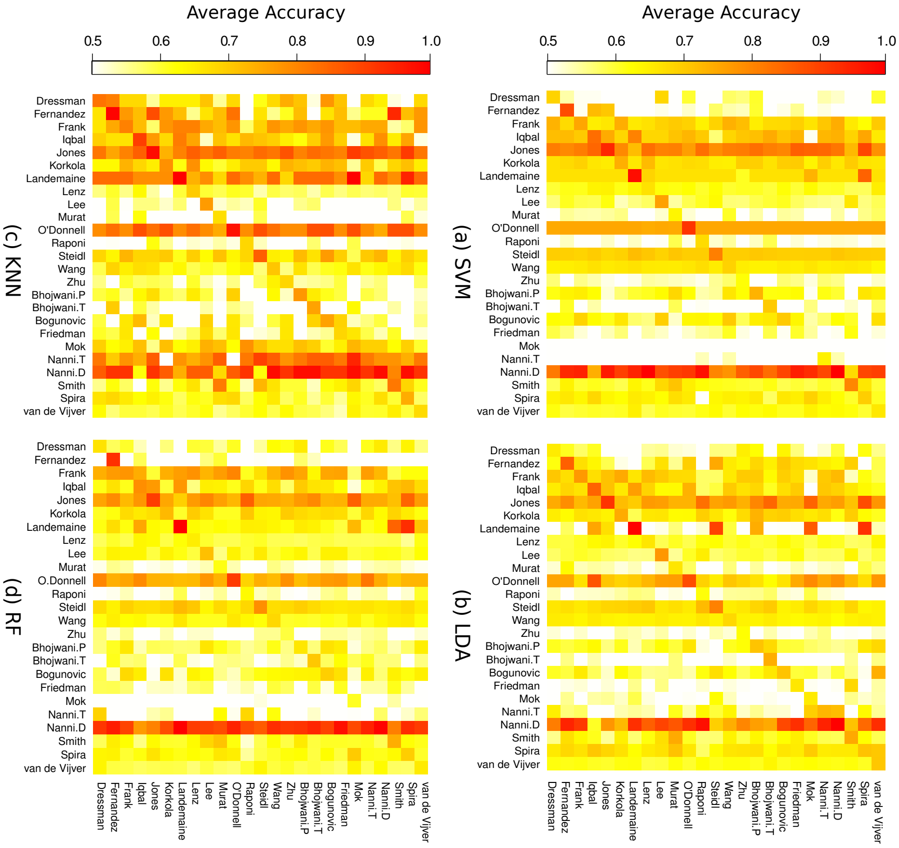


Figure 13: Accuracy heatmap of transferred signatures with different machine learning methods. Signatures were found with NetRank on Transfac. Rows and columns represent the predictive signatures and datasets, respectively. So, the diagonal shows the accuracy of the signature on the original dataset. The accuracies lower than 50% (worse than a random guess) are indicated in white color. Different machine learning techniques gave different accuracies. SVMs have an overall lower accuracy and fail to predict progression outcome of [122, 127]. For our datasets, the k-NN approach has an overall higher accuracy than SVM, LDA and RF (3%, 2% and 2 % average improvement). Nevertheless the differences are subtle, thus the performance of the different machine learning methods are comparable to each other. In addition, following the no-free lunch theorem, there is no ideal machine learning method providing high classification accuracy independent of context and data [53].

Test results

The results for SVM, LDA, k-NN and RF are given in Figure 13. DTs delivered low prediction accuracies. Therefore the results are not shown.

For our datasets, the k-NN approach has an overall higher accuracy than SVM, LDA and RF with 3%, 2% and 2 % average improvement, respectively. Over all datasets the k-NN approach achieves a prediction accuracy of 63 percent points. It generally outperforms the other approaches in almost all datasets. The k-NN is in 20, 22 and 22 out of 25 datasets better than RF, SVM and LDA, respectively. Interestingly, it is consistently worse than the other approaches in the dataset of Nanni *et al.* (T) [127] and Murat *et al.* [125], whereas it predicts the Jones *et al.* [90] and Mok *et al.* [122] dataset persistently better, with roughly 4 percent points. The signal in the Mok *et al.* dataset seems to be difficult to capture by RF, SVM and LDA as these methods obtain accuracies around 50% (white squares), whereas the k-NN approaches is able to handle the data. In general, is the performance of RF almost similar to LDA.

These experiments show that the prediction performance of the NetRank could be further improved by using several machine learning methods. Nevertheless, the differences are subtle. Thus, the performance of the different machine learning methods is comparable to each other. In the benchmark dataset the average cohort size is 103 patients, so differences in prediction accuracy of one percent lead to the correct prediction of only one extra patient.

In addition, as stated in the No-free lunch theorem [53], there is no ideal machine learning method providing high classification accuracy independent of context and data.

4.5 COMPUTATIONAL COMPLEXITY OF NETRANK

In the following section NetRank is analyzed regarding space and time complexity.

Space complexity

The space complexity of an algorithm is equal to the memory space used by the algorithm to solve a given computational problem as a function of the size of the input.

Space complexity grows linear to the number of edges and patients

For NetRank the patient and network data is loaded into memory for an efficient calculation, which introduces a memory-related overhead. The space complexity of the algorithm depends on three factors: number of nodes in the network (n), number of edges in the network (e) and number of patients (p). During this evaluation the number of nodes in the network is constant with approximately 22.000 human genes. Therefore the space complexity grows linear with the number of edges as well as with the number of patients in $\mathcal{O}(e * p)$.

Memory is the bottleneck

As in the protein networks more than 50% of the nodes do not have edges to any other nodes, the networks are implemented in a sparse matrix representation, resulting in a matrix primarily populated with zero. Sparse matrices are conserving space by representing only the non-zero entries. During calculation the sparse matrix has to be populated for the matrix-vector multiplication, resulting in an increased memory usage. In the current version, running the matrix-vector multiplication with a network of approximately 100.000 edges consumes at least 2 Gigabyte of memory. For current network sizes, NetRank can be run on commodity hardware having 4 GB RAM, but with increasing network sizes this is growing up to terabytes, espe-

Table 10: The running time of NetRank differs between the datasets, depending on cohort and network size. Table shows maximum running time in hour per process for datasets having different cohort sizes. Note: Lenz *et al.*was not evaluated on STRING and hPrint.

Dataset	**Cohort**	**Running time for 1 cross-validation**			
	size	Transfac	HPRD	STRING	hPrint
Lenz *et al.*	414	20h	40h	-	-
Spira *et al.*	129	12h	25h	57h	63h
Landemaine *et al.*	23	8h	12h	23h	27h

cially when testing several networks at the same time, thus making the memory the major bottleneck of the algorithm.

The memory usage of the machine learning algorithms is less significant.

Time complexity

The time complexity of an algorithm quantifies the amount of time an algorithm needs to run as a function of the length of the input.

Quadratic time complexity

At the core of the NetRank algorithm are two important operations, matrix-vector multiplications and machine learning algorithms. The machine learning algorithm used during NetRank evaluation are Support Vector Machines which run in a complexity of $\mathcal{O}(n)$ [35] and the matrix-vector multiplication is done in $\mathcal{O}(n^2)$ with n being the amount of nodes in a network. The multiplication is done 500 times in the outer validation and 500 times in the inner validation for each damping factor. This sums up to 250.000 matrix multiplications per task, but does not influence the time complexity. Therefore the complexity of NetRank grows quadratically with the size of nodes.

EVALUATION TIME One problem in biomarker discovery is the overestimation of prediction performance due to biased signatures. To overcome the problem of signatures being biased to either the training or test set, NetRank performs a supervised learning process combined with a Monte Carlo cross-validation scheme to predict biomarker genes. Therefore, each dataset is split 500 times. Table 10 shows the running time of NetRank for a single cross-validation on datasets having different cohort sizes. The evaluation of NetRank on the biggest dataset of Lenz *et al.* [106] took 30.000 CPU hours (3.4 CPU years). This estimates to approximately 60 CPU years for the evaluation of the benchmark dataset on HPRD and Transfac. With 26 datasets – each having roughly 22.000 genes – being evaluated on two networks (Transfac and HPRD) in a 500 fold cross-validation procedure 572 million PageRank's had to be calculated.

4.6 DISCUSSION

Although the first diagnostic assays for cancer outcome prediction are on the market allowing patients to personalize treatment options, the criticism remains that large signatures appear random, since other signatures perform equally well. Here, we aimed to address this problem by focusing on signatures that are compact (only 10 genes) and that were obtained with a network-based approach, which puts the signature genes into a context rather than selecting isolated genes.

Network information improves cancer outcome prediction

Indeed, we found on 26 datasets – covering different cancers, outcome tasks and cohort sizes – that there is a consistent and significant improvement of the network-based approaches over random signatures and a classical approach based on *t*-test or foldchange. Besides, we are approaching the accuracy level of the authors' signatures by applying a relatively unbiased but fully automated process for biomarker discovery.

NetRank performance depends on damping factor

We furthermore investigated several parameters influencing the result of NetRank. e.g. how the genes initial values or the damping factor influences the results. We show that the choice of the damping factor regulates the influence of the network on the results, as it balances the impact of expression and interaction on the prediction result. A weak correlation of d to accuracy improvement was observed, suggesting that increasing coverage of the interactome may also lead to further improvements in the prediction accuracy. The novelty of NetRank is the dynamical setting of the damping factor during signature development. As each dataset has its own dataset specific damping factor, the dynamical setting has to be maintained for each new biomarker creation.

Regulatory information best suited for outcome prediction

An important open question is also the influence of the network topology on the analysis. We experimented with a regulatory (Transfac) and physical interaction (HPRD) and two predicted networks (hPrint, STRING). The latter are several orders of magnitude larger than the first. Nonetheless, all can be considered as a fraction of the complete interactome that is currently not yet known. Interestingly, despite the size difference, they perform equally well. This suggests that regulatory information is particularly suited for outcome prediction. Most cancer types arise due to the mutations in regulatory elements, therefore the efficiency of Transfac in cancer outcome prediction is biologically reasonable.

Limited search space leads to worse prediction results for classical methods

Furthermore, we evaluated the disease bias in networks as genes in networks are often highly related to certain diseases. We show that this bias does not affect the prediction abilities of NetRank. In contrast, limiting the search space to only well-studied proteins decreases the result of the classical methods dramatically.

Diagnosis is easier to predict than progression or treatment

Predicting response to treatment and progression are generally more difficult than diagnosis and subtyping. We hypothesized that the former is more strongly influenced by external factors such as age and sex. Furthermore, we conclude that for diagnosis the gene expression signal is much stronger. This can be proven by comparing the accuracy obtained via random predictors in the prostate dataset with any other accuracy obtained by more sophisticated methods. The signal is strong enough to be captured even by random gene selection.

Signature size influences prediction performance

Another question was the correlation of cohort and signature size to prediction performance. On one hand we found a small correlation of signature size with prediction accuracy, but on the other hand no correlation for cohort size and accuracy could be observed.

Choice of machine learning method shows only small impact on accuracy

Finally, we investigated the influence of machine learning approaches on outcome prediction. We could clearly observe a difference between diverse machine learning approaches. Overall, the k-NN approach performs best. Nevertheless, the differences are subtle and not statistically significant.

NetRank is efficient

Furthermore, the time and space complexity of NetRank scales quadratically with the input size, resulting in an efficient method for biomarker discovery.

As a summary, network-based gene expression analysis is leading to a more detailed understanding of cancer and cancer-related processes by selecting highly relevant genes that are not just correlating with but actively influencing the outcome of a patient. Furthermore, putting prognostic signatures into the context of pathways and network neighborhood may provide crucial information to move from biomarkers to targets, whose modulation will influence outcome.

5 UNIVERSAL CANCER SIGNATURE

Single gene markers express several drawbacks and cannot reliably be used for cancer outcome prediction. For this reason several approaches use selections of predictive genes. Nevertheless, the overlap of these gene sets tends to be zero, even for the same outcome task.

In this chapter signatures obtained by original publications, greedy methods as well as NetRank are analyzed regarding their similarity. The overlapping genes of NetRank are furthermore analyzed for their biological context and meaningfulness. The similarity of signatures between several cancer types direct towards a Universal Cancer Signature. This notion is investigated and several possible combinations are tested.

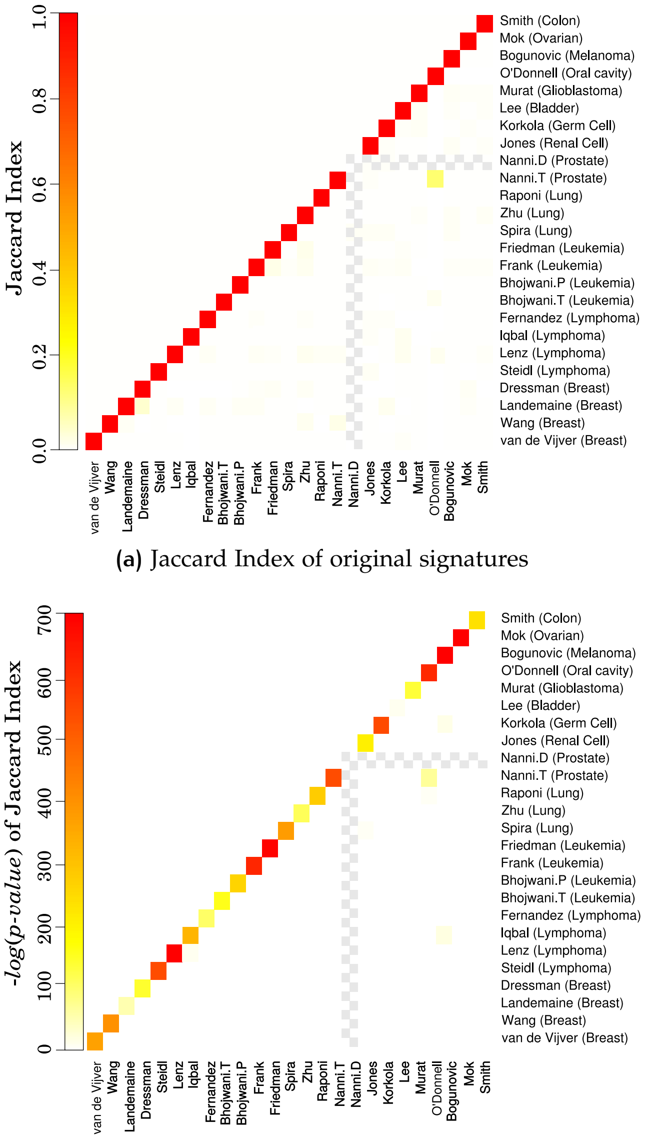
5.1 SIGNATURE OVERLAP – A SIGN FOR UNIVERSAL CANCER SIGNATURE

Elevated levels of single gene markers are often not only found in a single type of cancer. This indicates that several cancers share the same mechanism of survival, tumor growth and invasion. That cancers share the same mechanisms during development has been discussed before by Hanahan and Weinberg [79]. They discussed the idea of common cancer signals and defined the Hallmarks of cancer in which distinctive biological processes are responsible for tumor growth, invasion and metastatic dissemination.

By investigating signatures obtained by NetRank, a significant similarity between predictive genes was observed. In the following section, the similarity of signatures of the original studies with signatures obtained by NetRank is compared. Secondly, the effect of the similarity on the prediction accuracy is assessed and finally the overlapping genes are analyzed regarding their Gene Ontology terms.

5.1.1 Signatures from original cancer studies do not overlap – even for the same cancer type

The analysis of the signature similarity was performed on a benchmark dataset consisting of 25 literature derived datasets. The datasets are quite diverse as they cover 13 different types of cancer including breast and prostate cancer, leukemia and lymphoma. They differ in the outcome variable (diagnosis, response to treatment, progression



(a) Jaccard Index of original signatures

(b) $-log(p-value)$ of Jaccard Index

Figure 14: Signature similarity between original publications as a) Jaccard Index and b) -(log)((p-value)) of Jaccard Index. A value of 0 (white) indicates no overlap and 1 a strong overlap between the signatures. Only the signatures of O'Donnell *et al.* and Nanni (T) *et al.* show a slight similarity as they share 7 genes. Note: No signature was published for Nanni *et al.* (D).

or subtyping), the size of signatures (from 1 to 357 genes), and their prediction accuracy (according to the authors from 60% to 100%), as well as the methods employed. A summary of the datasets can be found in Table 4.

For the comparison of signature similarity, the signatures as stated in the original publications are taken. Note that, Nanni *et al.* did not publish a signature for their diagnosis of prostate cancer dataset. The pancreas cancer in-house dataset [196] is also not included.

No significant overlap for original signatures

To investigate the overlap between signatures the pairwise similarity of signatures represented as Jaccard Index is plotted against each other. In Figure 14a, a value of 0 (white) indicates no overlap and 1 a strong overlap between the signatures. As expected Figure 14a shows no significant overlap between the signatures published by the authors. Even for studies investigating the same cancer and outcome variable, no overlapping genes exist. Exceptions are Nanni *et al.* (T) [127] and O'Donnell *et al.* [132], which share seven transcription factor out of 205 genes. We also computed p-values of Jaccard indices using Fisher's exact test, which shows the probability of an overlap between the signatures by chance (see section 3.4). In Figure 14b a value of 0 (white) indicates that the signature could have occurred by chance. The different colors in the diagonal arise due to the different sizes of the signatures. The diagonal of Lee *et al.* [104] is almost white as the signature consists of only one gene and could have clearly arisen by chance. The figure also indicates the small overlap between Nanni *et al.* (prostate cancer) and O'Donnell *et al.* (oral cavity).

This discovery supports the finding of Ein-Dor and co-workers as they investigated two breast cancer signatures regarding their similarity and found an overlap of almost zero [55].

5.1.2 Greedy techniques do not trigger signature similarity

One reason for the lack of overlap might be the variety of techniques applied by the authors of the studies. None of the authors used exactly the same methods or thresholds as the other authors. We therefore aimed at comparing the resulting signatures while applying the same analysis to all datasets. Thus, we applied two standard greedy techniques – *t*-test and foldchange – on each dataset. An explanation of the methods can be found in section 3.5. These techniques do not take into account any relationships between genes; they solemnly analyze the relationship of the expression of a gene with the outcome variable. Hence, genes chosen by these techniques often tend to correlate in their expression; therefore no information gain is achieved by using several similar genes. Nevertheless, these standard techniques are applied to the 25 datasets and the resulting signatures are compared regarding their similarity.

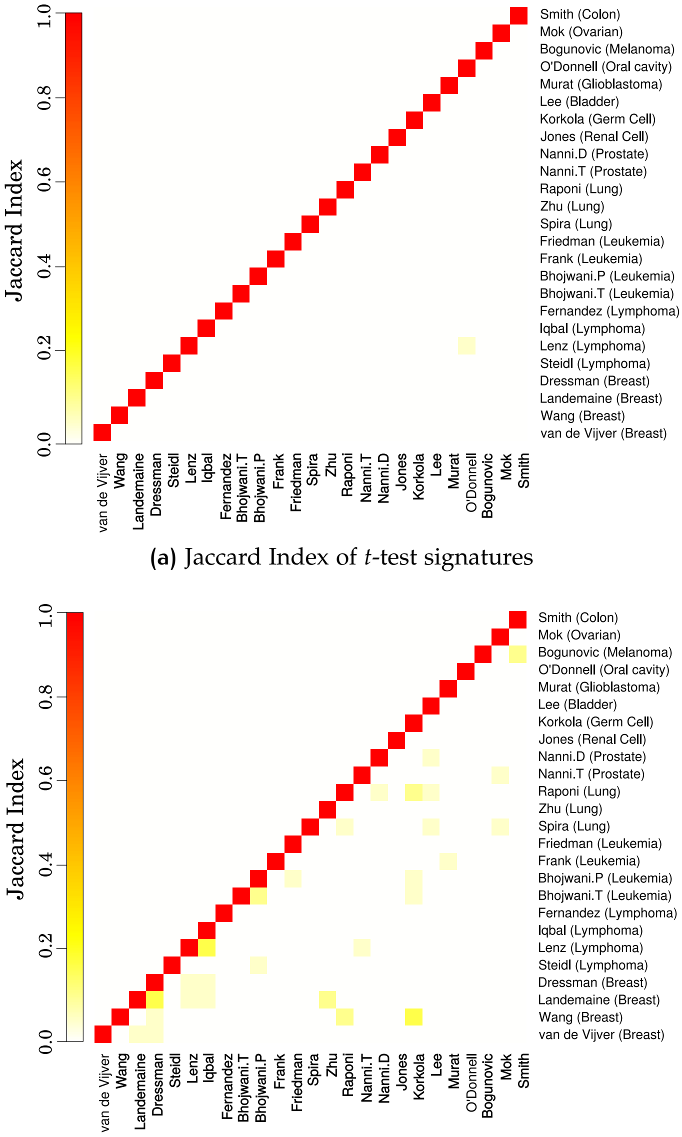
No significant overlap for signatures obtained by greedy methods

As the Figure 15a and Figure 15b show, both *t*-test and foldchange fail to construct stable signatures, with foldchange expressing small similarities between certain signatures. This supports the finding of Shi *et al.* [160, 161]. They showed in a comprehensive analysis that foldchange results in more reproducible gene list than the *t*-test.

As conclusion neither the authors themselves nor the same analysis methods are able to improve signature overlap to create stable predictive signatures. In the next section we investigate the signatures obtained by NetRank regarding their similarity.

5.1.3 NetRank signatures do strongly overlap

To compute stable signatures, dependencies between genes have to be considered and the data has to be aggregated. In former work, a network-based outcome prediction approach – NetRank – was devel-



(a) Jaccard Index of *t*-test signatures

(b) Jaccard Index of foldchange signatures

Figure 15: Signature similarity of a) *t*-test and b) foldchange represented as Jaccard Index. A value of 0 (white) indicates no overlap and 1 (red) a strong overlap between the signatures. Both *t*-test and foldchange fail to construct stable signatures, with foldchange having a small similarity between certain signatures.

oped that is able to cover these dependencies. In chapter 4 the algorithm was applied on several types of cancer gene expression data using 2 types of networks, a transcription factor network (Transfac [119]) and a protein-protein interaction network (HPRD [146]). The algorithm was not just having a better accuracy than random (13% improvement), it also boosted the prediction accuracies compared to classical methods like *t*-test and foldchange by 5%.

NetRank boosts signature similarity

In order to investigate the overlap of the resulting signatures, the pairwise similarity of signatures represented as Jaccard Index is again plotted against each other. Figure 16 and Figure 17 show the similarity of signatures obtained by NetRank applied on HPRD and Transfac on the 25 datasets. The overlap is statistically significant and different datasets covering either the same cancer type, as well as different cancer types overlap well.

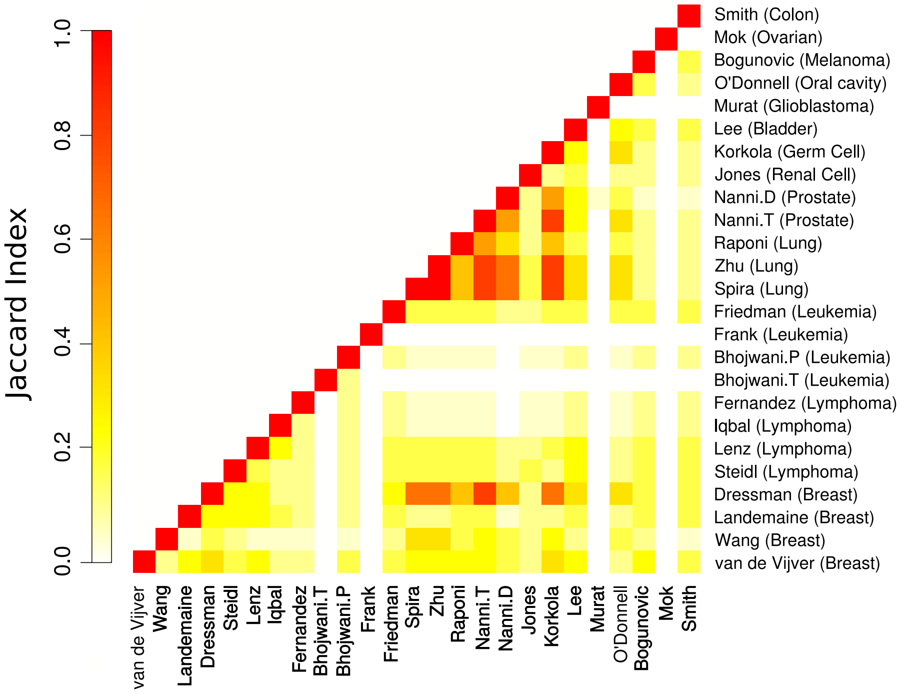


Figure 16: The NetRank signatures based on HPRD overlap nicely for the same type of cancer as well as for different cancer types. For example the signatures of lung cancer of Zhu *et al.* and Spira *et al.* overlap in all genes. HPRD-based NetRank signatures share in average 2 genes. In the figure a value of 0 (white) indicates no overlap and 1 a strong overlap between the signatures.

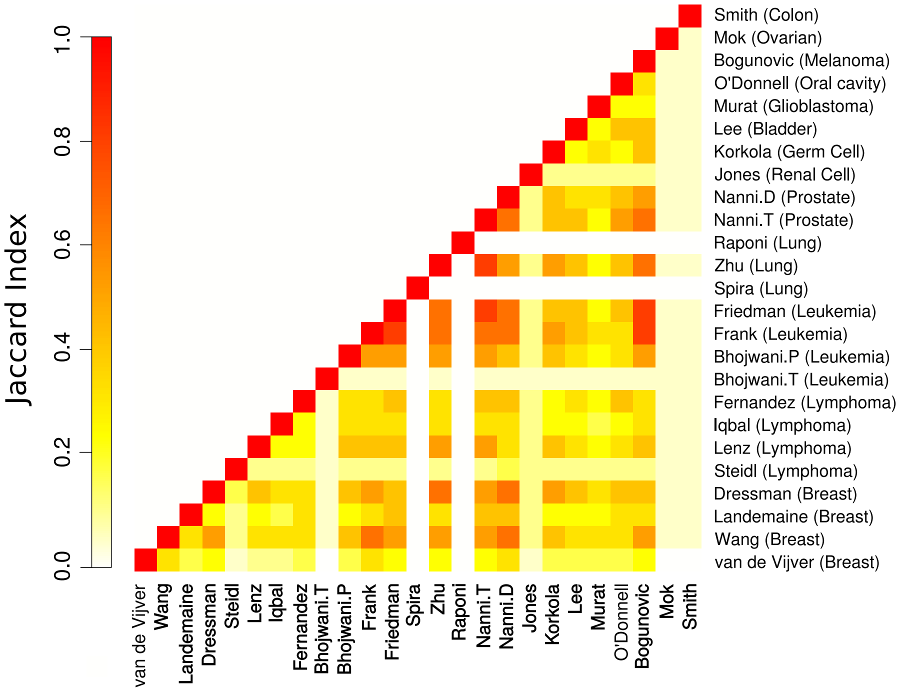


Figure 17: The signature obtained by NetRank based on Transfac overlap strongly. All signatures share in average 3.2 genes. Only for two (Spira *et al.*, Raponi *et al.*) out of 25 dataset there are no overlapping genes. A value of 0 (white) indicates no overlap and 1 a strong overlap between the signatures.

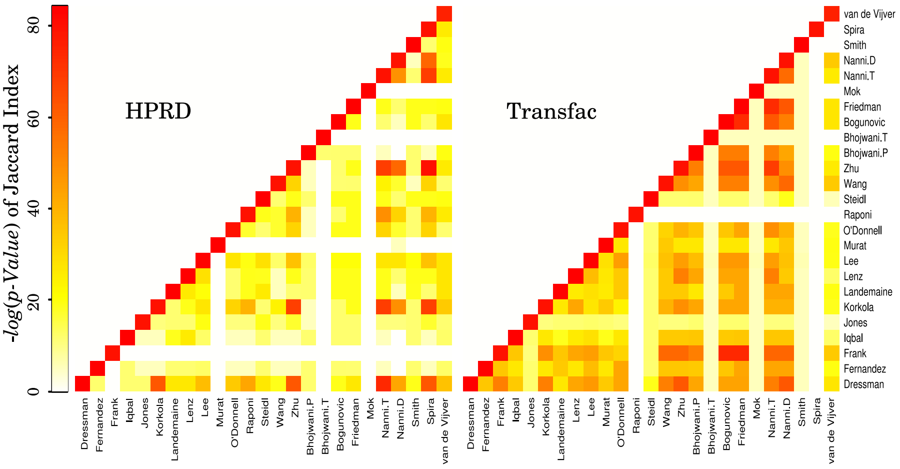


Figure 18: The heatmap of the $-log(p-value)$ of the Jaccard Index of Transfac and HPRD-based NetRank genes. Statistically significant signature similarities (not by chance) are represented by higher values and red colors. The signatures of NetRank with HPRD and Transfac are significantly overlapping for many datasets.

100% signature similarity for two lung cancer datasets

For example the HPRD-based lung cancer signatures for the Zhu *et al.* [206] and Spira *et al.* [168] datasets consist out of the same 10 genes. A similar high overlap for leukemia can be observed in the Transfac-based signature of Friedman *et al.* [65] and Frank *et al.* [64] that overlap in nine out of ten genes. Moreover, signatures do overlap which are not based on the same type of cancer. The signatures of Nanni *et al.* [127] and Korkola *et al.* [96] share 6 out of 10 genes, which might be surprising as they investigate prostate cancer and germ cell cancer outcome, respectively.

For the NetRank signatures we also computed p-values of Jaccard indices using Fisher's exact test, which shows the probability of an overlap between the signatures by chance. The $-log(p-value)$ of each pairwise similarity is plotted in Figure 18. The signatures of NetRank using either physical interaction (HPRD) or regulatory information (Transfac) are significantly overlapping for many datasets.

Taking all these facts together, NetRank signatures based on Transfac and HPRD share in average 3.2 and 2 genes, respectively. The question arises whether this overlap is biologically meaningful, or not. We hypothesize that the unique signature genes are cancer specific, whereas the overlapping genes represent general cancer signals.

The general principle of cancer development and growth has been investigated since several decades. A literature search in GOPubmed[1] yields 3 million publications annotated with the MESH term 'neoplasms', making it to one of the mostly used disease terms. We therefore assume, that general cancer genes are well-studied and known in literature. Hence, in the following section the most overlapping genes are investigated regarding their biological relation and reference to cancer growth and development in literature.

[1] http://gopubmed.org

5.2 NETRANK GENES ARE HIGHLY CONNECTED AND CONFIRMED IN LITERATURE

By using a network-based approach, signatures for 25 literature derived datasets are created. For a detailed dataset description see Table 4. While analyzing these datasets, a huge overlap between the signatures was discovered. Table 11 and Table 12 show the signatures obtained for each of the datasets applying NetRank on Transfac and HPRD. In the Transfac-based signatures the specificity proteins 1 and 3 (SP1 and SP3) are selected in more than 20 and 18 signatures, respectively. This is biologically reasonable as they are highly connected in the network and regulate a huge amount of different proteins. A similar picture can be drawn for the HPRD-based signatures. The following two sections describe in detail some of the highly connected proteins and shed light on the biological processes involved.

5.2.1 Transfac-based NetRank genes

Transcription factors are especially suitable for cancer outcome prediction. They act as signal transducers in signaling cascades by transferring information to genes by activating or repressing their expression. This key function renders transcription factors highly informative for signatures in malignant diseases. Table 13 shows the 13 genes which occur more than 3 times in the Transfac-based NetRank signatures. They are all master regulators with over 60 targets and as such involved in many biological processes. At the same time, these genes are known in literature as biomarkers for the majority of the 13 cancers of our datasets. Furthermore, all of them have been documented independently in literature as possible cancer biomarkers for at least two different cancers. The following paragraphs describe the biological function of six transcription factor (SP1, SP3, JUN, FOS, NFKB1 and Rela).

SP1 and SP3 were selected in more than 18 signatures

The specificity protein 1 (SP1) and 3 (SP3) are the most often chosen proteins. They have a predictive power in 22 and 20 out of 25 datasets and each of them interact with or regulate 421 and 164 other proteins, respectively. SP1 and SP3 play an important role in regulating cell proliferation and have high levels of expression in several cancer types, e.g., breast, pancreatic, colon and prostate cancer [1, 39, 81, 163]. Specifically, it was experimentally demonstrated that SP1 regulates proliferation and lipogenesis processes in colon, prostate, and breast cancer cells [113]. Therefore, drugs decreasing the expression level of the SP family would also reduce the level of genes related with cancer progression [38].

Jun and Fos form the AP1 complex

The Jun-like transcription factor (JUN) and the FOS protein has been chosen as predictive 17 and 14 times, respectively. They interact in the Transfac network with 154 and 138 different genes. JUN and

Table 11: Signatures for each dataset using NetRank on Transfac. The colours are indicating the frequency of a gene in a signature.

Author	Cancer	Acc	Signatures									
Bhojwani et al.	Leukemia	72	SP1	TP53	JUN	FOS	SP3	NFKB1	IRX1	CREB1	ETS1	IGFBP7
		64	SP1	IGJ	POU4F1	SOCS2	PRKCZ	SH3BP5	PECAM1	BCL2L11	MUC4	PHACTR3
Bogunovic et al.	Melanoma	70	SP1	SP3	JUN	FOS	NFKB1	CREB1	TP53	JUND	JUNB	TFAP2A
Dressman et al.	Breast	64	SP1	SP3	FOS	TP53	JUN	EGR1	CEBPB	NFKB1	RELA	STAT1
Fernàndez et al.	Lymphoma	81	SP1	SP3	TP53	HDGFRP3	RNGTT	FOS	HNF4A	JUN	PON2	FARP1
Frank et al.k	Leukemia	75	SP1	SP3	JUN	FOS	NFKB1	TP53	STAT1	CREB1	TFAP2A	JUND
Friedman et al.	Leukemia	55	SP1	SP3	JUN	FOS	HNF4A	CREB1	NFKB1	TP53	TFAP2A	JUND
Iqbal et al.	Lymphoma	82	SP1	SP3	CR2	NFKB1	CCL21	ALPK2	XKR4	TP53	FOS	CLU
Jones et al.	Renal Cell	97	SP1	SP3	MAOB	HSPB8	NR1H4	MINA	RARRES2	KBTBD11	SIRPA	USH1C
Korkola et al.	Germ Cell	71	SP1	SP3	TP53	NFKB1	CREB1	RELA	JUN	NFYA	STAT1	KLF4
Landemaine et al.	Breast	99	SP1	SP3	TP53	SFTPA2	SCGB2A2	HNF4A	JUN	FDCSP	TFAP2A	SFTPC
Lee et al.	Bladder	76	SP1	JUND	SP3	JUN	TBC1D1	TP53	FOS	ZNF2	EGR1	ZNF564
Lenz et al.	Lymphoma	65	SP1	CREB1	TP53	SP3	MYC	NFKB1	JUN	SPI1	EGR1	GABPA
Mok et al.	Ovarian	65	SP1	ITLN1	ABCA8	BCHE	ADH1B	HBB	OGN	AOX1	TCEAL2	MECOM
Murat et al.	Glioblastoma	59	SP1	JUN	EGFR	TP53	SP3	STAT1	NPR3	SEPT6	FAM89A	DDX1
Nanni et al.	Prostate	70	SP1	SP3	HNF4A	JUN	TP53	NFKB1	FOS	CREB1	TFAP2A	EGR1
		99	SP1	SP3	TP53	TFAP2A	NFKB1	JUN	STAT1	CEBPB	HNF4A	FOS
O'Donnell et al.	Oral cavity	81	SP1	SP3	JUN	EGR1	C9orf3	FOS	HNF4A	TP53	NEB	TTN
Raponi et al.	Lung	59	IGL@	KRT6A	SFN	KRT5	S100A8	IGKC	COL1A2	GPNMB	LOC 100130100	AKR1C1
Smith et al.	Colon	71	SP1	FABP4	SPP1	MAGEA6	PRAC	ADIPOQ	MAGEA3	CYP1B1	WDR72	MGP
Spira et al.	Lung	69	CCDC81	IL1R2	DEFB1	FGFR3	EPAS1	NR4A3	RGS2	DMBT1	ADH1C	CD55
Steidl et al.	Lymphoma	74	SP1	SP3	CLPS	HSPA5	AGRN	CEBPB	EPB41	PTX3	KTN1	RPL22
Wang et al.	Breast	67	SP1	SP3	MMP1	TP53	STAT1	NFKB1	TFAP2A	FOS	IRF1	JUN
Zhu et al.	Lung	58	SP1	SP3	JUN	TP53	FOS	TFAP2A	NFKB1	EGR1	RELA	CREB1
Vijver et al.	Breast	68	SP3	FOS	CEBPA	USF1	JUN	TFAP2A	POU2F1	STAT1	TK1	BARX2

Frequency in signatures: > 20 > 18 > 16 > 14 > 12 > 10 > 8 > 6 > 4 > 2

Table 12: Signatures for each dataset using NetRank on HPRD. The colours are indicating the frequency of a gene in a signature.

Author	Cancer	Acc	Signatures									
Bhojwani et al.	Leukemia	70	YWHAG	ATXN1	IRX1	IGJ	IGFBP7	AGPS	KANK2	MUC4	NPY	TGFBR1
		63	IGJ	POU4F1	SOCS2	PRKCZ	PECAM1	SH3BP5	BCL2L11	MUC4	PHACTR3	AKAP12
Bogunovic et al.	Melanoma	64	ATXN1	YWHAG	NR3C1	TRIP13	EWSR1	SDK1	RXRG	FOXN3	CDK1	SVIL
Dressman et al.	Breast	55	YWHAG	ATXN1	EWSR1	CREBBP	PRKACA	PRKCA	CSNK2A1	SMAD2	ATN1	SRC
Fernàndez et al.	Lymphoma	80	HDGFRP3	ATXN1	SOX11	YWHAG	CNR1	CNN3	PON2	ZNF711	DEF8	CRY1
Frank et al.	Leukemia	70	PRTN3	PRG2	TPSAB1	TPSB2	ELANE	CTSG	AZU1	IFIT1	VNN1	MPO
Friedman et al.	Leukemia	57	ATXN1	YWHAG	KIAA1210	PRKACA	CCDC24	MAN2B1	B4GALT2	PDIA5	CFP	EWSR1
Iqbal et al.	Lymphoma	81	CR2	YWHAG	CCL21	ALPK2	XKR4	CLU	FDCSP	SOX8	PCDHA6	ATXN1
Jones et al.	Renal Cell	97	ATXN1	EWSR1	NDRG1	HSPB8	SIRPA	MAOB	NR1H4	MINA	RARRES2	KBTBD11
Korkola et al.	Germ Cell	70	EWSR1	ATXN1	PRKACA	PRKCA	CSNK2A1	SMAD2	CREBBP	SRC	SMAD3	EP300
Landemaine et al.	Breast	100	YWHAG	ATXN1	SFTPA2	EWSR1	SCGB2A2	SCGB1D2	ESR1	FDCSP	MAGEA1	ATN1
Lee et al.	Bladder	73	ATXN1	YWHAG	TBC1D1	NDRG1	EWSR1	SKIL	ZNF2	ZNF564	SMAD2	PRKCA
Lenz et al.	Lymphoma	63	YWHAG	ATXN1	FOLR1	EWSR1	FDCSP	MS4A4A	CR2	NAA15	PRKCA	TMEM119
Mok et al.	Ovarian	64	ITLN1	ABCA8	BCHE	HBB	ADH1B	OGN	AOX1	TCEAL2	MECOM	ALDH1A2
Murat et al.	Glioblastoma	58	NPR3	SEPT6	FAM89A	EGFR	DDX1	AHCYL1	SYT6	TUSC3	SYTL4	BBS2
Nanni et al.	Prostate	72	PRKACA	PRKCA	EWSR1	ATXN1	CSNK2A1	CREBBP	SMAD3	ATN1	SRC	SMAD2
		100	EWSR1	PRKACA	PRKCA	EGFR	SMAD3	CREBBP	SRC	SMAD2	NDRG1	MAPK1
O'Donnell et al.	Oral cavity	81	ATXN1	EWSR1	C7	SVIL	CSNK2A1	TRAF2	PRKACA	SKIL	SMAD2	CHD3
Raponi et al.	Lung	64	ATXN1	EWSR1	PRKACA	ATN1	PTN	PRKCA	YWHAB	KAT5	CREBBP	SMAD3
Smith et al.	Colon	72	YWHAG	ATXN1	FABP4	SPP1	MAGEA6	PRAC	EWSR1	WDR72	MAGEA3	CYP1B1
Spira et al.	Lung	72	ATXN1	EWSR1	PRKCA	PRKACA	SMAD2	CREBBP	CSNK2A1	SMAD3	SRC	NDRG1
Steidl et al.	Lymphoma	73	YWHAG	XIST	ATXN1	RARA	EWSR1	PLTP	TMSB15A	RPS4Y1	NDRG1	ATN1
Wang et al.	Breast	70	CDK1	NDRG1	CSNK2A1	ATXN1	SMAD3	SMURF2	CD74	EEF1A1	TP53	PRKACA
Zhu et al.	Lung	53	ATXN1	PRKACA	CSNK2A1	CREBBP	NDRG1	SRC	PRKCA	EWSR1	SMAD3	SMAD2
Vijver et al.	Breast	66	YWHAG	ESR1	CREBBP	ATXN1	EP300	EWSR1	CDK1	GRB2	PRKCA	TGFBR1

Frequency in signatures: > 18 > 16 > 14 > 12 > 10 > 8 > 6 > 4 > 2

Table 13: Top NetRank genes are part of many signatures, are transcriptional master regulators, and are independently known in literature as cancer biomarkers (analysis based on 13 cancer types included in our analysis).

Gene	Signature in # of datasets	# of targets in Transfac	# of 13 cancers confirmed in literature
SP1	22	421	7
SP3	20	164	1
JUN	17	154	8
TP53	17	81	8
FOS	14	138	3
NFKB1	12	98	5
TFAP2A	9	76	6
CREB1	8	87	2
STAT1	7	61	4
EGR1	6	66	3
HNF4A	6	68	2
JUND	4	104	3
CEBPB	3	64	4
RELA	3	71	3

FOS families form the Activator protein 1 (AP-1) complex that controls the expression of genes, regulating several cell processes, such as proliferation, migration, survival, differentiation as well as apoptosis [112]. Overexpression of JUN was demonstrated in lung, leukemia, glioblastoma, melanoma and prostate cancer cells [111, 135, 136, 148, 164]. However, FOS shows overexpression in prostate cancer cells [136] and lower expression level in ovarian cancer [118]. The function of AP-1 changes based on its dimer composition (i.e., various combinations of JUN and FOS family proteins) and cell content [6]. Although some AP-1 complexes (e.g., c-JUN and FOS) may activate oncogenic target genes, some others (e.g., JUN and FRA1) may activate inhibitory pathways [159]. Drugs targeting JUN and FOS family members consequently disrupt the AP-1 complex formation and would provide an improvement in cancer treatments.

NFKB1 and Rela form the NF-kB complex

The nuclear factor kappa-light-chain-enhancer of activated B cells 1 (NFKB1) has been selected in 12 out of 25 signatures as predictive. It interacts with 98 other genes, including RELA, which is included in 3 out of 25 signatures. NFKB1 forms together with RELA the NF-kB complex that controls the expression of several genes regulating immune responses, cell cycle, proliferation, and apoptosis [50]. The promoting role of NF-kB in cancer progression has been extensively investigated in the last decade. The studies proved that it is either a tumor-suppressor or tumor-promoting oncogene based on mutations occurring in the upstream region of NF-kB. A recent study has shown that NF-kB has a tumor-suppressor function in ovarian cancer [203]. High activation of NF-kB has been detected in hepatocellular carcinoma, melanoma, leukemia, and breast cancer [49, 72, 145, 173]. Another study proved that the activation of NF-kB is mostly related with the inflammatory environment of tumor progression [92, 183]. Drugs disrupting function of NF-kB would lead to adverse effects for normal tissues, hence upstream interactors of NF-kB could be more effective drug targets for cancer treatments [92].

Transfac-based NetRank genes are highly interconnected

The Figure 19 shows the interconnectivity of 14 Transfac genes that are selected more than 3 times as predictive in one of the 25 datasets. As the figure shows, the genes are well interconnected, creating a highly connected cluster of genes regulating each other. The only exception is the CCAAT/enhancer binding protein, beta (CEBPB) which needs the matrix metallopeptidase 1 (MMP1) to be connected to the subgraph. MMP1 is known to be involved in cancer metastasis and is related to bladder [201] and oral cancer [57].

5.2.2 HPRD-based NetRank genes

Out of 25 signatures of NetRank based on the HPRD network, 23 genes got selected more than twice as predictive. These genes, their connectivity as well as their literature occurrence are listed in Table 14. These genes are generally highly connected to other genes in the HPRD network. Some of them are involved in the transcription of other genes or physically interact with master players of cancer development. In the following paragraphs, the functions of seven genes (PRKCA, SRC, TGFBR1, EGFR, ESR1, SMAD2, SMAD3), which are well-known targets in different cancer therapies, are summarized.

Protein kinase C alpha (PRKCA)

Protein kinase C alpha (PRKCA) is a member of the family of serine/threonine protein kinases and it regulates cardiac contractility, atherogenesis, and cancer [95]. PRKCA is activated by the tumor promoter phorbol ester [32] and its function is related with cell proliferation, apoptosis and cell motility in cancer. It is over- and under-expressed in different cancer types [95], hence it became a target for various cancer treatments. However several drugs targeting PRKCA activity have not provided a new adequate cancer therapy yet [121].

Proto-oncogene tyrosine-protein kinase Src

Proto-oncogene tyrosine-protein kinase Src (SRC) is discovered as first human oncogene and encodes a tyrosine kinase protein that regulates differentiation, survival, angiogenesis and motility in cancer. Significant activity of SRC was detected in colon, breast, lung, pancreatic and prostate cancers [194]. SRC interacts with several kinases, e.g. EGFR, VEGFR, PDGFR, CSF1R, HER2, HER3 and IGF1R. There are many ongoing studies and clinical trials that aim to inhibit activity of SRC in various cancers [147]. Due to synergistic effect between SRC and EGFR in tumor development [88, 114], new cancer therapies should target the inhibition of both genes concurrently [194].

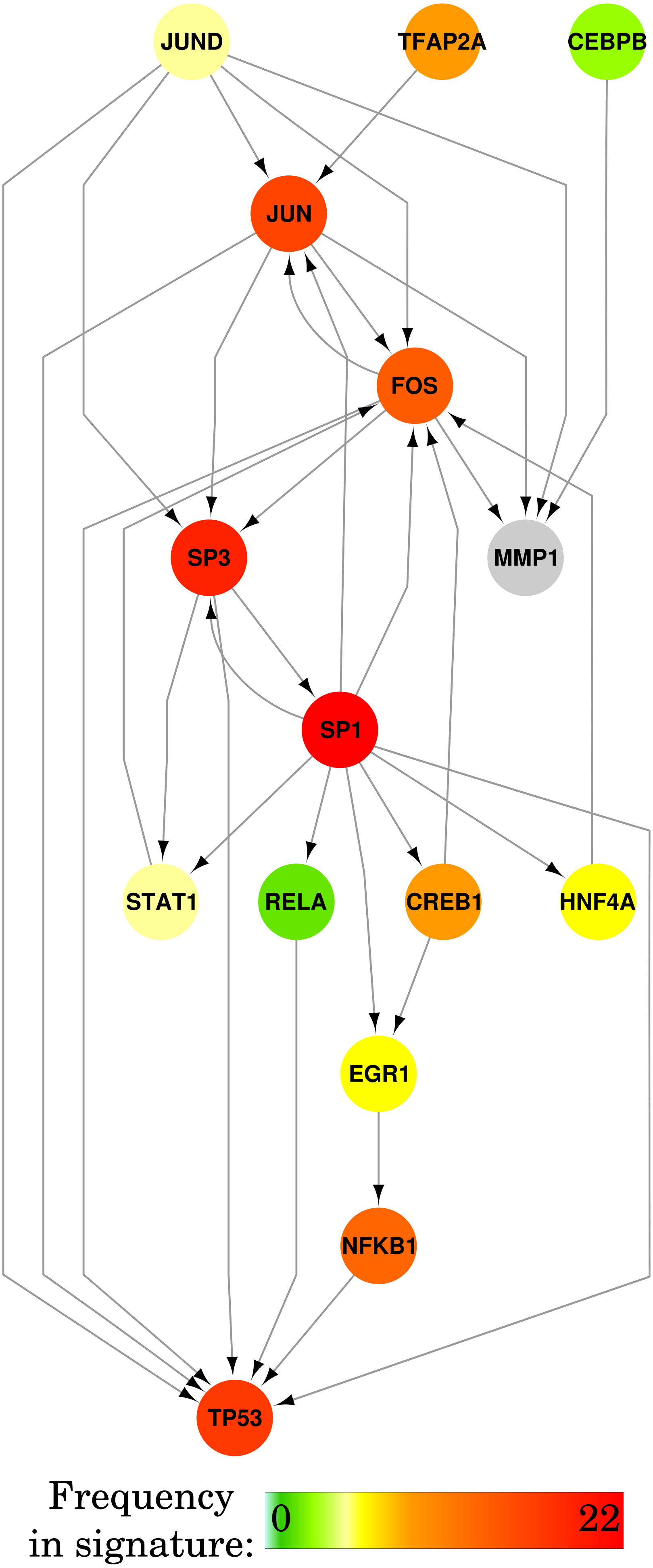


Figure 19: The figure shows the connectivity of the most chosen Transfac genes. The genes are generally well interconnected. Only one intermediate gene is needed (grey circle; MMP1) to connect all genes with each other. The color indicates the number of occurrences of a gene in the resulting signatures.

Table 14: Top NetRank genes are part of many signatures, are independently known in literature as cancer biomarkers (analysis based on 13 cancer types included in our analysis). All listed genes occur more than once in the results.

Gene	Signature in # of datasets	# of targets in HPRD	# of 13 cancers confirmed in literature
ATXN1	20	159	2
EWSR1	17	117	1
YWHAG	12	247	1
PRKCA	10	172	6
PRKACA	10	145	3
SMAD2	8	166	4
CREBBP	8	198	4
SMAD3	7	182	4
NDRG1	7	61	5
CSNK2A1	7	168	5
SRC	6	208	7
ATN1	5	88	-
FDCSP	3	0	-
CDK1	3	119	5
IGJ	2	2	1
MUC4	2	2	3
SKIL	2	73	-
SVIL	2	63	1
TGFBR1	2	154	1
CR2	2	10	3
EGFR	2	161	6
EP300	2	209	4
ESR1	2	188	4

Transforming growth factor, beta receptor I (TGFBR1) induces SMAD2 and SMAD3

Transforming growth factor, beta receptor I (TGFBR1) is a member of the transforming growth factor beta family of cytokines that have roles in proliferation, differentiation, adhesion, migration and apoptosis. SMAD2 and SMAD3 are signal transducer proteins that regulate several processes such as proliferation, differentiation and apoptosis. Activated TGFBR1 induces SMAD signaling by phosphorylation of SMAD2 and SMAD3, then SMAD4 translocates into the nucleus and starts transcription of target genes. Therefore, the SMAD pathway is important for the growth inhibitory action of TGFBR1 [43]. Loss of expression of TGFBR1 is frequently observed in gastric, colon, and bladder cancer [107]. Mutations on SMAD2 (frequent) and SMAD3 (less frequent) are detected in colon cancer [60]. Therapy options of TGF-β signaling could be either the inhibition of TGFBR1 by small molecules or disrupting SMAD family interactions by peptide aptamers [43].

Epidermal growth factor receptor (EGFR)

The epidermal growth factor receptor (EGFR) is a cell-surface receptor that has important roles in proliferation, cell cycle and migration processes [204]. Mutation or overexpression of this receptor may cause the development of lung, colon, breast, ovarian, glioblastoma and pancreatic cancers. Therefore EGFR is an important target for several cancer therapies. It has also a high prognostic power for various cancer types [14, 33, 130].

Estrogen receptor 1 (ESR1)

ESR1 encodes the estrogen receptor 1 that is activated by the estrogen hormone. The activated receptor translocates into the nucleus and initiates transcription of different genes. This receptor is active in round 60-70% of breast cancer tumors [108]. Hence, several breast cancer therapies try to decrease estrogen levels or to interrupt the estrogen receptor signaling pathway. Expression of ESR1 could be used as biomarker to predict survival time of ovarian and endometrial cancer patients [29].

HPRD-based NetRank genes are well interconnected

The Figure 20 shows the in-between connectivity of the top 23 HPRD genes. These genes got selected at least 3 times as predictive in one of the 25 datasets. As the figure shows, the proteins are well interconnected with each other. In contrast to the Transfac genes, there is a core of proteins that are more connected to each other and some peripheral proteins. The peripheral genes are IGJ, ATXN1 and MUC4.

The proteins immunoglobin J (IgJ) and ataxin 1 (ATXN1) are connected over FYN – an oncogene related to SRC, FGR, YES – to the core proteins. In addition needs IgJ one more connector, which is CD79A – an immunoglobulin-associated molecule, also known as IgA. Mucin 4 (MUC4), a cell surface associated protein, is the most distant gene and needs 3 intermediate proteins (ERBB2, CAV1, ABL1) to be connected to the core proteins.

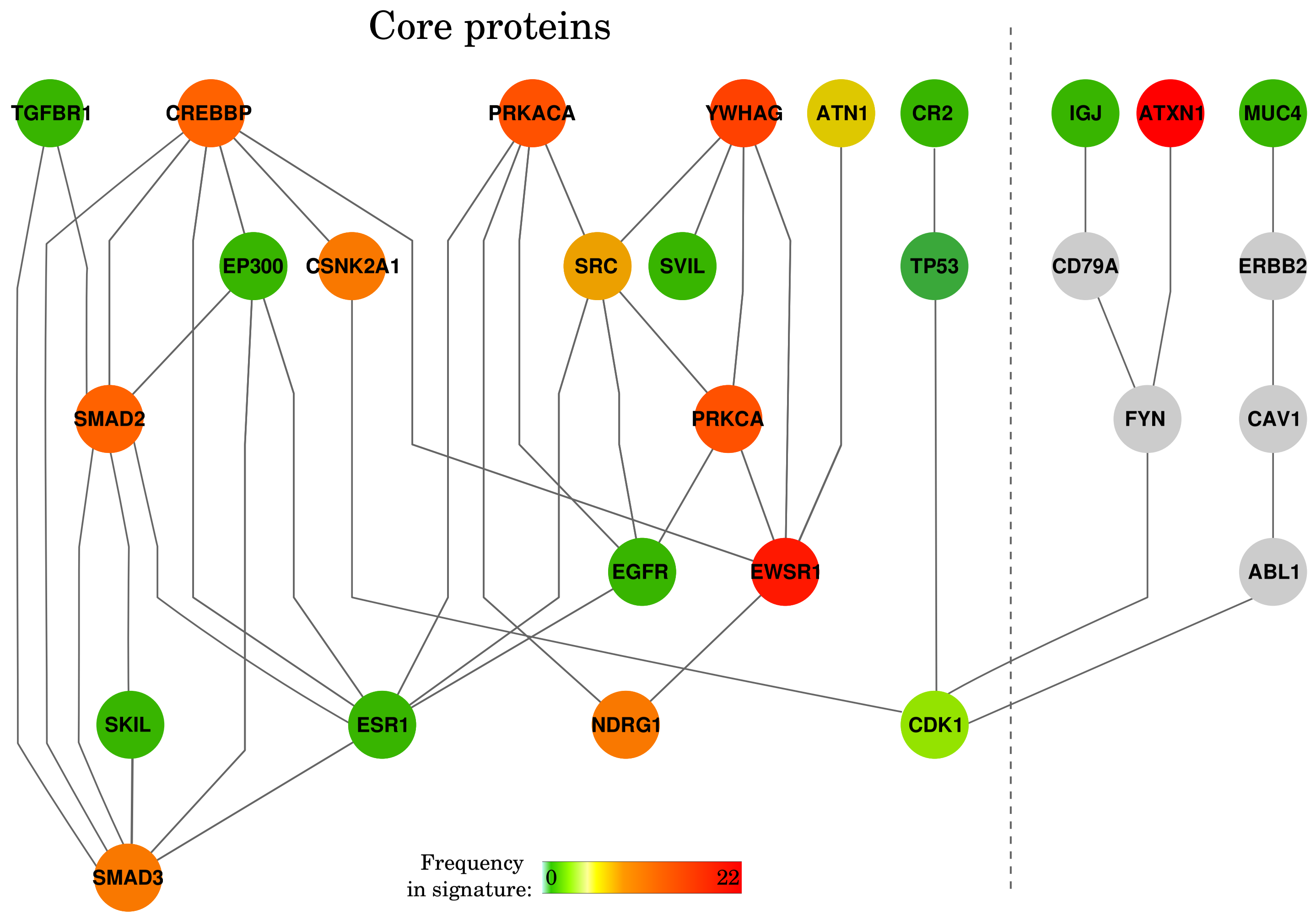


Figure 20: The figure shows the connectivity of the most chosen HPRD genes. The genes are interconnected. Only five extra genes are needed (grey circles) to connect all genes. Nevertheless there are core proteins that are better in-between connected and peripheral proteins. The proteins in the grey circles separate the two types of proteins. The node color indicates the number of occurrences of a gene in the resulting signatures.

5.3 HALLMARKS OF CANCER

In the previous section the similarity of signatures obtained by the authors, greedy methods and NetRank was investigated. We showed, that the original as well as the *t*-test and foldchange signatures do not overlap, whereas signatures retrieved by NetRank display a strong overlap. The overlapping genes are highly connected in Transfac and HPRD, respectively. Literature analysis showed, that these genes are related to cancer and have been mentioned already as biomarker for outcome prediction.

In the following section the overlapping Transfac and HPRD genes are analyzed regarding the 'Hallmarks' of cancer [79, 80].

Hallmarks of cancer summarize the underlying principles of cancer development

Hanahan and Weinberg claim that the complexity of cancer can be reduced to a small number of underlying principles [79]. They argue that all cancers share eight common 'Hallmarks', which govern the transformation from normal cells into cancer cells. Assuming these signature genes could form a possible Universal Cancer Signature, they should be involved in at least some of the Hallmarks suggested. We mapped the Gene Ontology annotations to the different Hallmarks of Cancer and could find associations to the following 8 Hallmarks[2]:

1. cancer cells stimulate their own growth;
2. they resist inhibitory signals that might otherwise stop their growth;
3. they resist their own programmed cell death (apoptosis);
4. they stimulate the growth of blood vessels to supply nutrients to tumors (angiogenesis);
5. they can multiply forever;
6. they invade local tissue and spread to distant sites (metastasis);
7. they are able to evade immune destruction;
8. they are able to reprogram their glucose metabolism.

An active immune system prevents cancer

The first six Hallmarks are general principles already published in 2000 [80]. The emerging Hallmarks – the evasion of the immune system as well as the reprogramming of the glucose metabolism – have been published recently. In preclinical studies it has been shown, that an active immune system continuously recognizes and eliminates the majority of cancer cells before they establish themselves and form a

[2] Images taken from original publication

Table 15: The Hallmarks of cancer are well represented by the signature genes based on the Transfac network. The Hallmarks of metabolic reprogramming and metastasis formation are sparsely covered by only HNF4A, whereas the Hallmark of resistance to inhibitory signals is covered by eight out of the 14 proteins and the remaining Hallmarks are at least represented by three genes. Except JunD – which only represents the Hallmark of blood vessel growth – all genes cover more than one Hallmark.

Gene	↱	⊘	✝	⌂	∞	👻	☸	⚛
SP1	✓	✓		✓				
SP3		✓						
JUN		✓	✓	✓	✓			
TP53	✓	✓	✓		✓		✓	
FOS		✓		✓				
NFKB1			✓	✓			✓	
TFAP2A	✓							
CREB1			✓					
STAT1		✓	✓					
EGR1	✓			✓				
HNF4A	✓	✓			✓	✓		✓
JUND				✓				
CEBPB			✓				✓	
RELA		✓	✓				✓	

tumor mass. Clinical examples also support this finding and demonstrate that colorectal and ovarian cancer patients with an increased immune response have a better prognosis than patients with a reduced immune response [79, 182].

Cancer cells are able to perform aerobic glycolysis

Due to the Warburg effect, cancer cells are able to reprogram their glucose metabolism and convert high amounts of glucose to lactate even in the presence of oxygen (aerobic glycolysis) [179]. This exhibits a sharp contrast to normal cells, which show a decrease of glucose uptake and inhibition of lactate production under aerobic conditions. A recent study found in addition, that glucose-starved colon cancer cells tend to be more aggressive [115].

NetRank genes cover Hallmarks of Cancer

Table 15 and Table 16 show the Transfac and HPRD-based genes mapped to Hanahan and Weinberg's Hallmarks of cancer. The relationship between the Hallmarks and a certain gene was assessed via Gene-Ontology terms. Gene-Ontology terms were extracted using the Database for Annotation, Visualization and Integrated Discovery (DAVID) [82]. Afterwards the Ontology-terms were annotated to one of the Hallmarks.

Table 16: The Hallmarks of cancer are covered by the signature genes based on the HPRD network, The majority of Hallmarks are represented by at least 3 proteins. The vast majority of genes caps at least 2 Hallmarks. Similar to Transfac, the Hallmark of metabolic reprogramming is sparsely represented only by the cAMP – dependent protein kinase catalytic subunit alpha (PRKACA).

Gene	↱	⊘	✚	∩	8	☗	✿	⚛
ATXN1			✓					
EWSR1				✓				
YWHAG					✓			
PRKCA		✓	✓	✓		✓		
PRKACA					✓		✓	✓
SMAD2		✓						
CREBBP							✓	
SMAD3	✓	✓	✓		✓	✓	✓	
NDRG1					✓			
CSNK2A1	✓	✓						
SRC						✓		
ATN1			✓					
FDCSP						✓		
CDK1			✓		✓			
IGJ							✓	
MUC4						✓		
SKIL			✓					
SVIL						✓		
TGFBR1	✓	✓	✓	✓		✓		
CR2							✓	
EGFR		✓	✓		✓	✓		
EP300	✓		✓					
ESR1			✓					

Transfac and HPRD-based NetRank genes cover Hallmarks of cancer

In the case of Transfac-based signature genes, each hallmark is covered by at least one signature gene. The Hallmarks of metabolic reprogramming and metastasis formation are sparsely covered by only one protein - Hepatocyte nuclear factor 4 alpha (HNF4A). All other Hallmarks are represented by at least three genes, with the resistance to inhibitory signals being covered by eight out of the 14 proteins. In addition, all genes, except JunD, represent more than one Hallmark. It was only possible to connect JunD to the Hallmark of blood vessel growth. Specifically, the most frequently observed Hallmarks are resisting to inhibitory signals, apoptosis and angiogenesis.

A similar picture can be drawn for genes obtained on the HPRD network (see Table 16). Again the Hallmark of metabolic reprogramming is sparsely represented only by the cAMP-dependent protein kinase catalytic subunit alpha (PRKACA). The remaining Hallmarks are covered by at least 3 proteins and the vast majority of genes cap at least 2 Hallmarks.

5.4 TESTING POSSIBLE UNIVERSAL CANCER SIGNATURES

In the previous section the relationship between the NetRank signatures and the Hallmarks of cancer have been discussed. Having these genes covering all Hallmarks of cancer, it might be possible to generate a Universal Cancer Signature able to predict any kind of cancer. This idea is reviewed in the following section.

5.4.1 Random signature as Universal Cancer Signature?

Venet and co-workers claim that 90% of all random signatures with more than 100 genes are a similar good predictor as any constructed signature [188]. This statement is tested on the in-house pancreas cancer dataset of Winter *et al.* [196]. Therefore, 1000 signatures of length 101 genes are created and during cross-validation it is measured how often the accuracy is better than the original NetRank accuracy of 72%.

90% of random signatures are NOT better than constructed signatures

RESULTS The random signatures where never better than the constructed signatures, nevertheless approaching the accuracy with a difference of 5% (68%) in 5 out of the 1000 cases. The overall prediction accuracy of the random signatures was 55.6%, which is an improvement over the random signatures of length 10 (52%). But this results indicates that 90% of random signatures with more than 100 genes are not better than constructed signatures, disproving the point of Venet and co-workers.

After all, following this study, maximal 10^{275} combinations have to be tested to discover a Universal Cancer Signature with 100 out of roughly 22.000 human genes. This is computationally demanding and not easily feasible at this point as testing the 1000 combinations for the pancreas cancer dataset in a Monte Carlo cross-validation procedure took already 200 CPU h. Therefore, selective testing is inevitable.

5.4.2 Combination of the most selected genes

Investigating signatures obtained by NetRank – based on Transfac and HPRD – a significant similarity between the different signatures could be observed and 37 genes have been chosen more at least three times as predictor. This limits the search space from roughly 22,000 human genes to 37. Nevertheless choosing 10 out of 37 still leads to 10^9 comparisons, assuming the best Universal Cancer Signature consist out of 10 genes. To test all possible combinations, 2^{37} comparisons have to be computed. This is computationally demanding and currently also not easily feasible.

5.4.3 Baseline accuracy - Top 10 highly connected transcription factors

To assess the performance of any signature, a baseline accuracy is needed. Highly connected genes are often well known cancer genes. As discussed before, most cancer types arise due to mutations in regulatory elements, creating alterations in transcription factor expression, which finally leads to unregulated cell growth and differentiation. In addition, transcription factors have been discussed as one of the main sources of cancer development [128] as well as suggested as targets for cancer therapy [47].

Following this, the investigation focuses on the most highly connected genes in the Transfac network (TF10) and chooses them as the baseline signature. The TF10 signature consisted of the following genes: `TP53, NFKB1, RELA, JUN, SP1, STAT1, CREB1, HNF4A, FOS, TFAP2A`.

Highly connected Transfac genes not useful as a Universal Cancer Signature

Table 17 column TF10 shows the accuracy of these signature on the benchmark dataset. It is better than randomly selected signatures in 12 out of 25 datasets, resulting in no average improvement. In comparison to the classical methods of *t*-test and foldchange, the TF10 is better in seven out of 25 datasets, nevertheless the resulting overall performance was 5% worse.

5.4.4 Top 10 most often used NetRank genes

NetRank chooses genes based on the gene expression and the connectivity of that gene in a network. Out of the 37 genes the top 10 most often used Transfac-based NetRank genes (NR10) were combined into a possible Universal Cancer Signature and tested on the 25 datasets. The NR10 signature consists of the following genes: `SP1, SP3, JUN, TP53, FOS, NFKB1, TFAP2A, CREB1, STAT1, EGR1`.

Table 17 column NR10 shows the accuracy of these signature on the benchmark dataset. It is better than random predictors in 13 out of 25 datasets, resulting in an average improvement of 3%. In

Table 17: Prediction accuracies achieved by the probable Universal Cancer signatures of NR10 and TF10 in comparison to random predictors (Rand), classical methods (FC and TT) and original NetRank signatures (NR).

Author		FC	TT	NR	NR10	TF10	Rand
Jones *et al.* [90]		86	85	89	82	77	79
O'Donnell *et al.* [132]		66	78	81	75	75	75
van de Vijver *et al.* [191]		61	58	67	-	-	62
Spira *et al.* [168]		66	67	71	65	62	65
Lee *et al.* [104]		70	69	75	55	57	56
Steidl *et al.* [170]		72	75	74	69	69	70
Bhojwani *et al.* [20]	T	64	58	64	42	39	51
	P	67	64	71	62	59	54
Bogunovic *et al.* [22]		53	63	67	63	69	56
Friedman *et al.* [65]		52	60	56	50	52	53
Raponi *et al.* [149]		59	55	62	50	50	54
Wang *et al.* [192]		66	65	69	64	65	66
Korkola *et al.* [96]		67	64	70	69	69	67
Landemaine *et al.* [100]		95	90	100	67	67	67
Nanni *et al.* [127]	D	89	96	100	90	89	90
	T	55	62	71	44	30	31
Iqbal *et al.* [86]		76	77	82	73	70	67
Fernandez *et al.* [59]		78	67	81	46	43	44
Frank *et al.* [64]		70	67	73	66	71	66
Smith *et al.* [166]		68	57	72	60	60	62
Lenz*et al.* [106]		60	59	64	61	60	60
Mok *et al.* [122]		60	54	65	66	54	65
Dressman *et al.* [52]		51	46	60	57	57	43
Murat *et al.* [125]		49	46	59	55	54	48
Zhu *et al.* [206]		50	46	56	57	58	50

Accuracy %:							
	30-34	35-40	41-45	45-50	51-55	56-60	61-65
	66-70	71-75	76-80	81-85	86-90	91-95	96-100

comparison to the classical methods, the NR10 is better in 10 out of 25 datasets, nevertheless the resulting overall performance was 7% worse. Note that the difference between TF10 and NR10 is quite low as they overlap in eight out of ten genes.

As the overall accuracy is not better than the individual signatures, the NetRank based NR10 fails to reliably predict different cancer outcome datasets.

Most often chosen Transfac genes not useful as a Universal Cancer Signature

Even though covering all Hallmarks of cancer, neither the NR10 nor the baseline TF10 signature is suitable as a Universal Cancer Signature.

5.4.5 Combination of HPRD and Transfac genes

Another possibility could be a combination of the most often chosen Transfac and HPRD genes. As transcription factors are especially appropriate for cancer outcome prediction, all Transfac genes are taken and combined them with the HPRD genes, which were at least selected 6 times as predictive in the 25 datasets. Table 18 describes the resulting signature (COMB10) in detail.

Table 19 shows the predictive results of each part of the signatures and the combination of the genes. TF indicates the accuracy obtained by using 14 Transfac genes and HPRD are the remaining 9 genes from Table 18. The NetRank column shows the accuracy obtained by the original NetRank signatures build on Transfac.

Combination of top genes not useful as a Universal Cancer Signature

The dataset of van de Vijver *et al.* [191] was excluded, as this dataset does not contain gene expression information for all genes in COMB10. As the table shows, in only five out of the remaining 24 datasets, COMB10 achieved a better accuracy then using Transfac or HPRD signatures alone (bold numbers). In another five datasets the results of the COMB10 signature are comparable to the original NetRank signatures. In the special case of Mok *et al.* [122] the accuracy achieved by the combination of highly connected HPRD and Transfac genes is better than the original NetRank signature (69% versus 65%).

For this comparison, several machine learning approaches are applied, but for sake of comparability only accuracies obtained with SVM are shown. Again the k-nearest neighbor method achieved the best prediction accuracy with 63%, 65%, 63% accuracy in Transfac, HPRD and combined, respectively. In contrast, the Support Vector Machines only reached an accuracy of 61%, 61% and 62%. Nevertheless the differences are subtle, thus the performance of the different machine learning methods are comparable to each other. In small datasets with less than 100 patients, differences in prediction accuracy of 1% lead to the not significant correct prediction of one extra patient.

As summary, the individual NetRank signatures are still better than a combination of the most important Transfac and HPRD genes.

5.4.6 Individual NetRank Signatures

As testing all possible combination of the NetRank genes is not feasible and also NR10, TF10 or COMB10 failed to reliably predict the outcome, maybe a NetRank signature obtained on the benchmark dataset could be a better predictor for other datasets. Therefore, signatures obtained on each dataset were tested on the remaining 24 datasets. The following analysis focuses on Transfac-based NetRank signatures.

Table 18: Transcription factors are especially appropriate for cancer outcome prediction, therefore we took all Transfac genes from Table 13 and combined them with the HPRD genes, that were at least selected 6 times as predictive in the benchmark dataset. The table summaries these signature genes (COMB10).

GeneID	Symbol	Description	HPRD	TF
6310	ATXN1	ataxin 1	✓	
1051	CEBPB	CCAAT/enhancer binding protein (C/EBP), beta		✓
1385	CREB1	cAMP responsive element binding protein 1		✓
1387	CREBBP	CREB binding protein	✓	
1457	CSNK2A1	casein kinase 2, alpha 1	✓	
1958	EGR1	early growth response 1		✓
2130	EWSR1	EWS RNA-binding protein 1	✓	
2353	FOS	FBJ murine osteosarcoma viral oncogene homolog		✓
3172	HNF4A	hepatocyte nuclear factor 4, alpha		✓
3725	JUN	jun proto-oncogene		✓
3727	JUND	jun D proto-oncogene		✓
10397	NDRG1	N-myc downstream regulated 1	✓	
4790	NFKB1	nuclear factor of kappa light polypeptide gene enhancer in B-cells 1		✓
5566	PRKACA	protein kinase, cAMP-dependent, catalytic, alpha	✓	
5578	PRKCA	protein kinase C, alpha	✓	
5970	RELA	v-rel avian reticuloendotheliosis viral oncogene homolog A		✓
4087	SMAD2	SMAD family member 2	✓	
4088	SMAD3	SMAD family member 3	✓	
6667	SP1	Sp1 transcription factor		✓
6670	SP3	Sp3 transcription factor		✓
6714	SRC	v-src avian sarcoma (Schmidt-Ruppin A-2) viral oncogene homolog	✓	
6772	STAT1	signal transducer and activator of transcription 1, 91kDa		✓
7020	TFAP2A	transcription factor AP-2 alpha		✓
7157	TP53	tumor protein p53		✓

Table 19: Results of a combined signature - TF are genes from Transfac, HPRD are genes from HPRD and combined indicates accuracy by combining both signatures (COMB10). The NetRank column gives the accuracy obtained by NetRank based on Transfac. In bold we marked, when COMB10 outperforms the TF and HPRD-based signatures. In only five out of the remaining 24 datasets, COMB10 achieved a better accuracy than using Transfac or HPRD signatures alone.

Dataset		TF	HPRD	COMB10	NetRank
Jones *et al.* [90]		83	81	83	**97**
Spira *et al.* [168]		62	67	**69**	69
Lee *et al.* [104]		55	56	**57**	76
Steidl *et al.* [170]		71	68	70	74
Bhojwani *et al.* [20]	T	43	55	49	64
	P	60	63	60	72
Bogunovic *et al.* [22]		63	57	59	70
Raponi *et al.* [149]		46	56	**58**	59
Wang *et al.* [192]		65	67	67	67
Korkola *et al.* [96]		69	68	**71**	71
O'Donnell *et al.* [132]		75	75	75	81
Friedman *et al.* [65]		56	54	55	55
Landemaine *et al.* [100]		67	67	67	99
Nanni *et al.* [127]	D	88	86	89	99
	T	36	31	36	70
Iqbal *et al.* [86]		71	68	69	82
Fernandez *et al.* [59]		42	44	44	81
Frank *et al.* [64]		70	67	66	75
Smith *et al.* [166]		60	62	60	71
Lenz *et al.* [106]		61	59	**63**	65
Mok *et al.* [122]		55	70	69	65
Dressman *et al.* [52]		59	40	42	64
Murat *et al.* [125]		50	44	49	59
Zhu *et al.* [206]		54	54	54	58

Accuracy %:							
	30-34	35-40	41-45	45-50	51-55	56-60	61-65
	66-70	71-75	76-80	81-85	86-90	91-95	96-100

The results can be found in Figure 21, where the x-axis shows the datasets and the y-axis shows the signatures sorted by best average accuracy. White boxes indicate accuracies worse than 50% (random guess). As the figure shows, nearly all signatures are better than random. Caution has to be taken in the case of van't Veer et al. [185], as the dataset does not provide information about all transcription factors used.

Signature obtained from Fernandez et al. dataset shows best performance

Figure 21 is sorted by average accuracy, indicating the best signature as the first row, which is the NetRank signature created on the data of Fernandez *et al.* for lymphoma subtyping [59]. The signature is composed of the following genes: `SP1`, `SP3`, `TP53`, `HDGFRP3`, `RNGTT`, `FOS`, `HNF4A`, `JUN`, `PON2`, `FARP1`. This signature appears to consist of general cancer proteins: `SP1`, `SP3`, `TP53`, `FOS`, `HNF4A`, `JUN`

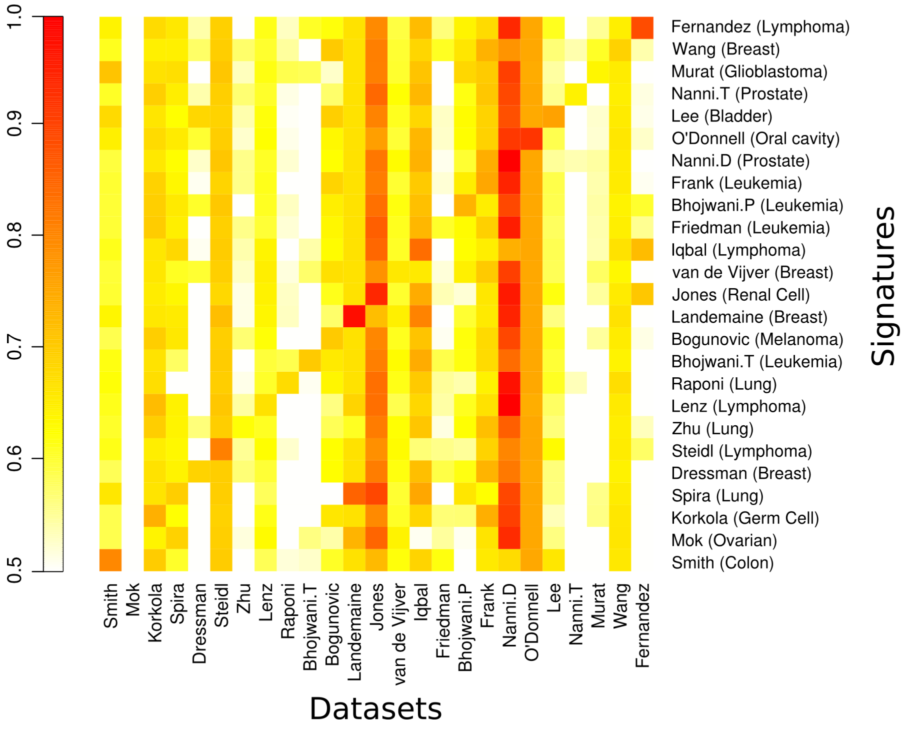


Figure 21: Accuracy of individual NetRank signatures against all datasets. The x-axis shows the datasets and the y-axis shows the NetRank signatures sorted by best average signature. White boxes indicate an accuracy worse than 50%. Signatures perform better than random (only a few white boxes) and the NetRank signature obtained on the Fernàndez et al. performs best on all datasets. Some datasets (Nanni D, Jones and Steidl) are solved well by all signatures.

and specific proteins: `HDGFRP3, RNGTT, PON2, FARP1`, which – in addition – cover all Hallmarks of cancer. The signature achieves on average a prediction accuracy of 62%, ranging across many different cancer types. In 11 out of 25 cases, this lymphoma signature outperforms the classical methods, creating cancer-specific signatures.

The diagonal represents the original NetRank accuracy, obtained and tested on the same dataset. Due to the significant similarity between predictive signatures, accuracies on the diagonal are similar to the rest of the comparisons.

Some datasets are difficult to capture

Furthermore, there are datasets which are easier to predict than others. The signal in the Nanni *et al.* diagnosis dataset [127] is strong enough to be easily captured by all signatures. In contrast, some datasets e.g. Mok *et al.* [122] and Nanni *et al.* (T) seem to be difficult to predict, as our method obtains average accuracies below 50%.

Although these results show a better prediction than a random guess, there is still little evidence for a generic Universal Cancer Signature. Nevertheless, it appears that a cancer-specific combination of general master regulators achieves the best results for all cancer types.

5.4.7 Influence of general and specific cancer markers

As no Universal Cancer Signature was detected so far, we investigated the notion of having signatures with general and specific cancer genes. This analysis was limited to Transfac-based NetRank signatures.

Table 20: Influence of general (Hub) and specific genes (noHub) on the prediction results. The Hub column indicates the prediction accuracy of the general cancer genes. These genes have more than 40 connections in the Transfac network. The NetRank column shows the accuracy obtained by the original NetRank signatures build on Transfac. For the majority of datasets, the combination of general and specific cancer genes lead to an improved prediction accuracy compared to using hub or non-hub genes alone.

Dataset		Hubs	noHubs	NetRank
Bhojwani *et al.* [20]	T	60	61	**64**
	P	68	**72**	**72**
Fernandez *et al.* [59]		60	**83**	81
Iqbal *et al.* [86]		71	79	**82**
Jones *et al.* [90]		67	92	**97**
Korkola *et al.* [96]		**71**	69	**71**
Lee *et al.* [104]		64	68	**76**
Lenz *et al.* [106]		63	59	**65**
Mok *et al.* [122]		55	63	**65**
Murat *et al.* [125]		58	**64**	59
O'Donnell *et al.* [132]		77	80	**81**
Smith *et al.* [166]		60	70	**71**
Steidl *et al.* [170]		70	72	**74**
Wang *et al.* [192]		66	**68**	67

Accuracy %:							
	30-34	35-40	41-45	45-50	51-55	56-60	61-65
	66-70	71-75	76-80	81-85	86-90	91-95	96-100

Hub genes are defined as genes with more than 40 connections in the Transfac network and include the following genes:

TP53	NFKB1	STAT3	CDKN1A	MYC	RELA	FOS
STAT1	CREB1	HNF4A	CEBPA	EGR1	CEBPB	JUNB
IRF1	YY1	GATA1	TFAP2A	SP3	ATF2	ETS1
JUND	GABPA	FOSB	JUN	SP1	USF1	

In order to see the influence of these two types of genes on the prediction result, we decided to return to the individual signatures and use solely genes for prediction that are in our definition either general or specific cancer genes. In this study general cancer genes are defined as genes having more than 40 connections in the Transfac network. Table 20 (bottom) shows genes that are, based on our definition, general cancer genes.

Not all signatures contain – following our definition – general and specific genes. Eleven out of 25 datasets do not fulfill the criteria, whereupon three datasets only contain specific genes (Raponi *et al.*, Spira *et al.*, van de Vijver *et al.*) and the remaining eight just contain hub genes (Bogunovic *et al.*, Dressmann *et al.*, Frank *et al.*, Friedman *et al.*, Landemaine *et al.*, Nanni *et al.*, Zhu *et al.*). The results might be slightly overestimated as the signatures were obtained using the entire dataset. During this analysis parts of the signature were reused, which might lead to overestimation of the predictive power.

Table 21: Overlap of the anti-profile signature of Bravo *et al.* [27] with the signature genes obtained with NetRank based on Transfac (NR-TF) and HPRD (NR-HPRD).

			Overlap anti-profile with	
Author	**Cancer**	**Var.**	**NR-HPRD**	**NR-TF**
Bhojwani *et al.* [20]	Leukemia	P	MUC4	-
		T	MUC4, POU4F1, PHACTR3	MUC4, POU4F1, PHACTR3
Fernandez *et al.* [59]	Lymphoma	S	SOX11	-
Jones *et al.* [90]	Renal cell	S	-	USH1C
Landemaine *et al.* [100]	Breast	P	MAGEA11	-
Raponi *et al.* [149]	Lung	P	-	KRT6A, KRT5
Smith *et al.* [166]	Colon	P	MAGEA6	MAGEA6, WDR72
Steidl *et al.* [170]	Lung	T	TMSB15A	-
Wang *et al.* [192]	Breast	P	-	MMP1

Table 20 shows the results of the separated signatures in comparison to the original results. The Hubs column indicates the prediction accuracy of the general cancer genes. These genes have more than 40 connections in the Transfac network. The NetRank column shows the accuracy obtained by the original NetRank signatures build on Transfac.

For the majority of datasets, the combination of general and specific cancer genes lead to an improved prediction accuracy compared to using hub or non-hub genes alone. In the case of Fernandez *et al.* [59] and Murat *et al.* [125] the non-hub gene signature outperforms the NetRank signature by 2% and 5%, which might be due to the performance overestimation. In the remaining cases NetRank performs similar or better than the general or specific cancer gene signatures.

The results indicate that there is no Universal Cancer Signature, but that a combination of cancer-specific genes and general master regulators achieves the best results for all cancer types.

UNIVERSAL CANCER SIGNATURE OF BRAVO *et al.* In a recent publication, Bravo and co-workers present a statistical technique that identifies genes showing normal variation in healthy samples, but hypervariability across tumor samples [27]. In the first step they created a predictor consisting of 542 genes for colon cancer and achieved an AUC of 0.89 on 30 patients. Han *et al.* developed a signature using the same patient samples and achieved with only 5 genes a similar result (AUC 0.88).

In a second step Bravo *et al.* developed an anti-profile Universal Cancer predictor consisting of 716 genes using 688 healthy and 4.138 cancer samples from 59 types. As they did not have enough samples for each cancer type, no accuracy for their signature is stated.

NetRank genes overlap with anti-profile Universal Cancer predictor

These two signatures are compared with the NetRank signatures and common genes in NetRank signatures based on HPRD and Transfac are observed with the colon cancer signature and the anti-profile universal predictor. Our benchmark dataset includes one colon cancer dataset, which overlaps in two and three genes for HPRD and Transfac, respectively. These genes are `MAGEA6`, `SPP1` and `WDR72`.

Twelve out of the 716 genes of the anti-profile signature are found in the 25 NetRank signatures. Table 21 shows the datasets which contain genes found in the anti-profile signature. Interestingly, in the signature of the leukemia dataset of Bhojwani *et al.* [20] three out of ten genes are found in the anti-profile signature. In addition, the genes `MUC4` and `MMP1` are found in the highly connected HPRD and Transfac cluster (see Figure 20 and Figure 19). The matrix metallopeptidase (MMP1) is in the Transfac network an interconnector for `CEBPB` to the remaining Transfac genes, whereas `MUC4` is found in the periphery of the HPRD cluster.

This result proves that these overlapping genes could be responsible for both initiation and development of specific cancer types.

5.5 CONCLUSION

When analyzing gene expression data with the purpose of finding clinical relevant biomarkers for cancer outcome prediction, genes selected for prediction should not just work on the initial dataset but also on other patient data. While comparing breast cancer progression signatures published by van't Veer *et al.* [185] and Wang *et al.* [192] no significant overlap between these two signatures was found. The authors concluded that many equally predictive lists could have been produced [55]. If signatures have genes in common, these overlapping genes should cover general signals in cancer development and accordingly overlap with the Hallmarks of cancer. These Hallmarks, introduced by Hanahan and Weinberg in 2000, reduce cancer development and growths to a small number of underlying principles [80].

No similarity between the published signatures of the benchmark dataset

We continued along these lines and compared signatures obtained from 25 publications regarding their similarity. We show that there is no significant overlap between these signatures. Even for studies investigating the same cancer and outcome variable, no overlapping genes exist, proving the point of Ein-Dor and co-workers [55].

No overlap between signatures obtained by greedy methods

This lack of similarity might be due to different statistical methodologies used during signature development. To investigate the effect of the methods on the resulting predictive signatures, we applied two standard methods *t*-test and foldchange on all datasets. We show that both techniques were not able to construct stable signatures as no significant overlap between signatures was visible.

Network-based methods trigger signatures overlap

Finally, we investigated whether network-based analysis of gene expression data is able to improve the overlap between predictive signatures. We therefore analyzed the similarity of signatures obtained by NetRank – a PageRank derivative – that combines the correlation of expression level of a gene with the outcome variable of the patient by using a network of known gene-gene relationships. NetRank was applied on two types of networks – a protein-protein interaction network (HPRD) and a transcription factor-target network (Transfac). The resulting signatures share in average 2 and 3.2 genes, respectively.

Overlapping genes are known cancer biomarker and cover Hallmarks of cancer

The overlapping genes are highly connected and a literature analysis showed, that these genes are related to cancer and have been mentioned already as biomarker for outcome prediction. We investigated the genes regarding the Hallmarks of cancer and show that all Hallmarks are covered by the overlapping genes. We in particular single out six transcription factor and seven proteins and discuss their specific role in cancer progression.

No general Universal Cancer Signature could be identified

Having this set of overlapping genes, we tried to identify a Universal Cancer Signatures that is able to predict cancer development and progression in a variety of different cancer types. So far, no Universal Cancer Signature could be identified.

A combination of hub and no-hub genes approximates a Universal Cancer Signature

Nevertheless, we investigated the signatures regarding general and cancer specific genes and concluded that a combination of both achieves the best result for all cancer types.

6 CLOUD-BASED BIOMARKER DISCOVERY

As NetRank offers a great value for cancer outcome prediction, I wanted to make the algorithm accessible to other researchers. Nevertheless, running NetRank on large gene expression datasets requires appropriate computational resources as well as the adjustment of the algorithm and a specialized data management.

Cloud services offer such resources, but security concerns rise when valuable research data is transferred to a public Cloud. Such a Cloud service could provide physicians a valuable tool for supporting the selection of treatment options, when it is possible to use the gene expression data obtained from the patient's tissue sample. However, patient data is confidential and thus has to be secured appropriately while being processed in the Cloud.

6.1 INTRODUCTION TO SECURE CLOUD COMPUTING

The use of High-performance computing (HPC) systems for biomedical research has a long history [98] and is in the transition to Cloud computing (e.g. Cloud storage systems for DNA sequencing [123]) – going away from in-house solutions. However, this brings new problems such as vendor lock-in and a potentially untrusted environment. Contrary to the increased economic efficiency, flexibility and all the other advantages of the Cloud computing, there are still a few major drawbacks related to privacy issues and possible loss, which prevent broad adoption.

There are many proposals on how to enhance data privacy in the public and hybrid Cloud computing scenario [150]. However, these techniques possess advantages and disadvantages that need to be considered before application. Some methods like homomorphic encryption provide a high security level, but data transformation is extremely time consuming, thus increasing execution time dramatically. In contrast, techniques – like anonymization – have only a small impact on the computational complexity, but do not provide a high level of security.

In the context of the project GeneCloud strategies are developed to overcome issues in data security on identical complexity levels for

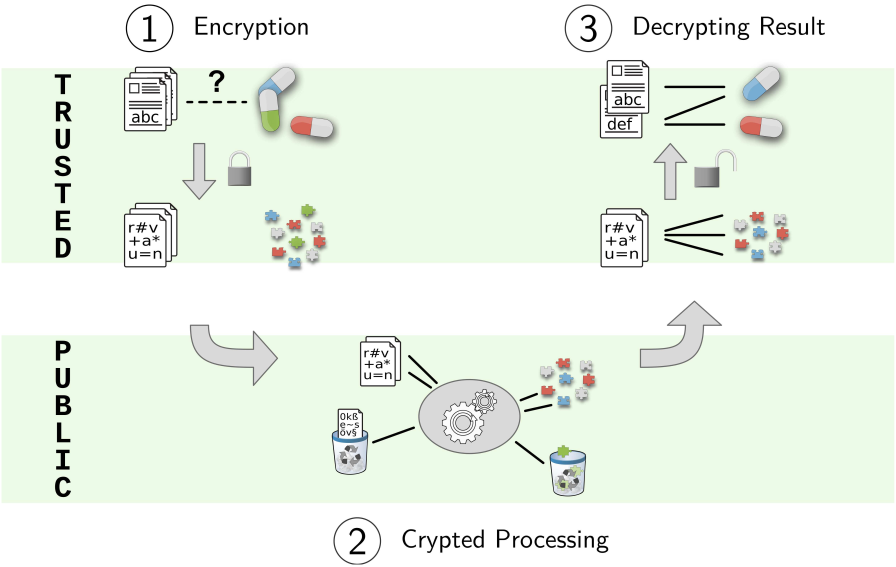


Figure 22: As a first step, data has to be encrypted in a trusted environment. Subsequently, the encrypted data is transferred to an untrusted environment e.g. Cloud, where processing is executed and encrypted results are obtained. After transferring the encrypted results back to a trusted environment, the data can be decrypted and the final results accessed.

services in a potentially untrusted environment (public Cloud) [17]. The GeneCloud project aims to develop model services to enable secure Cloud Computing in the drug development domain for small and medium size enterprises, which allows the application of high throughput methods in a secure manner.

Beck and co-workers present strategies – involving (homomorphic) encryption, minimization and anonymization – that are tailored to three case studies from different domains: Patent annotation (text mining), cancer outcome prediction (translational bioinformatics), and drug target prediction (structural bioinformatics) [18].

This chapter focuses on the results for the application of cancer outcome prediction.

A fully data-centric security approach is presented, utilizing customized algorithms to achieve a desired privacy-performance tradeoff. Figure 22 sketches the general idea of the GeneCloud project. In a first step, sensible data is preprocessed and secured locally in full control of the user. Afterwards, the encrypted data is subsequently transferred to the potentially unsafe Cloud platform. Ideally, the Cloud services run on the encrypted data, evading an access by an intruder and ultimately decreasing privacy related usage restrictions of Cloud infrastructure. After calculation, the encrypted result is transferred back to the trusted environment, where it is decrypted for further processing.

6.2 CANCER OUTCOME PREDICTION

The main goal of outcome prediction lies in defining the future state of a patient based on its current disease state. One method in cancer outcome prediction is the analysis of gene expression with the help of high-density oligonucleotide arrays. The gene signatures obtained in such analyzes could be addressed as biomarkers for cancer progression. In chapter 4 an algorithm – called NetRank – is described that employs protein-protein interaction networks and ranks genes by using the random surfer model of Google's PageRank algorithm. It has been shown, that NetRank improves the outcome prediction accuracy compared to classical methods by 7% for pancreas cancer patients [196] and provides an average 6% accuracy improvement on a larger benchmark dataset, consisting of 13 different cancer types.

Analysis of NetRank requirements

NetRank performs a supervised learning process combined with a Monte Carlo cross-validation scheme to predict biomarker genes. The main component of the algorithm is a matrix (network) – vector (patient data) multiplication, thus having a computational complexity of approximately $\mathcal{O}(N^2)$. Depending on the amount of patients and network size, the running time of one biomarker prediction can easily reach up to 30.000 CPU h. In order to reduce this running time, each cross-validation step is submitted to the Cloud as one worker process.

The gene expression data obtained from patients and some type of interaction networks are highly confidential. Therefore, the security of the input data and the predicted biomarkers should be guaranteed in the Cloud environment. Extra computation time is needed depending on applied security approach (e.g. blinding, homomorphic encryption, and randomization) on the data.

NetRank is memory intensive

For an efficient cross-validation the network and patient data is loaded into memory, which introduces a memory-related overhead. In the current version, running the matrix-vector multiplication with a network of approximately 100.000 edges consumes at least 2 Gigabyte of memory. For current network sizes, NetRank can be run on commodity hardware having 4 GB RAM, but with increasing network sizes this is growing up to terabytes, especially when testing several networks at the same time, thus making the memory the major bottleneck of the algorithm.

Since these factors cause the algorithm to be very memory intensive, an optimization is necessary before efficiently running it on a public Cloud.

6.3 SECURITY ANALYSIS

At the core of the NetRank algorithm are two important primitives, matrix-vector multiplications and machine learning algorithms working on confidential data.

Privacy-preserving solutions for matrix-vector multiplications are available based on different security assumptions like semi-honest multi-party computation [10] or blinding [9, 11]. Solutions upon blinding are implemented, which achieve a high efficiency, increased privacy, but still leak some information.

Blinding as a fast security approach

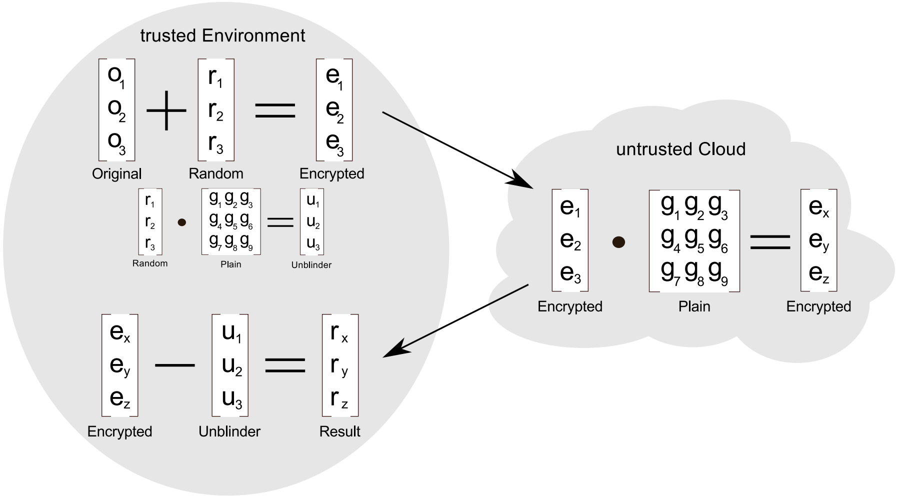


Figure 23: For the randomization of vectors, the vector-like patient data is transformed in a trusted environment by addition with a random vector. Afterwards, the matrix- vector multiplication is efficiently calculated in the untrusted Cloud environment. The blinded result is transferred back to the trusted systems and decrypted via subtraction by a prepared unblinding vector.

Figure 23 shows a schematic overview of the blinding process. For data randomization the vector-like patient data is transformed in a trusted environment by addition with a random vector. Afterwards, the matrix-vector multiplication is efficiently calculated in the untrusted Cloud environment. The blinded result is transferred back to the trusted systems and decrypted via subtraction by a prepared un-blinding vector.

Homomorphic encryption is less efficient but more secure

Another implemented solution is based on additive homomorphic encryption [139]. The encrypted approach is perfectly preserving privacy, but is less efficient than the blinding solution. It comes in two possible flavors, using an additive encryption scheme to protect one side of the matrix multiplication [139], or a scheme from pairing-based cryptography that protects all input data with less efficiency [24].

The machine learning part of NetRank was also targeted by specific privacy-preserving solutions. Graepel *et al.* propose a new class of machine learning algorithms based on leveled homomorphic encryption [99]. New binary classifiers are built with a multiplicative circuit depth of low polynomial degree. This ensures rather efficient evaluation through somewhat homomorphic encryption schemes. Another efficient and less privacy-preserving solution based on blinding is to randomize the data, while preserving important properties like the order and approximate relative distance. The blinding based approach also yields comparable classification results on the original data with nearly no performance impact. The achievable privacy margin is on the opposite rather low.

6.4 conclusion

Due to the growing amount of biological data, applications in the biomedical sector have a rising demand for computational power. Often the set-up of a whole computational infrastructure is too costly for small companies or research institutes. Cloud computing might be the solution to this problem. The cloud computing service provides running of client software on a remote server location, hence providing distributed computing and shared services in a converged cloud infrastructure. Unfortunately, data uploaded to the Cloud is often confidential and need to be secured.

Cancer outcome prediction tries to forecast disease progression from biological data (e.g. gene expression, sequencing etc.) and thus allows refined treatment options. By using the secured Cloud environment, new biomarkers found by NetRank could be applied in daily clinical routine and finally could suggest refined treatment options.

As the proposed security solution is independent of the actual Cloud infrastructure, NetRank can be run on virtually any Cloud platform.

7 CONTRIBUTIONS AND CONCLUSION

This chapter highlights the solutions to the open problems listed in chapter 1. It emphasizes the scientific contributions to analyze the applicability of network information for cancer outcome prediction.

7.1 CONTRIBUTIONS

Open Question 1 : Does network information improve cancer outcome prediction?
The goal is to show the general applicability of network-based prediction methods that incorporate network information in gene expression analysis in cancer outcome prediction. The performance should be assessed in comparison to standard outcome prediction methods on a representative benchmark dataset.

In this work a network-based approach – NetRank – is used to show the general applicability of network information in gene expression analysis to improve the performance in cancer outcome prediction. This analysis is conducted in chapter 4. For this purpose I created a benchmark dataset consisting of 26 dataset covering a broad variety of cancers and outcome variables. Before applying the algorithm on the benchmark dataset, I analyzed the influence of several parameters of NetRank e.g. the influence of the initial filtering, damping factor and noise on the prediction results. Following, I compared the performance of NetRank, classical methods and random predictors on the benchmark dataset. Therefore, NetRank had to be adapted to run on a cluster environment. Nevertheless, the evaluation of NetRank on two networks took 60 CPU years for the calculation of 500 million PageRanks.

This main analysis is followed by a secondary analysis regarding general parameters influencing outcome prediction like cohort size and outcome variable.

Finally, I conclude that network-based methods are especially suited for biomarker discovery. These methods improve prediction accuracies across a variety of cancer types by selecting genes that are not just correlating with but actively influencing the outcome of a patient.

Open question 2: Is there a Universal Cancer signature?
The goal is to assess the similarity between predictive signatures and how strongly published signatures are overlapping. Having strongly overlapping signature between different cancer types, the existence of a Universal Cancer Signature should be investigated that is able to predict the outcome of a patient independent of cancer type and outcome variable.

As the output of NetRank are highly overlapping signatures I investigated these similarities and compared them to signatures obtained with classical methods and the original signatures from a benchmark dataset consisting of 25 datasets. The results of this investigation are shown in chapter 5. The strongly overlapping genes are analyzed regarding literature occurrence, connectivity and Hallmarks of cancer. I analyze in detail six promising transcription factor and seven proteins regarding their biological relevance to cancer development and growth.

Subsequently, these genes are combined in different ways and several tests are conducted to discover a Universal Cancer Signature.

No Universal Cancer Signature could be identified so far. Nevertheless a combination of general and cancer specific genes leads to accurate prediction accuracies for a variety of cancers. Furthermore, I created a pool of promising Universal Cancer genes, whose combination could reflect a common predictive signature.

7.2 OPEN PROBLEM 1 REVISITED

In the last decade microarray studies developed to be a powerful tool to predict outcome of patients in various diseases. Many recent microarray studies have been accepted by the Food and Drug Administration (FDA) and are nowadays included in clinical routines, allowing personalized treatment options. Although the first diagnostic kits for cancer outcome prediction are in the market, the criticism remains that large signatures appear random, since other signatures perform equally well.

Network information can help to improve outcome prediction and reduce noise in microarray experiments. Many recent studies applied network-based methods and successfully improved outcome prediction of already published microarray data [40, 42, 45, 61, 129, 172, 181, 196].

Unfortunately, all these studies solely tested their approach on a small amount of datasets, questioning the general applicability of network-based methods for outcome prediction.

This is to my knowledge the first study to comprehensively test a network-based approach on a variety of different gene expression datasets. Our benchmark dataset is representative for all gene expression datasets as it covers 13 different cancer types, four distinct outcome variables, a huge variety of cohort sizes, signature sizes and statistical methodologies.

As representative of network-based methods, we adopted an algorithm recently proposed – NetRank –, which is a PageRank derivative working on gene-gene relationships. In this study NetRank employs known transcription factor-target relationships and protein-protein interactions for outcome prediction. For that purpose NetRank first assigns a score for each gene and then the network is used to spread this score to its neighbors and beyond. The genes with the highest NetRank score are then selected as signature genes. Here, we focus on signatures that are compact, consisting of only 10 genes and, therefore, can be easily tested in a clinical environment through inexpensive laboratory tests as RT-PCR or immunohistochemistry.

NetRank reliably predicts cancer outcome

We found NetRank reliably predicting the outcome in the benchmark dataset with high prediction accuracies. Furthermore, there was a consistent and significant improvement of the network-based approach over random signatures and classical approaches based on *t*-test or foldchange.

The purely computational approach of NetRank approaches the level of the authors' signatures

We compared the NetRank accuracies with the accuracies found by the original publications and showed that NetRank is approaching the accuracy level of the authors' signatures by applying a relatively unbiased but fully automated process for biomarker discovery. Taken together, a fully automated network-based approach – like NetRank – is able to obtain similar results as the hand-selected signatures of the authors. In addition, also other network-based methods successfully improved the prediction accuracy compared to the classical methods.

Regarding the open question, we conclude that NetRank is in particular suited for outcome prediction and that network-based methods in general improve prediction accuracies.

Signature size influences prediction performance

Another question was the correlation of cohort and signature size to prediction performance. On one hand we found a small correlation of signature size with prediction accuracy, but on the other hand no correlation for cohort size and accuracy could be observed.

Regulatory information best suited for outcome prediction

An important open question is, which type of interaction is best suited for outcome prediction. We experimented with a regulatory (Transfac), physical interaction (HPRD) and two predicted networks (hPrint, STRING). The latter are several orders of magnitude larger than the first. Nonetheless, all can be considered as a fraction of the complete interactome that is currently not yet known. Interestingly, despite the size difference, they perform equally well. This leads to the conclusion that regulatory information is particularly suited for outcome prediction. Cancer often arises due to alterations in transcription factor expression, leading to unregulated cell growth and differentiation. Therefore, the efficiency of Transfac in cancer outcome prediction is biologically reasonable. Transcription factors have been already discussed as one of the main source of cancer development [128] as well as suggested as targets for cancer therapy [47]. This implicates that future regulatory networks lead to further improvement in outcome prediction.

NetRank performance depends on many parameters

We furthermore investigated several parameters influencing the result of NetRank, e.g. how the genes initial values or the damping factor influences the results. The damping factor regulates the influence of the network on the results, as its balances the influence of the expression and interaction on the results. A weak correlation of d to accuracy improvement was observed, suggesting that increasing coverage of the interactome may also lead to further improvements in the prediction accuracy. The novelty of NetRank is the dynamical setting of the damping factor during signature development. As each dataset has its own dataset specific damping factor, the dynamical setting has to be maintained for each new biomarker creation.

The genes initial value plays also an important role on the prediction result. Initializing the genes with the same value instead with the gene expression leads already to an improvement over classical methods, but there is more prognostic information coming from the concrete expression values.

During evaluation of all open questions, we made the following discoveries regarding the outcome variable, the cancer bias and the machine learning method.

Diagnosis is easier than progression or treatment

While comparing the accuracies of the original publications, we discovered that predicting response to treatment and progression are generally more difficult than diagnosis and subtyping. We hypothesized that the former is more strongly influenced by external factors such as age and sex. Furthermore, we conclude that in diagnosis the gene expression signal is much stronger. This can be proven by comparing the random accuracy of the prostate dataset in the benchmark dataset with any other accuracy obtained by more sophisticated methods. The signal is strong enough to be captured even by random gene selection.

No bias due to well-studied proteins

Furthermore, we evaluated the disease bias that genes in the networks are often highly related to certain diseases. The results indicate that the information gain achieved by NetRank is not solely based on well-studied cancer proteins in networks but also on the interaction between them. Whereas, limiting the search space to only well-studied proteins for classical methods reduces the prediction performance dramatically.

Choice of machine learning method shows only small impact on accuracy

Finally, we investigated the influence of several machine learning approaches in outcome prediction. We could clearly observe a difference between diverse machine learning approaches. Overall, the k-NN approach performs best. Nevertheless the differences are subtle, thus the performance of the different machine learning methods are comparable to each other.

Network-based outcome prediction evaluation is computational expensive

Due to the cross-validation step and the dynamical setting of the damping factor included into NetRank, the calculation of a predictive signature takes several thousand CPU h. The evaluation of the algorithm on the whole benchmark dataset using two networks took 60 CPU years and the calculation of roughly 500 million PageRanks.

Regardless the quadratic running time, network-based gene expression analysis is leading to a more detailed understanding of cancer and cancer-related processes by selecting highly relevant genes that are not just correlating with but actively influencing the outcome of a patient. Furthermore, putting prognostic signatures into the context of pathways and network neighborhood may provide crucial information to move from biomarkers to targets, whose modulation will influence outcome.

7.3 OPEN PROBLEM 2 REVISITED

Cancer outcome prediction aims to forecast disease progression from gene expression data. Is has been shown, that single gene markers are not able to reliably predict cancer outcome. For that reason a combination of several genes combined in a predictive signature is nowadays used for outcome prediction. One success in this area is the diagnostic assay MammaPrint®, which predicts metastasis formation in breast cancer based on the expression of 70 genes. Ein-Dor and co-workers compared the breast cancer progression signature of MammaPrint® with another study and found no significant signature overlap between these two gene sets. They furthermore claimed that many equally predictive lists could have been produced [55].

When analyzing gene expression data with the purpose of finding clinical relevant biomarkers for cancer outcome prediction, genes selected for prediction should not just work on the initial dataset but also on other patient data. Therefore, signatures created for the same purpose should in general overlap. Having genes in common, these genes should cover general signals in cancer development and accordingly overlap with the Hallmarks of cancer. These Hallmarks, introduced by Hanahan and Weinberg in 2000, reduce cancer development and growths to a small number of underlying principles [80].

No similarity between the published signatures of the benchmark dataset

First, we followed the hypothesis of Ein-Dor and co-workers and compared the published signatures of the benchmark dataset regarding their similarity. We could not find a significant similarity between these signatures. Even for studies investigating the same cancer and outcome variable, no overlapping genes exist, proving the point of Ein-Dor *et al.*.

No overlap between signatures obtained by greedy methods

This lack of overlap might be due to different statistical methodologies employed during signature development. To investigate the effect of the methods on the resulting predictive signatures, we applied two standard methods *t*-test and foldchange on all datasets. We show that both techniques are not able to construct stable signatures as no significant overlap between signatures was visible.

Network-based methods trigger signatures similarity

Finally, we investigated whether network-based analysis of gene expression data is able to improve the similarity between predictive signatures. Thus, we analyzed the overlap of signatures obtained by NetRank, which was applied on two types of networks – a protein-protein interaction network (HPRD) and a transcription factor-target network (Transfac). A statistically significant overlap of predictive signatures was discovered and different datasets covering either the same cancer type, as well as different cancer types overlap well. The resulting signatures share 2 and 3.2 genes, respectively.

To summarize, NetRank, as a representative of network-based methods, is able to improve the similarity between signatures. The question arises if this overlap is meaningful.

Overlapping genes are biological meaningful and cover Hallmarks of cancer

To answer this question, the overlapping genes were investigated regarding their occurrence in literature as cancer-related or cancer biomarker. We found the overlapping genes are related to cancer and have been mentioned already as biomarker for outcome prediction. While examining the genes regarding the Hallmarks of cancer, we show that the overlapping genes cover all Hallmarks and thus all basic principles of cancer development. We in particular single out six transcription factors and seven proteins and discuss their specific role in cancer progression.

No Universal Cancer Signature could be identified so far

Having this set of overlapping highly cancer-related genes, we conducted several experiments to identify a Universal Cancer Signatures that is able to predict cancer development and progression in a variety of different cancer types. So far no Universal Cancer Signature could be identified. Nevertheless, we created a gene set consisting out of 37 genes which are highly promiscuous as a member of a Universal Cancer Signature.

A combination of hub and no-hub genes approximates Universal Cancer Signature

We investigated the signatures regarding general and cancer specific genes and concluded that a combination of both achieves the best result for all cancer types. Thus, the similarity between signatures could be improved by using genes indicating general cancer development and adding cancer specific genes.

Finally we compared our signatures to a signature published by Bravo *et al.* which contains 715 unique genes and is able to distinguish healthy and cancer patients [27]. We show that our signatures are partly overlapping with their anti-profile signature, indicating the general applicability of general gene sets for outcome prediction.

BIBLIOGRAPHY

[1] M. Abdelrahim et al. "Small Inhibitory RNA Duplexes for Sp1 mRNA Block Basal and Estrogen-induced Gene Expression and Cell Cycle Progression in MCF-7 Breast Cancer Cells." *Journal of Biological Chemistry* 277.32 (2002).

[2] A. Abdul-Wahid et al. "The carcinoembryonic antigen IgV-like N domain plays a critical role in the implantation of metastatic tumor cells." *Mol Oncol* (2013).

[3] Affymetrix. "Microarray Suite User's Guide." *5th edition* (2001).

[4] U. D. Akavia et al. "An integrated approach to uncover drivers of cancer." *Cell* 143.6 (2010).

[5] N. Akula et al. "A network-based approach to prioritize results from genome-wide association studies." *PLoS One* 6.9 (2011).

[6] M. Ameyar-Zazoua et al. "AP-1 dimers regulate transcription of the p14//p19ARF tumor suppressor gene." *Oncogene* (2005).

[7] G. Arpino et al. "Gene expression profiling in breast cancer: a clinical perspective." *Breast* 22.2 (2013).

[8] M. Ashburner et al. "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." *Nat Genet* 25.1 (2000).

[9] M. J. Atallah et al. "Secure outsourcing of scientific computations." *ADVANCES IN COMPUTERS* (1998).

[10] M. J. Atallah et al. "Securely outsourcing linear algebra computations." In: *Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security - ASIACCS '10*. New York, New York, USA: ACM Press, 2010.

[11] M. Atallah et al. "Secure outsourcing of some computations." *Computer* (1996).

[12] S. P. Balk et al. "Biology of prostate-specific antigen." *J Clin Oncol* 21.2 (2003).

[13] L. L. Bañez et al. "Obesity-related plasma hemodilution and PSA concentration among men with prostate cancer." *JAMA* 298.19 (2007).

[14] A. T. Baron et al. "Clinical implementation of soluble EGFR (sEGFR) as a theragnostic serum biomarker of breast, lung and ovarian cancer." *IDrugs* 12.5 (2009).

[15] J. A. Baron. "Screening for cancer with molecular markers: progress comes with potential problems." *Nat Rev Cancer* 12.5 (2012).

[16] R. C. Bast et al. "A radioimmunoassay using a monoclonal antibody to monitor the course of epithelial ovarian cancer." *N Engl J Med* 309.15 (1983).

[17] M. Beck et al. "Approximate Two-Party Privacy-Preserving String Matching with Linear Complexity" (2013).

[18] M. Beck et al. "GeneCloud: Secure Cloud Computing for Biomedical Research." *In Lecture Notes in Computer Science* (2014).

[19] F. Bertucci et al. "Gene expression profiles of poor-prognosis primary breast cancer correlate with survival." *Hum Mol Genet* 11.8 (2002).

[20] D. Bhojwani et al. "Gene expression signatures predictive of early response and outcome in high-risk childhood acute lymphoblastic leukemia: A Children's Oncology Group Study [corrected]." *J Clin Oncol* 26.27 (2008).

[21] H. Binder et al. "Calibration of microarray gene-expression data." *Methods Mol Biol* 576 (2010).

[22] D. Bogunovic et al. "Immune profile and mitotic index of metastatic melanoma lesions enhance clinical staging in predicting patient survival." *Proc Natl Acad Sci U S A* 106.48 (2009).

[23] M. Boivin et al. "CA125 (MUC16) tumor antigen selectively modulates the sensitivity of ovarian cancer cells to genotoxic drug-induced apoptosis." *Gynecol Oncol* 115.3 (2009).

[24] D. Boneh et al. "Evaluating 2-DNF formulas on ciphertexts." *Theory of Cryptography (TCC) '05* 3378 (2005).

[25] B. E. Boser et al. "A Training Algorithm for Optimal Margin Classifiers." In: *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*. ACM Press, 1992.

[26] A.-L. Boulesteix et al. "Evaluating microarray-based classifiers: an overview." *Cancer Inform* 6 (2008).

[27] H. C. Bravo et al. "Gene expression anti-profiles as a basis for accurate universal cancer signatures." *BMC Bioinformatics* 13 (2012).

[28] H. B. Burke. "Outcome prediction and the future of the TNM staging system." *J Natl Cancer Inst* 96.19 (2004).

[29] K. A. Burns et al. "Estrogen receptors and human disease: an update." *Arch Toxicol* 86.10 (2012).

[30] M. Buyse et al. "Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer." *J Natl Cancer Inst* 98.17 (2006).

[31] B. D. Camillo et al. "Effect of size and heterogeneity of samples on biomarker discovery: synthetic and real data assessment." *PLoS One* 7.3 (2012).

[32] M. Castagna et al. "Direct activation of calcium-activated, phospholipid-dependent protein kinase by tumor-promoting phorbol esters." *J Biol Chem* 257.13 (1982).

[33] D. G. de Castro et al. "Personalized cancer medicine: molecular diagnostics, predictive biomarkers, and drug resistance." *Clin Pharmacol Ther* 93.3 (2013).

[34] E. G. Cerami et al. "Pathway Commons, a web resource for biological pathway data." *Nucleic Acids Res* 39.Database issue (2011).

[35] C.-C. Chang et al. "LIBSVM: a library for support vector machines." *ACM Transactions on Intelligent Systems and Technology (TIST)* 2.3 (2011).

[36] L. Chen et al. "Identifying cancer biomarkers by network-constrained support vector machines." *BMC Syst Biol* 5 (2011).

[37] X. Chen et al. "Integrating Biological Knowledge with Gene Expression Profiles for Survival Prediction of Cancer." *J Comput Biol* 16.2 (2009).

[38] S. Chintharlapalli et al. "Betulinic acid inhibits colon cancer cell and tumor growth and induces proteasome-dependent and -independent downregulation of specificity proteins (Sp) transcription factors." *BMC Cancer* 11.1 (2011).

[39] S. Chintharlapalli et al. "Betulinic Acid Inhibits Prostate Cancer Growth through Inhibition of Specificity Protein Transcription Factors." *Cancer Research* 67.6 (2007).

[40] S. A. Chowdhury et al. "Identification of coordinately dysregulated subnetworks in complex phenotypes." *Pac Symp Biocomput* (2010).

[41] S. A. Chowdhury et al. "Subnetwork state functions define dysregulated subnetworks in cancer." *J Comput Biol* 18.3 (2011).

[42] H.-Y. Chuang et al. "Network-based classification of breast cancer metastasis." *Mol Syst Biol* 3 (2007).

[43] E. C. Connolly et al. "Complexities of TGF- targeted cancer therapy." *Int J Biol Sci* 8.7 (2012).

[44] Y. Cun et al. "Prognostic gene signatures for patient stratification in breast cancer: accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions." *BMC Bioinformatics* 13 (2012).

[45] P. Dao et al. "Inferring cancer subnetwork markers using density-constrained biclustering." *Bioinformatics* 26.18 (2010).

[46] P. Dao et al. "Optimally discriminative subnetwork markers predict response to chemotherapy." *Bioinformatics* 27.13 (2011).

[47] J. E. Darnell. "Transcription factors as targets for cancer therapy." *Nat Rev Cancer* 2.10 (2002).

[48] N. A. Davis et al. "Surfing a genetic association interaction network to identify modulators of antibody response to smallpox vaccine." *Genes Immun* 11.8 (2010).

[49] P. Dhawan et al. "The lymphotoxin-beta receptor is an upstream activator of NF-kappaB-mediated transcription in melanoma cells." *J Biol Chem* 283.22 (2008).

[50] V. Dixit et al. "NF-kappaB signaling. Many roads lead to madrid." *Cell* 111.5 (2002).

[51] K. K. Dobbin et al. "Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays." *Clin Cancer Res* 11.2 Pt 1 (2005).

[52] H. K. Dressman et al. "Gene expression profiles of multiple breast cancer phenotypes and response to neoadjuvant chemotherapy." *Clin Cancer Res* 12.3 Pt 1 (2006).

[53] R. O. Duda et al. *Pattern Classification*. 2nd ed. John Wiley & Sons, 2001.

[54] J. Dutkowski et al. "Protein networks as logic functions in development and cancer." *PLoS Comput Biol* 7.9 (2011).

[55] L. Ein-Dor et al. "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer." *Proc Natl Acad Sci U S A* 103.15 (2006).

[56] A. Elefsinioti et al. "Large-scale de novo prediction of physical protein-protein association." *Mol Cell Proteomics* 10.11 (2011).

[57] E. Erdei et al. "Cytokines and tumor metastasis gene variants in oral cancer and precancer in puerto rico." *PLoS One* 8.11 (2013).

[58] C. Fan et al. "Concordance among gene-expression-based predictors for breast cancer." *N Engl J Med* 355.6 (2006).

[59] V. Fernàndez et al. "Genomic and gene expression profiling defines indolent forms of mantle cell lymphoma." *Cancer Res* 70.4 (2010).

[60] N. I. Fleming et al. "SMAD2, SMAD3 and SMAD4 mutations in colorectal cancer." *Cancer Res* 73.2 (2013).

[61] K. Fortney et al. "Inferring the functions of longevity genes with modular subnetwork biomarkers of Caenorhabditis elegans aging." *Genome Biol* 11.2 (2010).

[62] K. Fortney et al. "Integrative computational biology for cancer research." *Human Genetics* 130 (4 2011). 10.1007/s00439-011-0983-z.

[63] A. Franceschini et al. "STRING v9.1: protein-protein interaction networks, with increased coverage and integration." *Nucleic Acids Res* 41.Database issue (2013).

[64] O. Frank et al. "Gene expression signature of primary imatinib-resistant chronic myeloid leukemia patients." *Leukemia* 20.8 (2006).

[65] D. R. Friedman et al. "A genomic approach to improve prognosis and predict therapeutic response in chronic lymphocytic leukemia." *Clin Cancer Res* 15.22 (2009).

[66] H. A. Fritsche et al. "CA 125 in ovarian cancer: advances and controversy." *Clin Chem* 44.7 (1998).

[67] N. Galanina et al. "Molecular predictors of response to therapy for breast cancer." *Cancer J* 17.2 (2011).

[68] L. García-Alonso et al. "Discovering the hidden sub-network component in a ranked list of genes or proteins derived from genomic experiments." *Nucleic Acids Res* 40.20 (2012).

[69] S. R. Gilman et al. "Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses." *Neuron* 70.5 (2011).

[70] P. Göcze et al. "[Ovarian carcinoma antigen (CA 125) and ovarian cancer (clinical follow-up and prognostic studies)]." *Orv Hetil* 134.17 (1993).

[71] M. D. Greicius et al. "Neuroimaging insights into network-based neurodegeneration." *Curr Opin Neurol* 25.6 (2012).

[72] J. Grosjean-Raillard et al. "ATM mediates constitutive NF-kappaB activation in high-risk myelodysplastic syndrome and acute myeloid leukemia." *Oncogene* 28.8 (2009).

[73] Z. Guo et al. "Edge-based scoring and searching method for identifying condition-responsive protein-protein interaction subnetwork." *Bioinformatics* 23.16 (2007).

[74] D. Gupta et al. "Pretreatment serum albumin as a predictor of cancer survival: a systematic review of the epidemiological literature." *Nutr J* 9 (2010).

[75] S. Gupta et al. "Test characteristics of alpha-fetoprotein for detecting hepatocellular carcinoma in patients with hepatitis C. A systematic review and critical analysis." *Ann Intern Med* 139.1 (2003).

[76] B. Gyorffy et al. "Evaluation of microarray preprocessing algorithms based on concordance with RT-PCR in clinical samples." *PLoS One* 4.5 (2009).

[77] D. T. Hai et al. "A pageranking based method for identifying characteristic genes of a disease." In: *Networking, Sensing and Control, 2008. ICNSC 2008. IEEE International Conference on.* 2008.

[78] B. V. Halldórsson et al. "Network-based interpretation of genomic variation data." *J Mol Biol* 425.21 (2013).

[79] D. Hanahan et al. "Hallmarks of cancer: the next generation." *Cell* 144.5 (2011).

[80] D. Hanahan et al. "The hallmarks of cancer." *Cell* 100.1 (2000).

[81] Y. Hosoi et al. "Up-regulation of DNA-dependent protein kinase activity and Sp1 in colorectal cancer." *Int J Oncol.* 25.2 (2004).

[82] D. W. Huang et al. "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." *Nat Protoc* 4.1 (2009).

[83] E. Huang et al. "Gene expression predictors of breast cancer outcomes." *Lancet* 361.9369 (2003).

[84] T. Ideker et al. "Boosting Signal-to-Noise in Complex Biology: Prior Knowledge Is Power." *Cell* 144.6 (2011).

[85] Igor et al. "DEGAS: De Novo Discovery of Dysregulated Pathways in Human Diseases." *PLoS ONE* 5.10 (2010).

[86] J. Iqbal et al. "Molecular signatures to improve diagnosis in peripheral T-cell lymphoma and prognostication in angioimmunoblastic T-cell lymphoma." *Blood* 115.5 (2010).

[87] R. A. Irizarry et al. "Exploration, normalization, and summaries of high density oligonucleotide array probe level data." *Biostatistics* 4.2 (2003).

[88] M. E. Irwin et al. "Src family kinases mediate epidermal growth factor receptor signaling from lipid rafts in breast cancer cells." *Cancer Biol Ther* 12.8 (2011).

[89] M. Johannes et al. "Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients." *Bioinformatics* 26.17 (2010).

[90] J. Jones et al. "Gene signatures of progression and metastasis in renal cell cancer." *Clin Cancer Res* 11.16 (2005).

[91] P. F. Jonsson et al. "Global topological features of cancer proteins in the human interactome." *Bioinformatics* 22.18 (2006).

[92] M. Karin. "NF-kappaB as a critical link between inflammation and cancer." *Cold Spring Harb Perspect Biol* 1.5 (2009).

[93] M. K. Kim et al. "Revised staging classification improves outcome prediction for small intestinal neuroendocrine tumors." *J Clin Oncol* 31.30 (2013).

[94] L. Klebanov et al. "How high is the level of technical noise in microarray data?" *Biol Direct* 2 (2007).

[95] O. Konopatskaya et al. "Protein kinase Calpha: disease regulator and therapeutic target." *Trends Pharmacol Sci* 31.1 (2010).

[96] J. E. Korkola et al. "Identification and validation of a gene expression signature that predicts outcome in adult men with germ cell tumors." *J Clin Oncol* 27.31 (2009).

[97] J. E. Korkola et al. "Identification of a robust gene signature that predicts breast cancer outcome in independent data sets." *BMC Cancer* 7 (2007).

[98] A. Krishnan. "GridBLAST: a Globus-based high-throughput implementation of BLAST in a Grid computing framework." *Concurrency and Computation: Practice and Experience* 17.13 (2005).

[99] T. Kwon et al. "ML Confidential: Machine Learning on Encrypted Data." *Information Security and Cryptology ,Äì ICISC 2012 Lecture Notes in Computer Science*. Lecture Notes in Computer Science 7839 (2013).

[100] T. Landemaine et al. "A six-gene signature predicting breast cancer lung metastasis." *Cancer Res* 68.15 (2008).

[101] D. A. Lashkari et al. "Yeast microarrays for genome wide parallel genetic and gene expression analysis." *Proc Natl Acad Sci U S A* 94.24 (1997).

[102] E. Lee et al. "Inferring pathway activity toward precise disease classification." *PLoS Comput Biol* 4.11 (2008).

[103] I. Lee et al. "Prioritizing candidate disease genes by network-based boosting of genome-wide association data." *Genome Research* (2011).

[104] J.-S. Lee et al. "Expression signature of E2F1 and its associated genes predict superficial to invasive progression of bladder tumors." *J Clin Oncol* 28.16 (2010).

[105] M. D. M. Leiserson et al. "Simultaneous identification of multiple driver pathways in cancer." *PLoS Comput Biol* 9.5 (2013).

[106] G. Lenz et al. "Stromal gene signatures in large-B-cell lymphomas." *N Engl J Med* 359.22 (2008).

[107] L. Levy et al. "Alterations in components of the TGF-beta superfamily signaling pathways in human cancer." *Cytokine Growth Factor Rev* 17.1-2 (2006).

[108] C. I. Li et al. "Incidence of invasive breast cancer by hormone receptor status from 1992 to 1998." *J Clin Oncol* 21.1 (2003).

[109] C. Li et al. "Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection." *Proc Natl Acad Sci U S A* 98.1 (2001).

[110] J. Lin et al. "A multidimensional analysis of genes mutated in breast and colorectal cancers." *Genome Res* 17.9 (2007).

[111] P. Lopez-Bergami et al. "c-Jun regulates phosphoinositide-dependent kinase 1 transcription: implication for Akt and protein kinase C activities and melanoma tumorigenesis." *J Biol Chem* 285.2 (2010).

[112] P. Lopez-Bergami et al. "Emerging roles of ATF2 and the dynamic AP1 network in cancer." *Nat Rev Cancer* (2010).

[113] S. Lu et al. "Sp1 coordinately regulates de novo lipogenesis and proliferation in cancer cells." *International Journal of Cancer* 126.2 (2010).

[114] Y. Lu et al. "Epidermal growth factor receptor (EGFR) ubiquitination as a mechanism of acquired resistance escaping treatment by the anti-EGFR monoclonal antibody cetuximab." *Cancer Res* 67.17 (2007).

[115] L. Ma et al. "Control of nutrient stress-induced metabolic reprogramming by PKC in tumorigenesis." *Cell* 152.3 (2013).

[116] X.-J. Ma et al. "A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen." *Cancer Cell* 5.6 (2004).

[117] S. Maestranzi et al. "The effect of benign and malignant liver disease on the tumour markers CA19-9 and CEA." *Ann Clin Biochem* 35 (Pt 1) (1998).

[118] S. Mahner et al. "C-Fos expression is a molecular predictor of progression and survival in epithelial ovarian carcinoma." *Br J Cancer* 99.8 (2008).

[119] V. Matys et al. "TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes." *Nucleic Acids Res* 34.Database issue (2006).

[120] S. Michiels et al. "Prediction of cancer outcome with microarrays: a multiple random validation strategy." *Lancet* 365.9458 (2005).

[121] D. Mochly-Rosen et al. "Protein kinase C, an elusive therapeutic target?" *Nat Rev Drug Discov* 11.12 (2012).

[122] S. C. Mok et al. "A gene signature predictive for outcome in advanced ovarian cancer identifies a survival factor: microfibril-associated glycoprotein 2." *Cancer Cell* 16.6 (2009).

[123] J. C. Morgan et al. "A next generation sequence processing and analysis platform with integrated cloud-storage and high performance computing resources." In: *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*. BCB '12. Orlando, Florida: ACM, 2012.

[124] J. L. Morrison et al. "GeneRank: using search engine technology for the analysis of microarray experiments." *BMC Bioinformatics* 6 (2005).

[125] A. Murat et al. "Stem cell-related "self-renewal" signature and high epidermal growth factor receptor expression associated with resistance to concomitant chemoradiotherapy in glioblastoma." *J Clin Oncol* 26.18 (2008).

[126] R. B. Nadler et al. "Effect of inflammation and benign prostatic hyperplasia on elevated serum prostate specific antigen levels." *J Urol* 154.2 Pt 1 (1995).

[127] S. Nanni et al. "Epithelial-restricted gene profile of primary cultures from human prostate tumors: a molecular approach to predict clinical behavior of prostate cancer." *Mol Cancer Res* 4.2 (2006).

[128] D. W. Nebert. "Transcription factors and cancer: an overview." *Toxicology* 181-182 (2002).

[129] R. K. Nibbe et al. "An integrative -omics approach to identify functional sub-networks in human colorectal cancer." *PLoS Comput Biol* 6.1 (2010).

[130] R. I. Nicholson et al. "EGFR and cancer prognosis." *Eur J Cancer* 37 Suppl 4 (2001).

[131] V. Nossov et al. "The early detection of ovarian cancer: from traditional methods to proteomics. Can we really do better than serum CA-125?" *Am J Obstet Gynecol* 199.3 (2008).

[132] R. K. O'Donnell et al. "Gene expression signature predicts lymphatic metastasis in squamous cell carcinoma of the oral cavity." *Oncogene* 24.7 (2005).

[133] M. J. Okoniewski et al. "Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations." *BMC Bioinformatics* 7 (2006).

[134] G. M. Osmani et al. "Using Google's PageRank algorithm to identify important attributes of genes." In: *Midwest Instruction and Computing Symposium 2007*. 2007.

[135] L. Ouafik et al. "Adrenomedullin promotes cell cycle transit and up-regulates cyclin D1 protein level in human glioblastoma cells through the activation of c-Jun/JNK/AP-1 signal transduction pathway." *Cell Signal* 21.4 (2009).

[136] X. Ouyang et al. "Activator protein-1 transcription factors are associated with progression and recurrence of prostate cancer." *Cancer Res* 68.7 (2008).

[137] L. Page et al. "The PageRank citation ranking: bringing order to the web." *Tech Rep Stanford Digital Library Technologies Project* (1998).

[138] S. Paik et al. "A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer." *N Engl J Med* 351.27 (2004).

[139] P. Paillier. "Public-Key Cryptosystems Based on Composite Degree Residuosity Classes." *Advances in Cryptography - Eurocrypt '99*. Lecture Notes in Computer Science 1592 (1999). Ed. by J. Stern.

[140] Y. H. Park et al. "Clinical relevance of TNM staging system according to breast cancer subtypes." *Ann Oncol* 22.7 (2011).

[141] J. S. Parker et al. "Supervised risk predictor of breast cancer based on intrinsic subtypes." *J Clin Oncol* 27.8 (2009).

[142] M. S. Patankar et al. "Potent suppression of natural killer cell response mediated by the ovarian tumor marker CA125." *Gynecol Oncol* 99.3 (2005).

[143] S. B. Paul et al. "Evaluating patients with cirrhosis for hepatocellular carcinoma: value of clinical symptomatology, imaging and alpha-fetoprotein." *Oncology* 72 Suppl 1 (2007).

[144] G. L. Perkins et al. "Serum tumor markers." *Am Fam Physician* 68.6 (2003).

[145] S. Pianetti et al. "Her-2/neu overexpression induces NF-kappaB via a PI3-kinase/Akt pathway involving calpain-mediated degradation of IkappaB-alpha that can be inhibited by the tumor suppressor PTEN." *Oncogene* 20.11 (2001).

[146] T. S. K. Prasad et al. "Human Protein Reference Database–2009 update." *Nucleic Acids Res* 37.Database issue (2009).

[147] L. N. Puls et al. "Current status of SRC inhibitors in solid tumor malignancies." *Oncologist* 16.5 (2011).

[148] J. Rangatia et al. "Elevated c-Jun expression in acute myeloid leukemias inhibits C/EBPalpha DNA binding via leucine zipper domain interaction." *Oncogene* 22.30 (2003).

[149] M. Raponi et al. "Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung." *Cancer Res* 66.15 (2006).

[150] K. Ren et al. "Security Challenges for the Public Cloud." *IEEE Internet Computing* 16.1 (2012).

[151] B. Z. Ring et al. "Novel prognostic immunohistochemical biomarker panel for estrogen receptor-positive breast cancer." *J Clin Oncol* 24.19 (2006).

[152] F. G. Rodrıguez-Gonzalez et al. "The challenge of gene expression profiling in heterogeneous clinical samples." *Methods* 59.1 (2013).

[153] F. B. Rogers. "Medical subject headings." *Bull Med Libr Assoc* 51 (1963).

[154] T. E. Royce et al. "Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification." *Nucleic Acids Res* 35.15 (2007).

[155] E. Rutgers et al. "The EORTC 10041/BIG 03-04 MINDACT trial is feasible: results of the pilot phase." *Eur J Cancer* 47.18 (2011).

[156] C. L. Sawyers. "The cancer biomarker problem." *Nature* 452.7187 (2008).

[157] M. Schena et al. "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." *Science* 270.5235 (1995).

[158] B. Schölkopf et al. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. Cambridge, MA: MIT Press, 2002.

[159] E. Shaulian et al. "AP-1 as a regulator of cell life and death." *Nat Cell Biol* 4.5 (2002).

[160] L. Shi et al. "Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential." *BMC Bioinformatics* 6 Suppl 2 (2005).

[161] L. Shi et al. "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements." *Nat Biotechnol* 24.9 (2006).

[162] M. Shi et al. "A network-based gene expression signature informs prognosis and treatment for colorectal cancer patients." *PLoS One* 7.7 (2012).

[163] Q. Shi et al. "Constitutive Sp1 Activity Is Essential for Differential Constitutive Expression of Vascular Endothelial Growth Factor in Human Pancreatic Adenocarcinoma." *Cancer Research* 61.10 (2001).

[164] Y. Shimizu et al. "Growth inhibition of non-small cell lung cancer cells by AP-1 blockade using a cJun dominant-negative mutant." *Br J Cancer* 98.5 (2008).

[165] R. Siegel et al. "Cancer statistics, 2013." *CA Cancer J Clin* 63.1 (2013).

[166] J. J. Smith et al. "Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer." *Gastroenterology* 138.3 (2010).

[167] C. Sotiriou et al. "Breast cancer classification and prognosis based on gene expression profiles from a population-based study." *Proc Natl Acad Sci U S A* 100.18 (2003).

[168] A. Spira et al. "Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer." *Nat Med* 13.3 (2007).

[169] P. R. Srinivas et al. "Proteomics for Cancer Biomarker Discovery." *Clinical Chemistry* 48.8 (2002).

[170] C. Steidl et al. "Tumor-associated macrophages and survival in classic Hodgkin's lymphoma." *N Engl J Med* 362.10 (2010).

[171] M. P. H. Stumpf et al. "Estimating the size of the human interactome." *Proceedings of the National Academy of Sciences* 105.19 (2008).

[172] J. Su et al. "Identification of diagnostic subnetwork markers for cancer in human protein-protein interaction network." *BMC Bioinformatics* 11 Suppl 6 (2010).

[173] D. I. Tai et al. "Constitutive activation of nuclear factor kappaB in hepatocellular carcinoma." *Cancer* 89.11 (2000).

[174] A. L. Tarca et al. "A novel signaling pathway impact analysis." *Bioinformatics* 25.1 (2009).

[175] A. A. Terentiev et al. "Alpha-fetoprotein: a renaissance." *Tumour Biol* 34.4 (2013).

[176] C. Thériault et al. "MUC16 (CA125) regulates epithelial ovarian cancer cell growth, tumorigenesis and metastasis." *Gynecol Oncol* 121.3 (2011).

[177] I. M. Thompson et al. "Prevalence of prostate cancer among men with a prostate-specific antigen level < or =4.0 ng per milliliter." *N Engl J Med* 350.22 (2004).

[178] R. Tibshirani et al. "Diagnosis of multiple cancer types by shrunken centroids of gene expression." *Proc Natl Acad Sci U S A* 99.10 (2002).

[179] J. S. TURNER et al. "Oxygen as a factor in photosynthesis." *Biol Rev Camb Philos Soc* 37 (1962).

[180] V. G. Tusher et al. "Significance analysis of microarrays applied to the ionizing radiation response." *Proc Natl Acad Sci U S A* 98.9 (2001).

[181] I. Ulitsky et al. "Detecting pathways transcriptionally correlated with clinical parameters." *Comput Syst Bioinformatics Conf* 7 (2008).

[182] C. M. Vajdic et al. "Cancer incidence and risk factors after solid organ transplantation." *Int J Cancer* 125.8 (2009).

[183] S. Vallabhapurapu et al. "Regulation and function of NF-kappaB transcription factors in the immune system." *Annu Rev Immunol* 27 (2009).

[184] F. Vandin et al. "Algorithms for detecting significantly mutated pathways in cancer." *J Comput Biol* 18.3 (2011).

[185] L. J. van't Veer et al. "Gene expression profiling predicts clinical outcome of breast cancer." *Nature* 415.6871 (2002).

[186] V. N. Vapnik. *Statistical Learning Theory*. New York: Wiley Interscience, 1998.

[187] V. M. Velonas et al. "Current status of biomarkers for prostate cancer." *Int J Mol Sci* 14.6 (2013).

[188] D. Venet et al. "Most random gene expression signatures are significantly associated with breast cancer outcome." *PLoS Comput Biol* 7.10 (2011).

[189] K. Venkatesan et al. "An empirical framework for binary interactome mapping." *Nat Methods* 6.1 (2009).

[190] R. G. W. Verhaak et al. "The effect of oligonucleotide microarray data pre-processing on the analysis of patient-cohort studies." *BMC Bioinformatics* 7 (2006).

[191] M. J. van de Vijver et al. "A gene-expression signature as a predictor of survival in breast cancer." *N Engl J Med* 347.25 (2002).

[192] Y. Wang et al. "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer." *Lancet* 365.9460 (2005).

[193] R. A. Weinberg. *The Biology of Cancer*. 2006.

[194] D. L. Wheeler et al. "The role of Src in solid tumors." *Oncologist* 14.7 (2009).

[195] R. Winnenburg et al. "Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies?" *Brief Bioinform* 9.6 (2008).

[196] C. Winter et al. "Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes." *PLoS Comput Biol* 8.5 (2012).

[197] D. Witten et al. "A comparison of fold-change and the t-statistic for microarray data analysis." *Stanford University* (2007).

[198] Z. Wu. "A review of statistical methods for preprocessing oligonucleotide microarrays." *Stat Methods Med Res* 18.6 (2009).

[199] Z. Wu et al. "Preprocessing of oligonucleotide array data." *Nat Biotechnol* 22.6 (2004).

[200] Z. Wu et al. "Network-based drug repositioning." *Mol Biosyst* 9.6 (2013).

[201] Y. Yan et al. "The MMP-1, MMP-2, and MMP-9 gene polymorphisms and susceptibility to bladder cancer: a meta-analysis." *Tumour Biol* (2014).

[202] B. Yang et al. "Network-based inference framework for identifying cancer genes from gene expression data." *Biomed Res Int* 2013 (2013).

[203] G. Yang et al. "The biphasic role of NF-kappaB in progression and chemoresistance of ovarian cancer." *Clin Cancer Res* 17.8 (2011).

[204] Y. Yarden et al. "Untangling the ErbB signalling network." *Nat Rev Mol Cell Biol* 2.2 (2001).

[205] L. Zhou et al. "Serum tumor markers for detection of hepatocellular carcinoma." *World J Gastroenterol* 12.8 (2006).

[206] C.-Q. Zhu et al. "Prognostic and predictive gene signature for adjuvant chemotherapy in resected non-small-cell lung cancer." *J Clin Oncol* 28.29 (2010).