Bioinformatics Approaches for Cancer Research

Dissertation

zur Erlangung des Grades des
Doktors der Naturwissenschaften
der Naturwissenschaftlich-Technischen Fakultäten
der Universität des Saarlandes

von
Christina Backes

Saarbrücken
April 2010

Datum des Kolloquiums: 09. Juli 2010

Dekan der Fakultät 6: Prof. Dr. Holger Hermanns

Mitglieder des Prüfungsausschusses:

Vorsitzender: Prof. Dr. Matthias Hein

Erster Gutachter: Prof. Dr. Hans-Peter Lenhof

Zweiter Gutachter: Prof. Dr. Eckart Meese

Wissenschaftlicher Beirat: Dr. Thomas In der Rieden

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form in einem Verfahren zur Erlangung eines akademischen Grades vorgelegt.

Saarbrücken, April 2010

(Christina Backes)

Abstract

Cancer is the consequence of genetic alterations that influence the behavior of affected cells. While the phenotypic effects of cancer like infinite proliferation are common hallmarks of this complex class of diseases, the connections between the genetic alterations and these effects are not always evident. The growth of information generated by experimental high-throughput techniques makes it possible to combine heterogeneous data from different sources to gain new insights into these complex molecular processes. The demand on computational biology to develop tools and methods to facilitate the evaluation of such data has increased accordingly. To this end, we developed new approaches and bioinformatics tools for the analysis of high-throughput data. Additionally, we integrated these new approaches into our comprehensive C++ framework GeneTrail. GeneTrail presents a powerful package that combines information retrieval, statistical evaluation of gene sets, result presentation, and data exchange. To make GeneTrail's capabilities available to the research community, we implemented a graphical user interface in PHP and set up a web-server that is world-wide accessible. In this thesis, we discuss newly integrated algorithms and extensions of GeneTrail, as well as some comprehensive studies that have been performed with GeneTrail in the context of cancer research. We applied GeneTrail to analyze properties of tumor-associated antigens to elucidate the mechanisms of antigen candidate selection. Furthermore, we performed an extensive analysis of miRNAs and their putative target pathways and networks in cancer. In the field of differential network analysis, we employed a combination of expression values and topological data to identify patterns of deregulated subnetworks and putative key players for the deregulation. Signatures of deregulated subnetworks may help to predict the sensitivity of tumor subtypes to therapeutic agents and, hence, may be used in the future to guide the selection of optimal agents. Furthermore, the identified putative key players may represent oncogenes, tumor suppressor genes, or other genes that contribute to crucial changes of regulatory and signaling processes in cancer cells and may serve as potential targets for an individualized tumor therapy. With these applications, we demonstrate the usefulness of our GeneTrail package and hope that our work will contribute to a better understanding of cancer.

German Abstract

Krebs ist eine Folge von tiefgreifenden genetischen Veränderungen, die das Verhalten der betroffenen Zellen beeinflussen. Während phänotypische Effekte wie unaufhörliches Wachstum augenscheinliche Merkmale dieser komplexen Klasse von Krankheiten sind, sind die Zusammenhänge zwischen genetischen Veränderungen und diesen Effekten oftmals weit weniger offensichtlich. Mit der stetigen Zunahme an Daten, die aus Hochdurchsatz-Verfahren stammen, ist es möglich geworden, heterogene Daten aus verschiedenen Quellen zu kombinieren und neue Erkenntnisse über diese Zusammenhänge zu gewinnen. Dementsprechend sind auch die Anforderungen an die Bioinformatik gewachsen, geeignete Applikationen und Verfahren zu entwickeln, um die Auswertung solcher Daten zu vereinfachen. Zu diesem Zweck haben wir neue Ansätze und bioinformatische Werkzeuge für die Analyse von entsprechenden Daten für die Krebsforschung entwickelt, welche wir in unser umfangreiches C++ System GeneTrail integriert haben. GeneTrail stellt ein mächtiges Softwarepaket dar, das Informationsgewinnung, statistische Auswertung von Gen Mengen, visuelle Darstellung der Resultate und Datenaustausch kombiniert. Um GeneTrail's Fähigkeiten der Forschungsgemeinschaft zugänglich zu machen, haben wir eine graphische Benutzerschnittstelle in PHP implementiert und einen Webserver aufgesetzt, auf den weltweit zugegriffen werden kann. In der vorliegenden Arbeit diskutieren wir neu integrierte Algorithmen und Erweiterungen von GeneTrail, sowie umfangreiche Untersuchungen im Bereich Krebsforschung, die mit GeneTrail durchgeführt wurden. Wir haben GeneTrail angewendet, um Eigenschaften von Tumorantigenen zu untersuchen, um aufzuklären, welche dieser Eigenschaften zur Selektion dieser Proteine als Antigene beitragen. Des Weiteren haben wir eine umfangreiche Analyse von miRNAs und deren potentiellen Zielpfaden und -netzen in verschiedenen Krebsarten durchgeführt. Im Bereich differentieller Netzwerkanalyse kombinierten wir Expressionswerte und topologische Netzwerkdaten, um Muster deregulierter Teilnetzwerke und mögliche Schlüsselgene für die Deregulation zu identifizieren. Signaturen deregulierter Teilnetzwerke können helfen die Sensitivität verschiedener Tumorarten gegenüber Therapeutika vorherzusagen und damit zukünftig eine optimal angepasste Therapie zu ermöglichen. Außerdem können die identifizierten potentiellen Schlüsselgene Oncogene, Tumorsuppressorgene, oder andere Gene darstellen, die zu wichtigen Änderungen von regulatorischen Prozessen in Krebszellen beitragen, und damit auch als potentielle Ziele für eine individuelle Tumortherapie in Frage kommen. Mit diesen Anwendungen untermauern wir den Nutzen von GeneTrail und hoffen, dass unsere Arbeit in Zukunft zu einem besseren Verständnis von Krebs beiträgt.

German Summary

Krebs ist eine der häufigsten Todesursachen in Industrieländern. Im Jahr 2004 starben weltweit ca. 7,4 Millionen Menschen an dieser Gruppe von Krankheiten. Die Projektionen für das Jahr 2030 erwarten sogar noch weiter steigende Zahlen an Todesfällen. Daher arbeiten weltweit viele Forschungsinstitute daran, wie es zur Entstehung von Krebs kommt, bis hin zur Diagnose und Therapie von Krebs. Moderne Therapiestrategien zur Behandlung von Krebs greifen direkt in komplexe zelluläre Prozesse ein, wobei die Nebenwirkungen oftmals nicht vorhersehbar sind. Daher ist Grundlagenforschung zu einem besseren Verständnis dieser molekularen Prozesse immer noch notwendig um in Zukunft eine gezieltere individuelle Behandlung zu ermöglichen.

Durch den explosiven Anstieg verfügbarer experimenteller Daten, der durch die technologischen Fortschritte im Bereich der Hochdurchsatz-Verfahren entstanden ist, ist die Analyse molekularer Prozesse in Krebs auf verschiedenen Ebenen mit bioinformatischen Methoden ermöglicht worden. In der vorliegenden Arbeit stellen wir computergestützte Verfahren vor, die die Auswertung und Interpretation von Daten aus Hochdurchsatz-Verfahren erleichtern sollen. Zu diesem Zweck haben wir ein umfangreiches System zur Gen-Mengen Analyse – genannt GeneTrail – entwickelt, dessen Funktionalität und Vielseitigkeit wir unter Beweis stellen, indem wir verschiedenen aktuellen Fragestellungen im Bereich Krebsforschung nachgehen.

GeneTrail selbst ist ein modular aufgebautes C++ System, dessen generelle Funktionalität der Detektion von statistisch angereicherten oder abgereicherten biologischen Kategorien mit Genen untersuchter Datenmengen dient. Unser System wurde ständig weiterentwickelt, um eine möglichst große Vielzahl an biologischen Kategorien, Organismen, und statistischen Methoden zu unterstützen. Des Weiteren wurde die Funktionalität von GeneTrail beispielsweise durch die Vorverarbeitung von Microarray Roh-Daten durch GeneTrailExpress, einer dynamischen Netzwerk-Visualisierung durch BiNA, sowie der Fähigkeit zur Durchführung differentieller Netzwerkanalysen ergänzt. Um GeneTrail's Fähigkeiten der Forschungsgemeinschaft zugänglich zu machen haben wir eine graphische Benutzerschnittstelle in PHP implementiert und einen Webserver aufgesetzt, auf den weltweit zugegriffen werden kann.

Während GeneTrail nicht nur zum Zwecke der Krebsforschung entwickelt wurde und auch

in anderen Bereichen eingesetzt werden kann, untersuchen wir in der vorliegenden Arbeit dennoch ausschließlich aktuelle Themengebiete der Krebsforschung. Als erste Anwendung führen wir eine umfangreiche Untersuchung potentieller Eigenschaften von Antigenen durch, die möglicherweise dafür verantwortlich sind, dass diese Antigene in Tumor- oder Autoimmun-Erkrankungen eine Immunantwort auslösen. Unsere Resultate zeigen Gemeinsamkeiten und Unterschiede zwischen Tumor- und Autoantigenen auf. Außerdem, weisen die untersuchten Antigene eine gewisse Prävalenz an Sequenzähnlichkeiten zu Proteinen in anderen Organismen auf, welches eine mögliche Begründung für das begrenzte Autoantikörper Repertoire darstellen könnte.

Als nächstes untersuchen wir die möglichen Zielpfade und -netwerke von miRNAs in verschiedenen Krebsarten. miRNAs sind eine Gruppe von nicht-codierender RNA, die direkt in die Genregulation komplementärer RNA eingreifen können. Wir führen eine Untersuchung mit Expressionsprofilen verschiedener Tumorarten durch und können zeigen, dass die Zielgene verschiedener miRNAs signifikant angereichert oder abgereichert sind. Des Weiteren finden wir Hinweise darauf, dass die Regulation durch miRNAs vermutlich eher auf Wechselwirkungen zwischen deren Konzentrationen basiert, statt auf Regulation einzelner wichtiger Hubs im regulatorischen Netzwerk. Unsere Resultate bestätigen die Rolle von miRNAs als Schlüsselkomponenten bei der Genregulation in Krebserkrankungen.

Als letztes Beispiel führen wir differentielle Netzwerkanalysen mit zwei verschiedenen neu entwickelten Algorithmen durch. Der erste Algorithmus – FiDePa – basiert auf der dynamischen Programmierung für die Berechnung von exakten Wahrscheinlichkeiten bei dem ungewichteten “Gene Set Enrichment Analysis” Verfahren. FiDePa findet deregulierte Pfade in einem regulatorischen Netzwerk, die statistisch signifikant sind. Für die differentielle Netzwerk Analyse mit diesem Algorithmus setzen wir Expressionsprofile von hochgradigen Glioma Erkrankungen in Vergleich zu Normalgeweben und zeigen, dass es möglich ist, Patienten-spezifische Teilnetzwerke aus der Vereinigung der berechneten signifikant deregulierten Pfade herzuleiten. Unser zweiter Algorithmus wendet ein ILP an und berechnet direkt den am meisten deregulierten Teil eines regulatorischen Netzwerkes, der zudem von einem Wurzelknoten ausgeht, von dem alle anderen Knoten des Teilnetzes erreichbar sind. Dieses Modell erzwingt, dass der Wurzelknoten die Eigenschaften einer potentiellen Schlüsselkomponente im Netzwerk hat, die direkten Einfluss auf die beobachteten Unterschiede der betrachteten Konditionen hat. Um die Stärke dieses Ansatzes unter Beweis zu stellen, berechnen wir das deregulierte Netzwerk für Expressionsprofile von BRCA1 Mutationsträgern im Vergleich zu Nicht-Mutationsträgern. Unsere Auswertung deutet darauf hin, dass oxidativer Stress eine wichtige Rolle in den Epithelzellen von BRCA1 Mutationsträgern spielt, welcher möglicherweise zu der späteren Entwicklung von Brustkrebs beiträgt. Beide Anzätze könnten in Zukunft in solch komplexen und heterogenen Krankheiten wie Krebs eingesetzt werden, um die Auswahl optimal angepasster Therapeutika zu erleichtern, sowie neue potentielle Zielmoleküle für eine individuelle Therapie zu identifizieren.

Zusammenfassend ist GeneTrail ein umfangreiches System zur Analyse und Auswertung von Daten aus Hochdurchsatz-Verfahren. Der modulare Aufbau unseres Systems erlaubt ein einfaches Erweitern und Anpassen, um aktuellen Fragestellungen aus unterschiedlichsten wissenschaftlichen Bereichen nachgehen zu können wie mit den Anwendungen in der vorliegenden Arbeit gezeigt. Ein weiteres Ziel unserer Arbeit ist auch anderen Forschungsgruppen die Möglichkeit zu geben, von unseren Entwicklungen zu profitieren, um damit neue Erkenntnisse für ihre eigenen Forschungen erlangen zu können.

Danksagung

Ich danke Herrn Prof. Dr. Hans-Peter Lenhof für die Vergabe des Themas, sowie für die Betreuung dieser Arbeit. Weiterhin möchte ich mich bei Herrn Prof. Dr. Eckart Meese für seine hilfreichen Ratschläge und Kommentare bedanken, sowie bei seiner Arbeitsgruppe, allen voran Nicole Ludwig, für die zur Verfügung gestellten Daten. Meinem wissenschaftlichen Begleiter Dr. Dirk Neumann danke ich, dass er stets versucht hat mich zu motivieren und immer ein offenes Ohr für meine Probleme hatte.

Ganz besonderen Dank möchte ich auch Dr. Andreas Keller für die erfolgreiche Zusammenarbeit an verschiedenen gemeinsamen Projekten wie GeneTrail und FiDePa aussprechen. In diesem Zusammenhang danke ich auch Cedric Laczny für seinen Beitrag zum GeneTrail Projekt während seiner Bachelorarbeit, sowie Maher Al-Awadhi, der als Bachelorarbeit das GeneTrailExpress Preprocessing implementiert hat. Des Weiteren möchte ich Andreas Gerasch für die Entwicklung des BiNA Visualisierers und dessen Anpassung an unsere Fragestellungen danken. Dr. Jan Küntzer danke ich für seine Hilfe, BN++/BNDB für GeneTrail nutzbar zu machen und seine vielen guten Ideen und Ratschläge während meiner Doktorarbeit. Nicht zuletzt möchte ich Alexander Rurainski für die Zusammenarbeit bei der Entwicklung des ILP Algorithmus zur Detektion deregulierter Teilnetzwerke danken.

Weiterhin möchte ich allen Mitarbeitern des Lehrstuhls von Prof. Lenhof, sowie der Nachwuchsgruppe Hildebrandt meinen Dank aussprechen. Dabei möchte ich besonders Benny Kneissl und Sophie Weggler erwähnen, die mich auch in schwierigen Zeiten unterstützt haben. Meinem “SC” Team danke ich für die vielen netten Abende, an denen wir “Forschung und Entwicklung” etwas anders als tagsüber betrieben haben.

Schließlich möchte ich mich noch bei meiner Familie bedanken, ohne deren jahrelange Unterstützung diese Arbeit nicht möglich gewesen wäre.

”Progress in science depends on new techniques, new discoveries, and new ideas, probably in that order.”

- *Sidney Brenner*

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Cancer is one of the most common causes of death in industrial countries promoted by the increase of age in the population. In 2004, cancer accounted for 7.4 million deaths (approximately 13%) worldwide. The projection for deaths from cancer for 2030 estimate rising numbers of over 12 million deaths worldwide as demonstrated in Figure 1.1 [1].

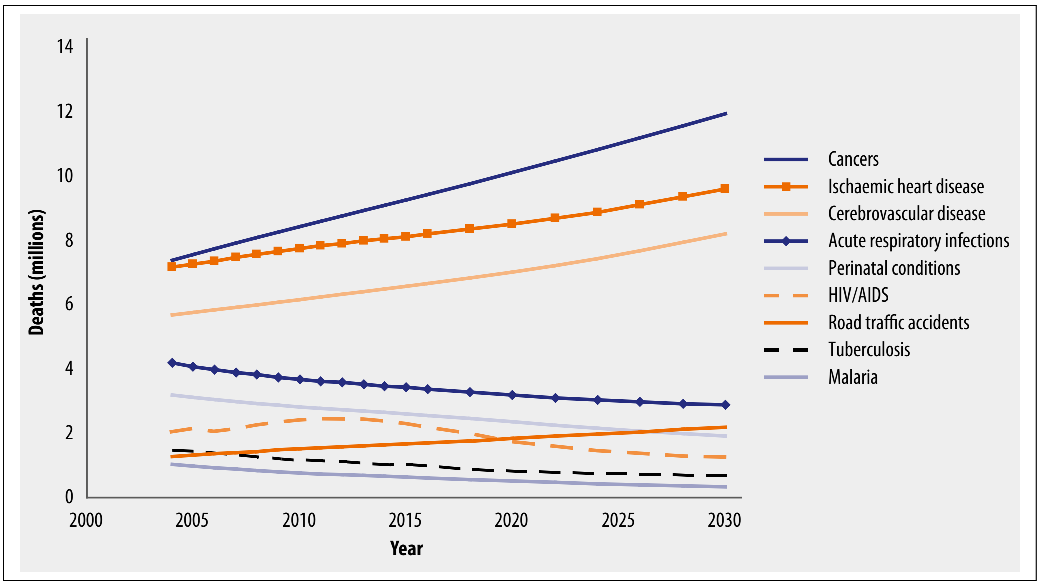


Figure 1.1: Projected global deaths for selected causes, 2004–2030.
Source: World Health Organization - Global Burden of Disease[1]

As a consequence, many research institutes are engaged in cancer research comprising cancer development, cancer diagnosis, and cancer therapy. Today, cancer treatment tries to interfere with the complex biochemical processes and signaling cascades in cancer cells. However, the aftermath of these treatments on the cells' behavior and the influence on healthy tissue are rarely exactly known as exemplified in the clinical trial of the mono-

[1] `http://www.who.int/healthinfo/global_burden_disease/GBD_report_2004update_part2.pdf`

clonal antibody TGN1412, originally intended for the treatment of B-cell chronic lymphocytic leukemia, which gained notoriety as the so-called TGN1412 tragedy [2]. The reasons for this incidence were allegedly a lack in knowledge about the biological processes as summarized to the point by the following citation:

> *"Where mAbs have produced unpleasant surprises in the clinic, it is usually because of insufficient grasp of the biology of the target antigen, especially in murky, relatively unexplored areas." [3]*

Therefore, basic research concerning the molecular processes in diseased tissue is still necessary to develop individual therapies.

With the advent of experimental high-throughput techniques (e.g. microarray expression profiling) to screen samples on a large scale, it became possible to study cancer associated processes on different levels with bioinformatics approaches. In the following, we give an overview of the current statistical methods, available tools, and sophisticated algorithms for evaluating high-throughput data. A timeline and summary of the presented approaches is illustrated in Figure 1.2.

In the early stages, microarray studies tried to identify single differentially expressed genes to explain the observed differences between the investigated conditions. However, when differences in gene expression are marginal, the task to distinguish the key players of the alterations from noise is very difficult. To this end, so-called Gene Set Analysis (GSA) approaches have been developed taking into account that genes do not act individually but in a coordinated fashion. In general, these approaches compute a score for a list of genes that is dependent on their expression values and their occurrence in a pre-defined biological category (e.g. the genes of a biochemical pathway) and estimate the significance of this score with permutation tests. The most popular of these methods known as Gene Set Enrichment Analysis (GSEA) was developed by Mootha et al. [4], which is similar to the simultaneously proposed method by Lamb and coworkers [5]. Furthermore, comparable methods such as SAM-GS [6] and the maxmean approach of Efron and Tibshirani [7] exist, which differ in the computation of scores for estimating the significance. For a detailed review and comparison of gene set enrichment methods the interested reader is kindly referred to [8].

Most of these GSA methods or variations thereof have been integrated in various web-based applications or downloadable programs like FatiScan [9], GeneTrail [10], GSEA-p [11], and SAM-GS [6]. The demand on computational biology for the development of easy-to-use analysis methods and applications for evaluating high-throughput data is also mirrored in the vast amount of publications describing such tools in recent years (reviewed in [12, 13]).

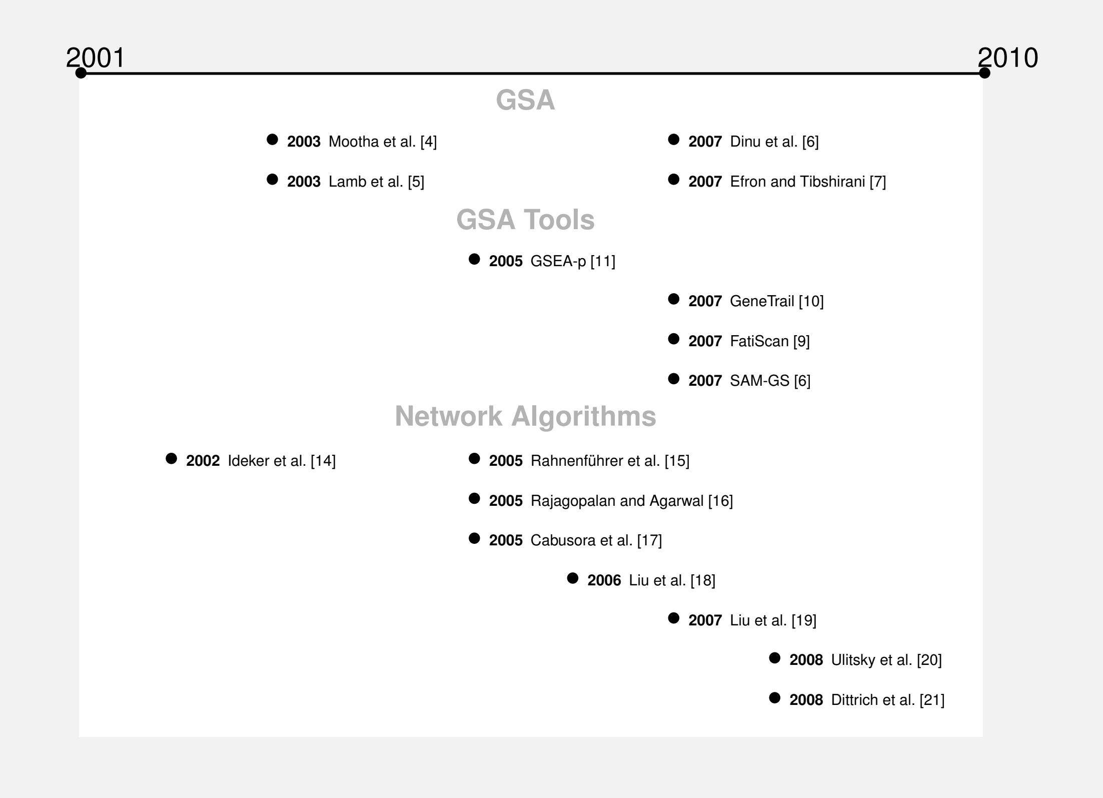
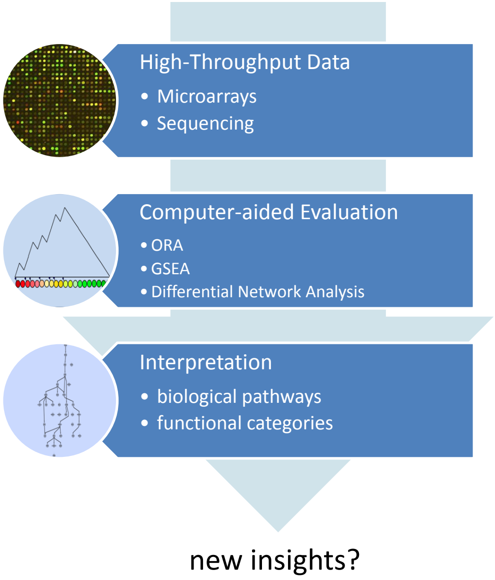


Figure 1.2: Timeline of presented approaches and algorithms

However, the above described GSA methods can only reveal the enrichment of genes in pre-defined gene sets, e.g. canonical biological pathways. Therefore, the research focus has shifted towards analysis methods that integrate topological data mirroring the biological dependencies and interactions between the involved genes or proteins. Several approaches for integrating network and gene expression data are described in the literature. Ideker et al. proposed a method for the detection of active subnetworks by devising a scoring function and an algorithm for detecting high-scoring subnetworks [14]. Similar methods, which are based on scoring networks given experimental data, have also been published by other groups [15–17, 22]. Additionally, topology-based classification technologies have been successfully applied to cancer [18, 23]. Recently, Liu et al. published a method called 'Gene Network Enrichment Analysis', which is similar to standard 'Gene Set Enrichment Analysis' and applies hypothesis testing to evaluate pathways [19]. In 2008, Ulitsky and coworkers presented an algorithm for detecting disease-specific dysregulated pathways by using clinical expression profiles [20]. Since the underlying combinatorial problem of finding high-scoring subnetworks is NP-hard, usually all described approaches use heuristics to solve this problem. By contrast, Dittrich et al. devised the first approach to solve the maximal-scoring subgraph problem optimally by integer-linear programming (ILP) [21].

In this work, we focus on the development and application of tools and algorithms for the evaluation of high-throughput data to contribute to a better understanding of cancer as depicted in the general information flow in Figure 1.3. To this end, we developed and implemented the gene set analysis framework GeneTrail [10]. GeneTrail has evolved since its publication in 2007 to one of the most comprehensive web-based applications due to its statistical capabilities, the variety of biological categories available for analysis, and its sophisticated graphical output for displaying results. Our tool has been designed for the evaluation of high-throughput data, but is not limited to a special type of experimental data, since the input consists solely of lists of “interesting” genes. To facilitate the usage of experimental raw data like microarray expression values, we added a pre-processing pipeline for this type of data, called GeneTrailExpress [24]. GeneTrailExpress provides comprehensive normalization and scoring functions for pre-processing microarray data. The processed data is directly passed to GeneTrail for statistical evaluation in an extensive gene set analysis.


High-Throughput Data

- Microarrays
- Sequencing

Computer-aided Evaluation

- ORA
- GSEA
- Differential Network Analysis

Interpretation

- biological pathways
- functional categories

new insights?

Figure 1.3: Information flow

Furthermore, since the GeneTrail C++ framework already supported information retrieval from the Biochemical Network Database BNDB [25], we extended our framework with a graph data structure to make use of the network topology. Using this functionality, we developed two approaches for detecting differentially regulated components of a regulatory network. The first approach, called FiDePa (Finding Deregulated Paths) [26], is a dynamic programming algorithm and relies on a statistical test similar to a standard gene set enrichment analysis. The results of the FiDePa algorithm are the most significant paths of a chosen length. Applying FiDePa to expression profiles of 100 high-grade glioma samples in comparison to 158 profiles of normal tissue samples, we demonstrated that it is possible to derive patient specific deregulated subnetworks from the union of computed significant paths. Our second approach is an ILP algorithm that reveals the most deregulated subnetwork of a certain size. The computed subnetwork is furthermore rooted in a special node from which all other nodes in the subnetwork are reachable. This way, it is much easier for researchers to interpret the resulting subnetworks and to verify whether this node can serve as a potential target for therapies. We employed the ILP algorithm on expression profiles of BRCA1 mutation carriers and non-mutation carriers. Our evaluation indicates that oxidative stress plays an important role in epithelial cells of BRCA1 mutation carriers that may contribute to the later development of breast cancer.

Beside the development of bioinformatics tools for gene set analysis, we attach great importance in this thesis to the application of those tools to different fields of cancer research. One of those fields focuses on miRNAs and their putative targets in the context of cancer. While the expression of gene coding mRNAs was the primary target of research over the last three decades, the emphasis of research shifted lately to the analysis of non-coding RNAs, especially so-called microRNAs (miRNAs). These miRNAs play a crucial role in regulating gene expression, e.g. through binding to mRNA and enabling the degradation or silencing of their target mRNAs [27]. Furthermore, their function as potential tumor suppressors or oncogenes has been demonstrated [28, 29]. To further elucidate the methods of action of miRNAs in cancer, we performed a comprehensive study of different cancer expression profiles which showed that targets of specific miRNAs were significantly enriched or depleted in these sets. Our findings confirm the important role of miRNAs as key players of gene regulation in cancer.

Another field of cancer research covered in this thesis is the immunogenicity of tumor associated antigens (TAAs). TAAs have the potential to function as biomarkers for early detection of human neoplasms [30]. However, the reasons why these antigens become immunogenic remains for the most part elusive. In this work, we test different hypotheses as causes for the immunogenicity, e.g., if mutations, SNPs, or similarities to proteins in other organisms play a role. Our results suggest that there is a certain prevalence of sequence similarities to proteins of other organisms in the tested antigen sets, which may be a possible cause why the autoantibody repertoire seems restricted to a limited number of self-proteins.

Taken together, we focus on the detection of molecular changes in cancer and their effects on the level of mRNA expression, miRNA expression, the immune system, and related regulatory networks. Our tools and algorithms have been successfully applied by our group and researchers worldwide to contribute to a better understanding of cancer. This knowledge may help identifying mechanisms of disease origin, progression, and ultimately detecting new starting points for individual therapies.

This thesis is structured as follows: in the next chapter we give a detailed overview of the bioinformatics tools we implemented. Afterwards, we describe in Chapters 3–5 our findings concerning tumor associated antigens, miRNAs and cancer, and our pathway and network algorithms for finding deregulated paths/subgraphs in regulatory networks. Finally, this thesis concludes with a summary of our contribution to cancer research.

CHAPTER 2

TOOLS FOR CANCER RESEARCH

In this chapter, we introduce the newly developed bioinformatics tools for cancer research that have been implemented in the course of this thesis. As previously mentioned in the introduction, computer-aided methods for the statistical evaluation of high-throughput data are nowadays essential for gaining insights into complex molecular processes. In general, these high-throughput experiments serve to quantify changes in the genome and proteome in response to a given condition, e.g., in which way gene expression in a tumor tissue differs from gene expression in normal tissue. The subsequent challenge is to group, analyze, and interpret the vast amount of heterogeneous data provided by these methods. Computer-aided gene set analysis tools are tailored for grouping and analyzing such data by identifying significantly enriched functional categories, which facilitates the interpretation.

To study the enrichment of gene sets, two basic approaches have been developed. The first method, the so-called "Over-Representation Analysis" (ORA), compares the set of interest to a reference set. When considering a certain functional category, e.g. a Gene Ontology (GO) [31] term, this method tries to detect if this category is over-represented or under-represented in the respective set and estimates how likely this is due to chance. The second method is called "Gene Set Enrichment Analysis“ (GSEA) [4]. Here, the input set is sorted by some specific criteria (e.g. gene expression values). When considering an arbitrary functional category, GSEA tests if the genes in the set that belong to the category are uniformly distributed or accumulated on top or on bottom of the sorted input list. Additionally, the usage of other statistical tests like the Wilcoxon-Mann-Whitney test [32,33] or a Monte Carlo permutation test can be applied for evaluating whether the parent populations of two samples of observations (e.g., the number of SNPs in a test set compared to the number of SNPs in the reference set) are identical.

The demand on computational biology for the development of easy-to-use applications for evaluating high-throughput data is also mirrored in the vast amount of publications describing such tools in the recent years (reviewed in [12, 13]). However, most of these tools are either restricted to a certain type of experimental data or to a few biological categories. Gene Ontology based tools, e.g., FatiGO [34], BiNGO [35], and GOstat [36] to name a few, rank among the most frequent type of developed applications. Tools that focus on certain types of high-throughput data (e.g. microarray expression data) are ErmineJ [37], CRSD [38], or GSEA-P [39]. Furthermore, some tools, like Catmap [40], do not include biochemical categories and it is left to the user to define these categories. A few tool packages, however, allow for the analysis of different types of functional categories, e.g. WebGestalt [41] and Babelomics [42].

The central part of this thesis was the development of the comprehensive gene set analysis framework GeneTrail [10], which is described in detail in Section 2.1. GeneTrail is not only a user-friendly web-based online application, but has also evolved into a sophisticated C++ framework that efficiently combines information retrieval from various data sources, statistical evaluation, and a suitable presentation of the results, which are essential prerequisites for a state-of-the-art gene set analysis tool. In Section 2.2, we describe an extension of GeneTrail, called GeneTrailExpress [24], tailored for pre-processing data from microarray experiments. Finally, this chapter concludes with the description of GraBCas [43] in Section 2.3, a tool for the prediction of granzyme B and caspase cleavage sites, which has also been integrated in GeneTrail.

2.1 GeneTrail

In this section, we give a detailed description of the features of our gene set analysis framework GeneTrail [10]. GeneTrail has been developed to facilitate the statistical evaluation of arbitrary high-throughput data by providing support for ORA and GSEA approaches (described in Section 2.1.3). Our implementation of the unweighted GSEA method includes a novel algorithm that computes the correct p-value instead of estimating it by permutation tests. Since our tool relies to some extent on the comprehensive integrative system BN++ [44], GeneTrail allows the evaluation of a broad range of functional categories. The capabilities of the GeneTrail C++ framework are impressively demonstrated in the development of the so called FiDePa algorithm (see Section 2.1.4.6 and Chapter 5.1), which efficiently combines the information retrieval, construction of a regulatory network, and statistical evaluation of deregulated paths in this network. Furthermore, the online version of

GeneTrail provides a user-friendly interface and visualizes the computed results in a clear and concise manner (Section 2.1.5). GeneTrail has been developed in collaboration with Andreas Keller, who developed and implemented the dynamic programming algorithms for the unweighted GSEA method and FiDePa.

2.1.1 Workflow

GeneTrail is implemented in the programming language C++ and can be used either directly via command line or indirectly via the graphical user interface (GUI) that is written in PHP and accessible with a web-browser. However, the web-application offers not all functions that the command line version does. The input for GeneTrail consists of a list of interesting or sorted gene/protein IDs when performing an ORA or GSEA, respectively. In the former case, a reference set is additionally needed. The different input steps of the web-application are presented in Figure 2.1.

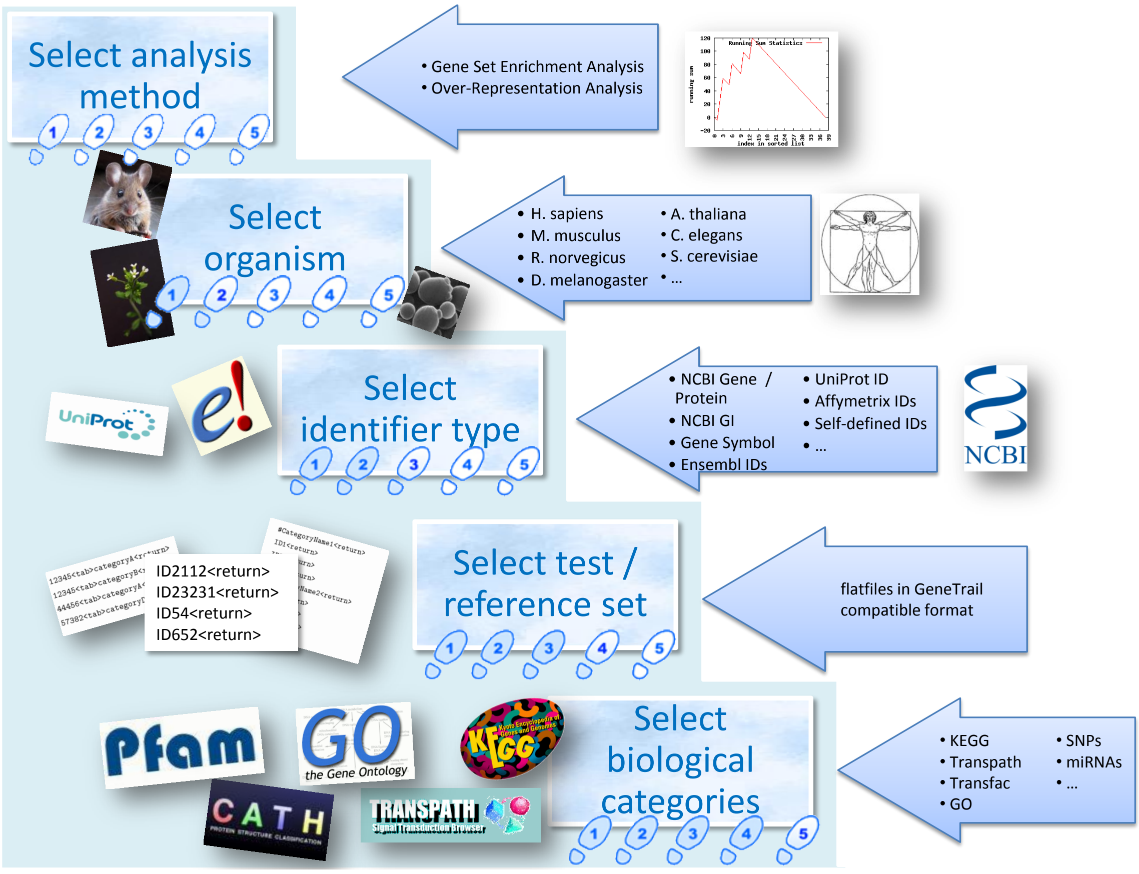


Figure 2.1: Overview of the subsequent input steps for the GeneTrail web-application

After successful computation of the statistical significance for the selected biological categories, an output in HTML, plain text, PDF, and XML is created illustrating the results.

2.1.2 Integrated resources

GeneTrail provides several pre-defined biological categories for statistical evaluation. In this section, we briefly summarize the most important data sources of these pre-defined categories. In addition, we give an overview of the supported gene/protein identifiers and organisms. Further details can be found in [10, 45] or in Appendix C.

2.1.2.1 Biological categories

The different biological categories supported by GeneTrail stem from various data sources, e.g., MySQL databases or downloadable flatfiles. For all of these data sources, we generate flatfiles in special formats (see Section 2.1.4.2) during the update process (Section 2.1.4.8). This guarantees a fast access to the information that is independent of the availability of a database connection or external resources. An overview of GeneTrail's integrated pre-defined biological categories is illustrated in Figure 2.2.

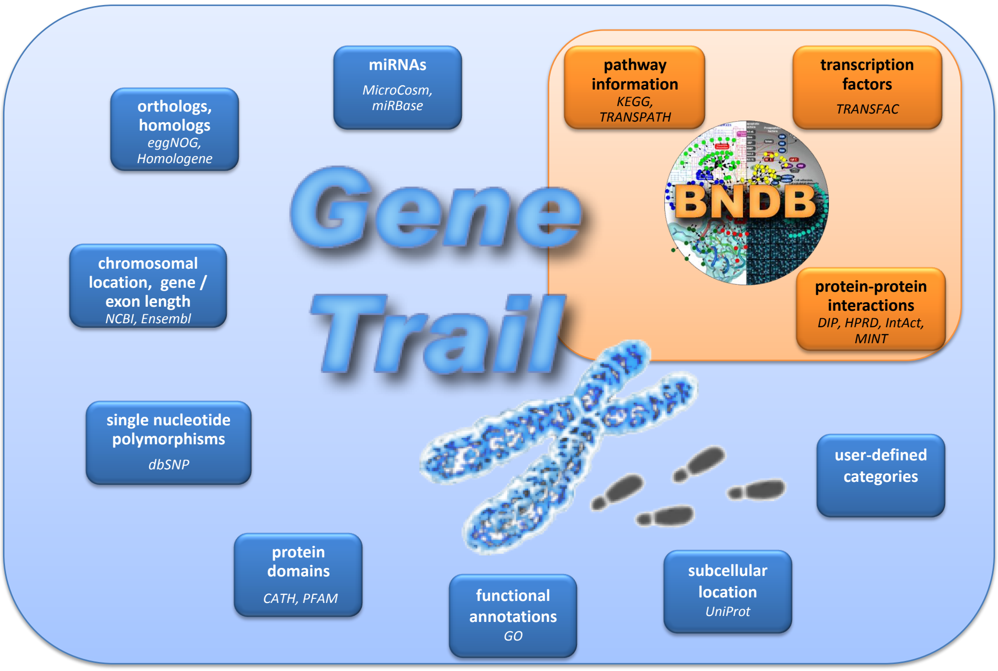


Figure 2.2: Pre-defined biological categories integrated in GeneTrail

BNDB: The Biochemical Network database (BNDB) [25] is part of the biological information retrieval system BN++ [44, 46]. BN++ is a C++ library tailored for modelling biochemical networks and is based on a comprehensive and easily extensible data model, called BioCore. The BioCore model has been implemented as C++ and Java framework, as presented by BN++ and BiNA (see also 2.1.5.1), respectively, and additionally as a relational database (BNDB). Figure 2.3 illustrates the architecture of BN++ and the coherences between the data model and its implementations.

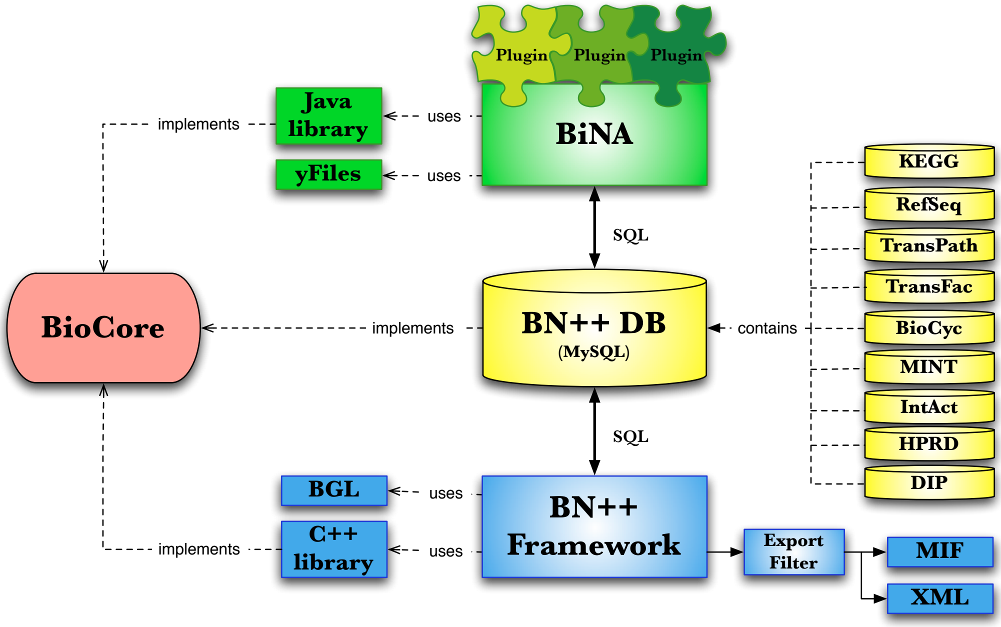


Figure 2.3: Architecture of BN++. Source: PhD thesis Jan Küntzer [46]

The BN++ framework provides importers for various databases, e.g. the pathway databases KEGG [47] and TRANSPATH [48], the transcription factor database TRANSFAC [49], and the protein interaction databases DIP [50], MINT [51], IntAct [52], and HPRD [53]. GeneTrail uses the BNDB as interface to retrieve the pathway and interaction information from these different databases. Hence, it is not necessary that GeneTrail itself has to support different database formats and importers. However, the disadvantage of using the BNDB is that building this database from external sources is very time-consuming. The download, import and merging of data can take several days and is not suitable for performing automatic and regular updates.

KEGG: The Kyoto Encyclopedia of Genes and Genomes[1] (KEGG) database [47] represents a comprehensive resource of metabolic and regulatory pathways. The KEGG database contains canonical pathways for different organisms. For using the pathway information from KEGG, we have implemented two variants. We can either access the information from KEGG using the BNDB or directly retrieve the information from the KEGG homepage via their SOAP interface. The second variant is preferred for doing updates on a regular basis when only the membership of genes/proteins to their respective metabolic or regulatory pathways is needed. If the topological information, in which way the genes or proteins interact, is required, we still use the BNDB.

Gene Ontology: The Gene Ontology (GO) [31] database consists of a controlled vocabulary that can be used to describe the attributes of a gene product in an organism. GO comprises three sets of independent vocabularies or ontologies: the molecular function, the biological process, and the cellular component. A gene product can be associated with one or more GO terms and belong to different GO ontologies. The structure of the GO hierarchy can be visualized by a directed acyclic graph (DAG). For GeneTrail, we use a local version of a current MySQL dump of the GO database.

MicroCosm: One of the more recent extensions of GeneTrail's pre-defined categories comprises miRNAs and their putative target genes. We integrated the MicroCosm targets[2] (formerly miRBase targets) that are identified by the miRanda algorithm [54–56]. The putative miRNA targets are labelled with a p-value which is an estimated probability of the same miRNA family hitting multiple transcripts for different species in an orthologous group. The lower this p-value, the more specific are the predicted targets of a miRNA for an organism. In GeneTrail, we included three miRNA target thresholds: p-value < 0.01, p-value < 0.001, p-value < 0.0001. An extensive analysis of putative targets of miRNAs and their pathways in cancer is described in Chapter 4.

User-defined categories: In addition to GeneTrail's pre-defined biological categories, we provide the possibility to use GeneTrail's statistical evaluation capabilities with user-defined categories. We support several standard gene set formats as the .gmx and .gmt format, as well as a simple self-defined format (.gtf) besides the formats that GeneTrail uses internally. The different file formats are described in more detail in Section 2.1.4.2.

[1] `http://www.genome.jp/kegg/`

[2] `http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/`

2.1.2.2 Supported gene/protein identifiers

GeneTrail was designed to analyze sets of genes. Due to the various biological databases, many different accession numbers for a single gene exist. We decided to use NCBI Entrez Gene IDs (see also Appendix C.1) as central gene identifier for GeneTrail, because these IDs are unique for one gene and are closely connected to other information and identifiers made available by NCBI. Additionally, this data can be downloaded as flatfiles from NCBI's ftp-server, which is regularly updated.

Besides the different NCBI identifier types (Entrez Gene, RNA/Protein RefSeq, RNA/Protein GI, UniGene), we additionally support the official gene symbols from the HUGO Gene Nomenclature Committee[3], UniProt accession numbers, and Ensembl Gene/Protein IDs. Furthermore, we provide transcript IDs for different popular Affymetrix microarray platforms and the Amersham Whole Genome Human array. However, not all of these ID types are available for all supported organisms. For some organisms we added special IDs that are only available for a specific organism, e.g. the TAIR[4] IDs for *A. thaliana* or the SGD[5] ORF IDs for *S. cerevisiae*.

As described above, GeneTrail can evaluate user-defined categories. To further enhance this capability, the user-defined categories can consist of user-defined IDs instead of the standard supported IDs. This enables the user to be completely independent of the provided IDs and categories in GeneTrail. However, in this case, no mapping to NCBI Gene IDs can be performed, and therefore, this identifier type is not applicable with the predefined biological categories.

Additional information about external data sources that are integrated in GeneTrail can be found in Appendix C.

2.1.2.3 Organisms

GeneTrail covers a large number of model organisms to make efficient gene set analysis available to a broad clientele of users working in different research areas. Until now, we have added support for *Homo sapiens*, *Saccharomyces cerevisiae*, *Mus musculus*, *Rattus norvegicus*, *Staphylococcus aureus N315*, *Corynebacterium glutamicum ATCC 13032*, *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Aspergillus fumi-*

[3] `http://www.genenames.org/index.html`
[4] `http://www.arabidopsis.org/`
[5] `http://www.yeastgenome.org/`

gatus Af293, and *Danio rerio*. Of course, not all pre-defined biological categories and IDs are available for all these organisms. However, we try to provide as many analyses for the supported organisms as possible. KEGG, GO, and MicroCosm targets are the best covered biological categories for these organisms. The most pre-defined biological categories are available for *H. sapiens*.

2.1.3 Statistics

In this section, we introduce the statistical tests provided by GeneTrail. First, we present the general approaches ORA and GSEA. In the latter case, we will also describe the dynamic programming algorithm for the computation of exact p-values in more detail. Additionally, we provide information about the Wilcoxon-Mann-Whitney (WMW) test and the Monte Carlo permutation test that can be applied to non-binary biological categories. At last, we describe the implemented methods for performing a multiple testing adjustment.

2.1.3.1 Over-Representation Analysis

The ”Over-Representation Analysis“ (ORA) compares a set of interesting genes (test set) to a background distribution (reference set) concerning a certain biological category (e.g. a canonical pathway). The ’interesting’ genes are determined in a selection step after performing a high-throughput experiment (Figure 2.4). Given a set of n test set genes, of which k belong to a category C, and a reference set of m genes, of which l belong to C. The probability to find exactly k genes that belong to C can be modelled as a sampling problem without replacement for a random variable X using the formula of the hypergeometric distribution:

$$P(X = k) = \frac{\binom{l}{k}\binom{m-l}{n-k}}{\binom{m}{n}}$$

Since l elements of the reference set belong to C, we expect to find $E(X) = \frac{n \cdot l}{m}$ elements in the test set belonging to C. If $k > E(X)$, C is said to be enriched or over-represented, otherwise C is said to be depleted or under-represented in the test set. Depending on the expectation value, we can compute a one-tailed p-value for the probability of having at most (at least) k genes belonging to C:

$$\text{p-value} = \begin{cases} \sum_{i=k}^{n} \frac{\binom{l}{i}\binom{m-l}{n-i}}{\binom{m}{n}} & ifk > E(X) \\ \sum_{i=0}^{k} \frac{\binom{l}{i}\binom{m-l}{n-i}}{\binom{m}{n}} & ifk \leq E(X) \end{cases}$$

If the computed p-value is smaller than the previously defined α-level, we consider the result as significant.

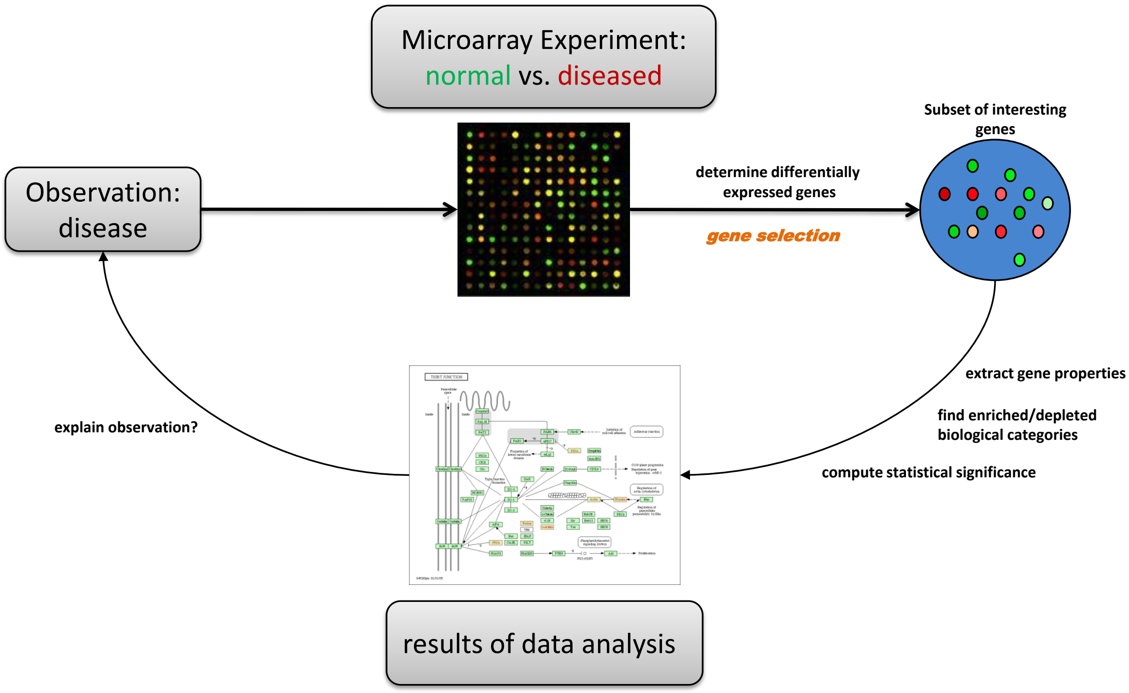

Microarray Experiment:
normal vs. diseased
Subset of interesting genes
Observation: disease
determine differentially expressed genes
gene selection
extract gene properties
find enriched/depleted biological categories
compute statistical significance
explain observation?
results of data analysis

Figure 2.4: Workflow of an ORA. After performing a high-throughput experiment, e.g., a microarray analysis, a gene selection step follows that filters genes meeting a certain threshold (e.g. genes that are 2-fold over-expressed). The selected genes are compared to the background distribution (e.g. all genes on the microarray) concerning certain biological categories. For enriched or depleted biological categories, the significance is estimated by a suitable statistical test. If the significant categories give new insights concerning the observations of the initial microarray analysis, more refined experiments can be performed and the process starts over.

2.1.3.2 Gene Set Enrichment Analysis

In contrast to the ORA, the 'interesting' genes for a Gene Set Enrichment Analysis (GSEA) are not selected by some arbitrary threshold, but are altogether sorted by a criterion that mirrors the differences in expression between the investigated states (Figure 2.5). For computing the statistical significance of an arbitrary biological category C given a sorted list of genes of size m, we apply the so-called unweighted GSEA as proposed by Lamb et al. [5]. Using a Kolmogorov-Smirnov-like test that computes whether the genes in C are equally distributed in the sorted list or accumulate on top or on bottom of the list, we determine if the considered category is significantly enriched or depleted. If l genes of the sorted list belong to C, we compute the running sum by processing the input list, adding $m - l$ to the running sum if the considered gene belongs to C, or subtracting l otherwise. This means, we sum up $l \cdot (m - l)$ for the genes in C in total and $(m - l) \cdot (-l)$ for the genes not in C in total. Therefore, the running sum's final value will always be zero and we can reach a maximal possible sum of $l \cdot (m - l)$ and a minimal possible sum of $(m - l) \cdot (-l)$. The value of interest is the running sum's maximal deviation from zero, denoted as RS_C. An example of the procedure is provided in Figure 2.6. The significance value (p-value) is computed as the probability that any running sum reaches a greater absolute value than RS_C. Such a probability can either be approximated with permutation tests or exactly calculated by a dynamic programming algorithm that computes the exact number of possible running sum statistics with greater deviation than RS_C as described in the following. We adopt here the presentation of concepts from our BMC Bioinformatics publication [57].

Computation of exact p-values for unweighted GSEA

As mentioned above, the p-value for a GSEA can be computed by nonparametric permutation tests, i.e., RS_C is calculated for permuted gene lists and compared to the value of the original list. In general, two ways have been proposed for performing permutation tests in this case. First, the sorted gene list can be randomly permuted. Second, if the list is sorted by the median expression quotient of expression values in one group divided by the median expression value in another group, the samples are randomly assigned to the two groups, the median fold quotient of the new groups is computed and thereby permuted gene lists are generated. Notably, these methods do not always yield the same results. The permutation procedure is repeated t times and the running sum statistics together with the corresponding maximal deviations from zero, denoted as RS_i, $i \in 1, ..., t$, are computed. Usually, the p-value is computed as the fraction of RS_i values that are larger or equal than the original RS_C value:

$$\text{p-value} = \frac{1}{t} \sum_{i=1}^{t} I(RS_i \geq RS_C)$$

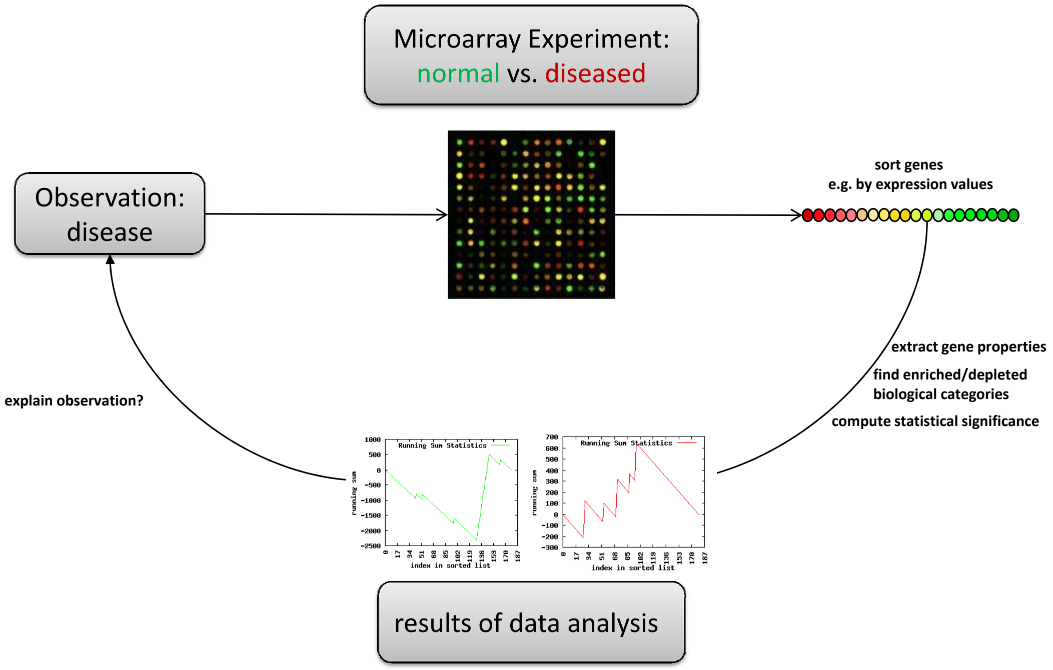


Figure 2.5: Workflow of an GSEA. After performing a high-throughput experiment, e.g., a microarray analysis, the genes are sorted by a criterion that mirrors the differences in gene expression between the considered states. This sorted list serves directly as input for a GSEA, which determines if genes that belong to a certain biological category are significantly accumulated on top or on bottom of this list. If the significant categories give new insights concerning the observations of the initial microarray analysis, more refined experiments can be performed and the process starts over.

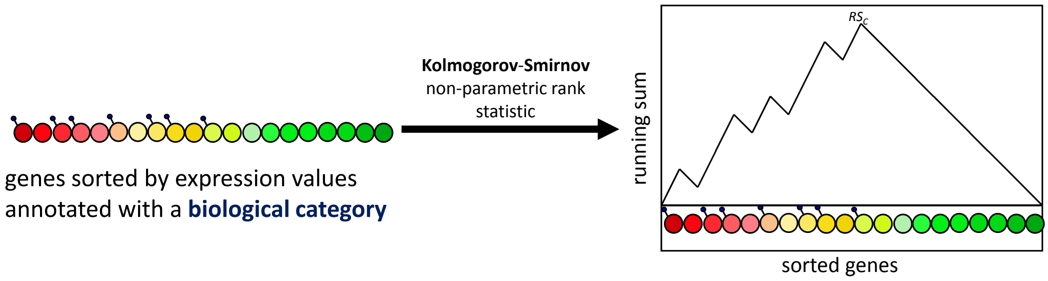


Figure 2.6: Example of a running sum statistic. The list of sorted genes is traversed from left to right. If a gene belongs to the considered biological category (genes having a ’flag’ in the picture), the running sum increases, otherwise it decreases.

Here, I is an indicator function:

$$I(RS_i \geq RS_C) = \begin{cases} 1: & RS_i \geq RS_C \\ 0: & RS_i < RS_C \end{cases}$$

Such permutation tests are widely used for estimating the significance, however, such tests entail three disadvantages:

First, repeated runs of the permutation test algorithm may lead to different significance values because of the random sampling.

Second, the permutation test procedure causes problems if the significance values are small. Given a running sum statistic whose true p-value is 0.00001. If, as usual, 1000 permutation tests are performed, probably none will have a higher maximal deviation as the original running sum statistics. According to the formula given above, the p-value would compute as 0 = 0/1000, which may be a bad estimation. Since GSEA is often applied to many biological categories, p-values have to be adjusted for multiple testing by suitable methods (e.g. Bonferroni, Benjamini & Hochberg). However, given the above estimation and the known multiple testing methods, the p-value cannot be adjusted in an appropriate way.

Third, it is difficult to estimate how many permutations should be performed to obtain a sample of reasonable size. Obviously, if $m = 20000$ and $l = 2000$, a sample size of 1000 permutations may be by far too small. Remarkably, the number of possible different running sum statistics amounts to $\binom{m}{l}$. On the example given above, the number of different running sums adds up to approximately $4 \cdot 10^{2821}$, emphasizing that 1000 permutation represent a very small sample. The required large number of permutation tests leads to an unacceptable computational effort, especially if thousands of biological categories are tested. An alternative, parametric method is the so called Parametric Analysis of Gene Set Enrichment (PAGE) method [58] that calculates a z-score for a given gene set and infers the significance value of this z-score against standard normal distribution.

In this section, we address the exact and efficient p-value computation for unweighted Gene Set Enrichment Analysis. Unweighted means that the number by which the running sum statistic is increased if a gene of C is found and the number by which the running sum statistic is decreased if the gene does not belong to C are constants. The dynamic programming method is similar to the ”DRIM” approach [59] that computes the optimal partition of a gene set in a target and a background set.

As mentioned before, the value of interest is the running sum's maximal deviation from zero, denoted as RS_C. The p-value can be computed as the probability that any running sum reaches a maximal deviation greater or equal than RS_C. We compute this probability via the complement of the event as:

$$\text{p-value} = 1 - \frac{X}{Y},$$

where X is the number of running sum statistics with a maximum deviation of at most $RS_C - 1$. Y is the number of all possible different running sum statistics that can be computed as $\binom{m}{l}$. To compute X, we count all running sum statistics that have a maximum deviation of at most $RS_C - 1$.

We use a matrix M of dimension $(2l(m-l)+1) \times (m+1)$, where the different rows represent all possible values of the running sum and the columns represent the indices of the sorted list from $1, ..., m$ and an initialization column with index 0. Let $M(j,i)$ denote the number of running sum statistics with value j in step i whose maximum deviation of zero is less than $RS_C - 1$. The entries of M are computed using dynamic programming, starting with the first column. $M(0,0)$ is set to 1 and all other values are set to 0.

We fill the matrix column by column, where the matrix entry $M(j,i)$ is recursively computed as:

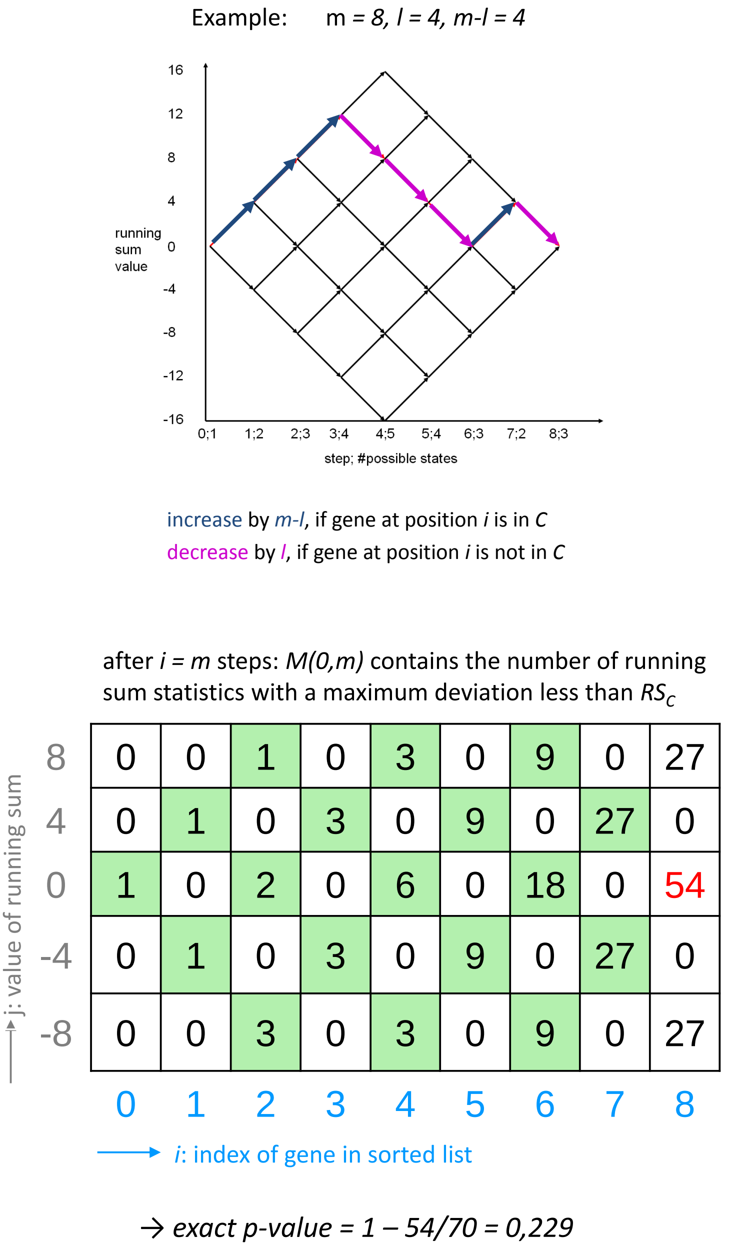
$$M(j,i) = \begin{cases} M(j-m+l, i-1) + M(j+l, i-1) & \text{if } (*) \\ 0 & \text{else} \end{cases} \tag{2.1}$$

where the constraint

$(*) - |RS_C| < j < |RS_C|$

ensures that only the running sum statistics with maximal deviation smaller than RS_C are counted. The total number of running sum statistics with maximum deviation smaller than RS_C can be found at matrix entry $M(0,m)$. An computation example is provided in Figure 2.7.

At first glance, the presented algorithm seems to be inefficient concerning both, space requirement and runtime, which are of order $O(m^2 l)$. For example, if $m = 20000$ genes and



Example: m = *8*, *l = 4*, *m-l = 4*

increase by *m-l*, if gene at position *i* is in *C*

decrease by *l*, if gene at position *i* is not in *C*

after *i = m* steps: *M(0,m)* contains the number of running sum statistics with a maximum deviation less than RS_C

8	0	0	1	0	3	0	9	0	27
4	0	1	0	3	0	9	0	27	0
0	1	0	2	0	6	0	18	0	54
-4	0	1	0	3	0	9	0	27	0
-8	0	0	3	0	3	0	9	0	27
	0	1	2	3	4	5	6	7	8

→ *exact p-value = 1 – 54/70 = 0,229*

Figure 2.7: Computation example with the dynamic programming algorithm for unweighted GSEA. The upper figure shows all possible running sum statistics for a sorted list of 8 genes of which 4 belong to an arbitrary biological category. The colored running sum statistic has an RS_C value of 12. Below, the corresponding dynamic programming matrix is depicted. Matrix entries unequal zero are highlighted in green. The matrix entries in the upper and lower right corner do not have to be computed due to the extended side constraints. The number of running sum statistics with a smaller deviation from zero (RS_C value) than 12 add up to 54. Given this result, the p-value can be computed and amounts to 0.229.

a functional category with $l = 2000$ genes is considered, M would have about $1.44 \cdot 10^{12}$ entries. As the recurrence Equation 2.1 implies, filling the ith column of M only requires the values of the $i - 1$th column. Thus, the dynamic programming approach requires only two columns of the matrix reducing the memory requirements to $O(ml)$. Additionally, the matrix M is sparse, i.e., it contains many entries of 0 and certain parts of M do not have to be computed at all as described in the following.

The running time of the algorithm can be further reduced by adding a second constraint

$(**) - m^2 + l \cdot m + i \cdot m - i \cdot l \leq j \leq l \cdot m - i \cdot l$

for each column i to the recurrence equation. The right side of the constraint holds because, for column i, the value j of the running sum can be computed as

$j = a \cdot (m - l) + (i - a) \cdot (-l)$

where a is the number of genes that belong to C up to index i in the ordered list. Since a can be at most l, the following inequality holds

$$\begin{aligned} j &\leq l \cdot (m - l) + (i - l) \cdot (-l) \\ \Leftrightarrow j &\leq l \cdot m - i \cdot l \\ \Leftrightarrow j &\leq l \cdot (m - i) \end{aligned}$$

Equivalently, for the left side of constraint $(**)$ and column i the following equation holds:

$$\begin{aligned} j &= (i - b) \cdot m - i \cdot l \\ \Leftrightarrow j &= -b \cdot m + i \cdot m - i \cdot l \end{aligned}$$

where b is the number of genes that do not belong to C up to index i in the ordered list. Since b can be at most $m - l$:

$$\begin{aligned} j &\geq -(m-l)\cdot m + i\cdot m - i\cdot l \\ \Leftrightarrow j &\geq -m^2 + m\cdot l + i\cdot m - i\cdot l \end{aligned}$$

Although the additional constraint does not lead to an asymptotically improved runtime, an increased performance has been measured, especially for small p-values.

The dynamic programming algorithm was integrated in GeneTrail with some minor adaptations. Since we are in general interested if a biological category is enriched or depleted, the algorithm was accordingly adjusted to compute an one-tailed p-value instead of the above described computation that corresponds to a two-tailed p-value.

2.1.3.3 Wilcoxon-Mann-Whitney test

The Wilcoxon-Mann-Whitney (WMW) test [32, 33] is a nonparametric statistical test for comparing the medians of two distributions. It is used to test the null hypothesis that two independent samples were drawn from the same population. For a sample S_1 of size m and a second sample S_2 of size n, a test statistic U can be computed as follows:

$$U = m\cdot n + \frac{m(m+1)}{2} - T, \tag{2.2}$$

where T is the rank sum of sample S_1.

The rank sum T is computed by sorting the values in both samples and summing up the resulting ranks for the values in sample S_1. For large samples, the distribution of the test statistic approximates the normal distribution, with known mean $\mu = \frac{m\cdot n}{2}$ and standard deviation $\sigma = \sqrt{\frac{m\cdot n(m+n+1)}{12}}$. Thus, we can compute a z-score:

$$Z = \frac{U-\mu}{\sigma}, \tag{2.3}$$

that can be directly used to determine the corresponding p-value with the cumulative standard normal distribution. The z-score expresses the divergence of U from the most probable result μ in numbers of standard deviations. The larger the value of the z-score is, the less probable is the value of the test statistic U due to chance. In the presence of ties (equal values in the samples), median ranks are assigned to these values. This influences the standard deviation that can be corrected as follows:

$$\sigma = \sqrt{\frac{m \cdot n}{N(N-1)} \cdot \left(\frac{N^3 - N}{12} - C\right)}, \tag{2.4}$$

where $N = m + n$ and $C = \sum_i (\frac{t_i^3 - t_i}{12})$ [60]. C is the so called tie correction and t_i is the number of values with equal ranks i. If there are no ties, C equals 0 and the standard deviation in Equation 2.4 equals the uncorrected standard deviation.

We included the WMW test in GeneTrail to be able to test non-binary categories (e.g. numbers of SNPs, gene length) for enrichment.

2.1.3.4 Monte Carlo permutation test

In addition to the WMW test, we implemented a permutation test to estimate the significance of an observed difference of means between two samples. In brief, we compute the difference of the means of the values in the test set (of size n) and the reference set. Then, the test set and the reference set are combined to one common set, from which we randomly draw n values. This way, we obtain a new distribution of the values in a new test and reference set. We compute the differences of means of the values in the two new sets. This procedure is repeated at least 1000 times. A p-value can be computed by counting the number of resampled differences with a better score than the difference of the original sets and dividing by the number of permutations.

2.1.3.5 Multiple testing adjustment

The multiple testing corrections are used when several independent statistical tests are performed simultaneously. When testing many hypotheses, the probability for false positive predictions increases. If we perform a testing of n independent hypotheses to a specified significance level α, we can expect $n \cdot \alpha$ hypotheses to be significant by chance. Therefore, an adjustment of p-values is necessary for multiple hypotheses testing. We implemented two p-value correction algorithms in GeneTrail, the Bonferroni and the false discovery rate (FDR) correction.

The Bonferroni correction

The Bonferroni p-value correction is a very conservative method, which means that the reduction of false positives is bought with an increase of false negatives. This method is often too restrictive for biological questions because of the high information loss.

The Bonferroni adjustment can be performed in two ways: Either the significance level is adjusted by dividing by the number of tested hypotheses or alternatively, the p-values can be adjusted after computation with the unadjusted significance level by multiplying with the number of tested hypotheses. The Bonferroni correction thereby controls the probability of committing any type I error considering all tests.

The FDR correction

The FDR correction was developed by Benjamini and Hochberg (1995) [61]. Instead of controlling the probability of committing any type I error, the FDR controls the expected proportion of errors among the rejected null hypotheses and is therefore less strict than the Bonferroni method.

$$FDR = E\Big(\frac{\text{number of falsely rejected null hypotheses}}{\text{number of rejected null hypotheses}}\Big)$$

The FDR controlling procedure works as follows:

Let $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}$ be the ordered p-values for the tested hypotheses $H_1, H_2, ..., H_m$ and their corresponding p-values $p_1, p_2, ..., p_m$, and denote by $H_{(i)}$ the null hypothesis corresponding to $p_{(i)}$. To control FDR at level α, reject the hypothesis $H_{(j)}$ for $j = 1, ..., j^*$, where $j^* = \max\{j : p_{(j)} \leq \frac{j}{m}\alpha\}$.

Adjusted p-values can be computed as follows [62]:

$$p^*_{(j)} = \min_{k=j,...,m} \Big\{ \min\Big(\frac{m}{k} p_{(k)}, 1\Big)\Big\}$$

This method controls the FDR for independent test statistics as well as under certain dependence structures (positive regression dependency) as shown by Benjamini & Yekutieli (2001) [63].

2.1.4 C++ framework

The GeneTrail C++ framework provides all necessary components for performing efficient gene set analyses comprising information retrieval, data integration, statistical evaluation, result presentation, and data exchange. In the following, we will present the basic concepts of the GeneTrail data model and its implementation. Furthermore, we discuss some special features of GeneTrail and the way of extending GeneTrail's integrated biological categories, identifier types and organisms.

2.1.4.1 Data model

In this section, we briefly describe the most important base classes of GeneTrail and their function. The class `DataObject` is the parent class of GeneTrail's internal data structures. We implemented object-oriented data structures for gene set analyses that are hierarchically constructed (see Figure 2.8). The `Parameter` class provides parsing of the command line options, collects all necessary information for performing the statistical evaluation, and finally contains the results of the analysis. An instance of `Parameter` stores the `Testset`(s) and the `Referenceset`. These data sets contain the information which genes belong to them and the categories that are statistically evaluated. For each biological category that is analyzed an instance of `Category` is created which itself is filled with the different subcategories. A `Subcategory` has crosslinks to the genes of the data set that are contained in this subcategory. The statistical evaluation is performed for each subcategory and the computed raw p-value or adjusted p-values are stored within the corresponding `Subcategory` instance. The class `Analysis` is the central base class for all derived specific analyses and provides methods for filling the `Category` data structure by parsing the corresponding flatfiles. These methods can be overridden if necessary in the derived classes. Furthermore, the specific derived analysis classes comprise methods for generating and updating their flatfiles. The class `Statistics` is the base class for all statistical evaluations. The derived classes are responsible for a specific statistical test. Furthermore, the base class provides methods for performing the multiple testing adjustment. For the result presentation and data exchange, we have implemented the base class `Serializer` and its derived classes that provide the options to serialize the data structures in different data formats as XML, HTML, LaTeX, and plain text.

A class diagram of GeneTrail's most important classes can be found in Appendix B. The red classes are specific analysis classes derived from the `Analysis` base class. The

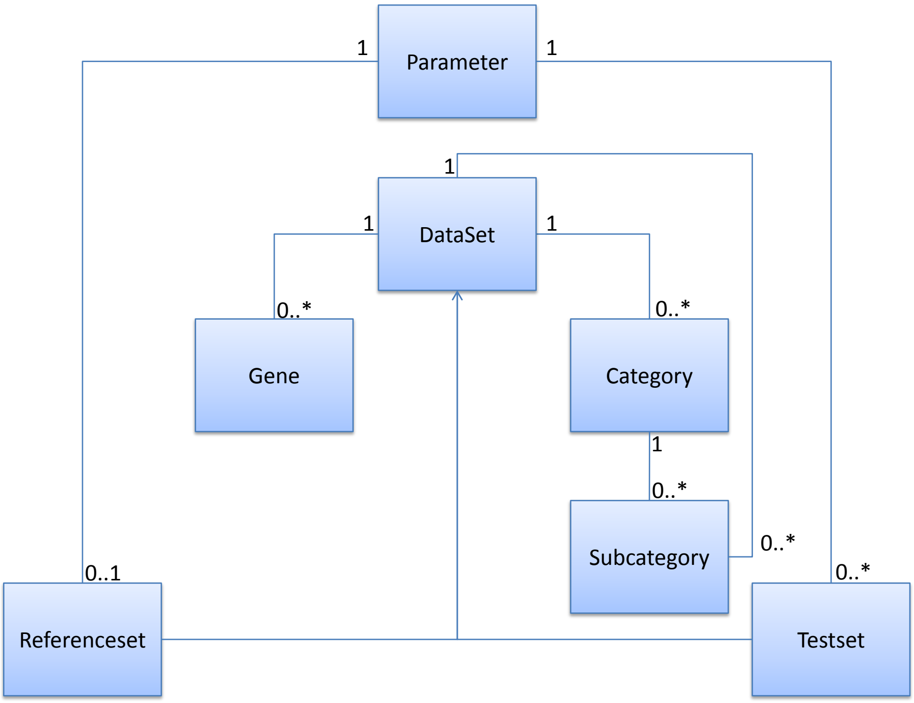


Figure 2.8: Simplified UML diagram of the internal data structures in GeneTrail

statistics classes derived from the base class `Statistics` are colored in yellow. Green colored classes derived from `DataObject` represent the internal data structure classes in our model. The classes colored in blue are responsible for generating the output of the results of the analyses in different file formats. The classes derived from the `Output` base class are obsolete and are replaced by the new `Serializer` derived classes. The latter classes are better integrated in the data model and GeneTrail's data structures and can be more easily extended for (de-)serialization of various file formats.

2.1.4.2 File formats

GeneTrail handles different file formats as input that are adjusted to their usage sites. We describe here the simple file formats for test and reference sets, identifier mapping files, binary and non-binary biological category files, and the supported standard gene set file formats.

File format for test / reference sets: The input file format for test and reference sets consists of a plain text file that contains one ID per line.

```
23744<return>
51<return>
872<return>
```

All identifier types can be used that are supported by GeneTrail for the selected organism. In any case, we insist that for all test and reference sets that are involved in an analysis the same ID type is used.

File formats for pre-computed categories and identifier mapping: Depending on the type of statistical evaluation, we use different file formats for the pre-defined and user-defined biological categories in GeneTrail. For binary categories, we designed a simple file format containing the Gene ID and tab-separated the category this gene belongs to, e.g.:

```
12345<tab>categoryA<return>
12345<tab>categoryB<return>
44456<tab>categoryA<return>
57382<tab>categoryD<return>
```

All pre-computed binary biological categories are deposited in this format during the update process (Section 2.1.4.8). We name these flatfiles according to this convention: <analysis>_annotated_<opt>_<org>_<taxid>.txt, where <analysis> is the biological category, <org> the three letter code for an organism, and <taxid> the NCBI taxonomy ID for the organism. The <opt> parameter is only needed in special cases, e.g. we provide for GO flatfiles containing either all annotations (<opt> = "all") or only manually curated annotations (<opt> = "manu"). This file format is also utilized for mapping Gene IDs to other identifier types. In this case, the naming convention is the following: map_geneid_<id>_<org>_<taxid>.txt. For analyzing user-defined (binary) categories, we support three different additional file formats:

- The GeneTrail format (.gtf): The category name starts with a #-sign; the IDs are listed underneath their category separated by newline symbols

  ```
  #CategoryName1<return>
  ID1<return>
  ID2<return>
  ID43<return>
  #CategoryName2<return>
  ```

```
ID23<return>
ID2<return>
ID54<return>
ID4<return>
```

- The gene matrix transposed file format (.gmt): Each row in this file format represents a gene set. The first column consists of the gene set names, the second column can contain a description for the gene set, the remaining columns comprise the genes belonging to the gene set. All columns are delimited by tabs.

```
CategoryName1<tab>na<tab>ID1<tab>ID2<tab>ID43<return>
CategoryName2<tab>na<tab>ID23<tab>ID2<tab>ID54<tab>ID4<return>
```

- The gene matrix file format (.gmx): Each column in this file format represents a gene set. The first row consists of the gene set names, the second row can contain a description for the gene set, the remaining rows comprise the genes belonging to the gene set. All columns are delimited by tabs.

```
CategoryName1<tab>CategoryName2<return>
 na <tab> na <return>
ID1 <tab>ID23<return>
ID2 <tab>ID2 <return>
ID43<tab>ID54<return>
    <tab>ID4<return>
```

In the case of non-binary categories, where the genes are mapped to a number (e.g. number of SNPs for a gene or gene length), we need to use a different format. This kind of data is deposited in a tab-delimited matrix format, where the first column contains the Gene IDs, the first row the category names, and at position (i, j) of the matrix the numerical property of gene i for category j.

```
GENEID<tab>SNPs in exons<tab>SNPs in introns<tab>SNPs in total<return>
12345<tab>       4      <tab>      2         <tab>     6       <return>
44456<tab>      23      <tab>     10         <tab>    33       <return>
57382<tab>       8      <tab>      5         <tab>    13       <return>
```

The naming convention for these non-binary flatfiles is similar to the above, but we replace ”annotated“ with ”flatfile“: <analysis>_flatfile_<opt>_<org>_<taxid>.txt

2.1.4.3 Information retrieval: MySQL databases

GeneTrail supports a wide range of biological categories. The original data sources of these categories are present in various data formats or deposited in databases. For the latter, we integrated the information retrieval from MySQL databases in GeneTrail. Up to now, it was sufficient to provide an interface for MySQL databases, but this functionality can be extended to other relational database systems as Oracle, DB2, or PostgreSQL, since we are using Qt[6] as external library that supports all major database drivers. An analysis class that is dependent on a database connection is derived from the class `DatabaseDerived` in our data model that provides the necessary functions for establishing a database connection and querying the database.

2.1.4.4 Serialization concept

Since we are storing the results of the computation in the internal data structures of GeneTrail's C++ framework, we need a way to output the results in a user-friendly format. To this end, we integrated a serialization concept that is capable to output the data structures in various formats. So far, we implemented serializer for XML, HTML, plain text, and LaTeX. For XML, we also provide a deserializer that re-creates the data objects in memory when parsing an XML file in our own format. This is a very useful feature, because we can, e.g., filter the information later on or re-compute the p-values if some thresholds have changed. The serialization concept is realized in a way that allows for easy modification of our data structure classes and facilitates adding new serializers. In brief, each data structure class tells the serializer which information should be serialized. The serializer itself does not need to know about the internal structure of a data structure class and is, therefore, independent of the data structure classes. If new information is added to a data structure class, only this class has to be adapted, the serializer does not need to be changed.

2.1.4.5 Graph data structure for topology usage

The GeneTrail C++ framework supports information retrieval from the BNDB [25] as described previously in Section 2.1.2.1. In order to take full advantage of the provided information, we extended our framework with a graph data structure using the efficient boost graph library (BGL) [64]. We implemented the generation of a compound graph, a bipartite

[6] `http://qt.nokia.com/products/library/modular-class-library`

reaction graph, and an event graph from the data in BNDB as adapted from the original code in BN++. The different graph representations of a simple metabolic reaction are illustrated in Figure 2.9.

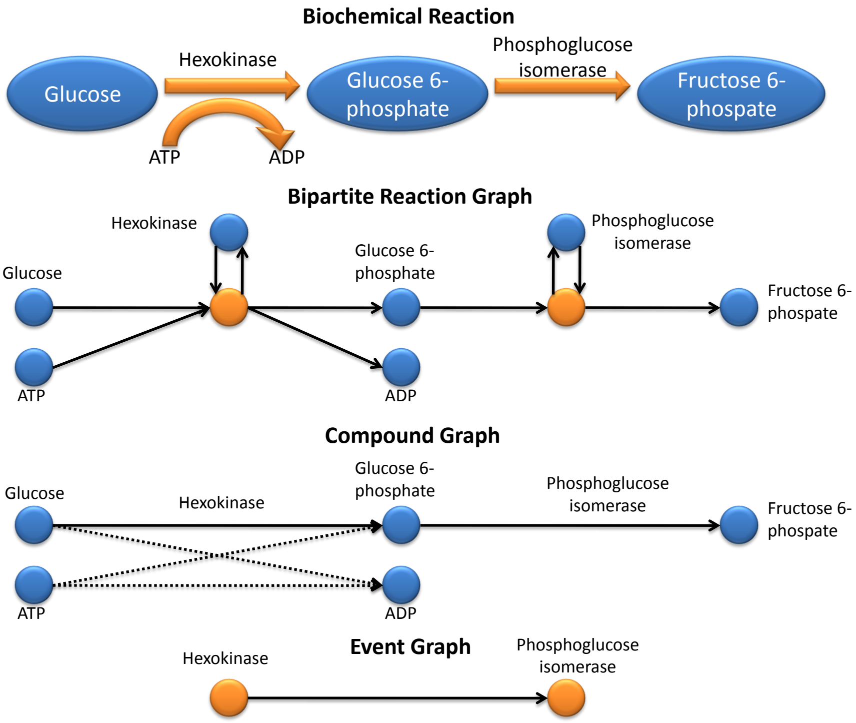


Figure 2.9: The different graph representations of the first two steps in the glycolysis pathway. Blue nodes depict participants, orange nodes events as, e.g., reactions. The edge direction is derived from the role the participant is playing in the event. In the compound graph representation the dashed arrows illustrate the optional edges for side educts and side products. Adapted from [46].

The `Topology` class and its implementation for building a compound graph for the regulatory network derived from KEGG has been applied in our ILP approach for finding deregulated subnetworks using expression profiles (described in detail in Chapter 5.2). The nodes in this network correspond to proteins, protein families, or protein complexes, the edges represent either directed reactions, e.g., an activation or inhibition, or interactions that are undirected for which we add two directed edges in both directions. Our graph class provides the possibility to split protein families and protein complexes that contain protein families into their components if desired. Given a protein family, we replace the family node by a set of nodes where each node represents a family member. Each new node is connected to all neighbors of the original family node, i.e., it has the same set of incoming and outcoming edges as the original family node. Here, we assume that all family members interact in the same manner with the neighboring nodes of the original family node.

Furthermore, our Topology class provides some additional features. We support the Cytoscape [65] format for networks (.sif), as well as for node and edge attributes (.na, .ea). We can output a graph in this format and reconstruct the graph when reading back the output. For analyzing network characteristics, we implemented some measures that are often applied: computation of the graph diameter, clustering coefficient, average distance, degree distribution. Additionally, we provide methods for filtering the original graph and extracting subgraphs, computation of connected components, and converting directed to undirected graphs to mention a few. Furthermore, we can map a selection of genes/proteins to the graph nodes and compute the resulting subnetwork of shortest paths between the selected nodes.

2.1.4.6 Combination of topology and statistics: FiDePa

To demonstrate the capabilities of our GeneTrail C++ framework, we combined the information retrieval, the network topology, and the statistics of the unweighted GSEA dynamic programming algorithm to develop a novel method for finding deregulated paths (FiDePa). As described in Section 2.1.3.2, we integrated a variant of unweighted GSEA in GeneTrail to compute a p-value given a sorted list of input genes and a biological category. With this test we can verify whether there is a significant enrichment or depletion of the biological category, which means that the genes belonging to the category are accumulated on top or on bottom of the sorted list. The disadvantage of this method is that we can only test predefined biological categories. FiDePa identifies deregulated paths in regulatory networks. These deregulated paths can consist of different parts of various canonical pathways that are connected in the regulatory network. The FiDePa algorithm and an application of FiDePa to glioma expression profiles is described in detail in Chapter 5.1.

2.1.4.7 Testing

The GeneTrail C++ framework must ensure that the results when performing statistical evaluations are correct. As a consequence, our framework must provide a way for automatic testing of the implemented objects and functions. To this end, we use the CppUnit[7] testing framework. For the most important data structures and functions, as well as for the statistical computations, we provide test cases assuring the reliability of GeneTrail's computations. Furthermore, we have implemented a program which compares the XML

[7] http://sourceforge.net/projects/cppunit/

output of two GeneTrail executions and states the differences if there are any. This is a useful feature, if we compare the GeneTrail developer version with the version running as web-application before updating the web server to the new version. We have written over 60 different test calls to assure the high quality and reliability of our program.

2.1.4.8 Updates

A crucial issue for gene set analysis tools is to keep the data they use up-to-date. For GeneTrail, we implemented an `Update` class, which handles the updates for the different analysis classes (biological categories) and identifier types. The main program for performing the updates provides several command line options so that, e.g., only single analyses or some specific organisms can be updated. First, the new data is downloaded from their original sources. The file ”update_urls.txt“ in the ’resources’ folder contains all necessary organism specific and general download URLs for this task. Second, the downloaded data is processed and GeneTrail compatible flatfiles for the biological categories are generated. To this end, the update program creates instances of each analysis class to update and calls the method for generating flatfiles from the original data. Mapping files for the different identifier types can also be updated if desired. The file ”files_to_parse.txt“ in the ’resources’ folder contains the information which columns to parse during the update process for tab-separated flatfiles as, e.g., the files downloaded from NCBI (gene2accession, gene_info) that provide most of the identifier mapping informations for GeneTrail. An additional feature of the update process is the option to compare the newly generated flatfiles to the files in another ’data’ folder. This way, we can asses how many changes between two GeneTrail updates occurred and can also detect error or failures during the update process.

2.1.4.9 Adding new analyses / IDs / organisms

Since GeneTrail is modularly constructed, an extension for additional analyses is straightforward. A new specific analysis class must only be created if, e.g., special update and parsing methods are necessary for generating the flatfiles of this analysis. For testing new biological categories with GeneTrail, it is only necessary that flatfiles in the above described format(s) are available. Adding new mappings for different identifier types to NCBI Gene ID is just as simple. The new mapping file must be available in the corresponding format, be named according to our convention, and be present in the designated GeneTrail ’data’ folder. In the GeneTrail ’resources’ folder, we provide a file named ”organisms.txt“, which contains the information which analyses and ID mappings are available for the organisms in GeneTrail. This file is parsed during the update process to generate organism specific flatfiles and in the web-application to dynamically create the websites. If additional analyses and identifier types should be made available for an organism, this file must be edited.

2.1.5 Web-based application

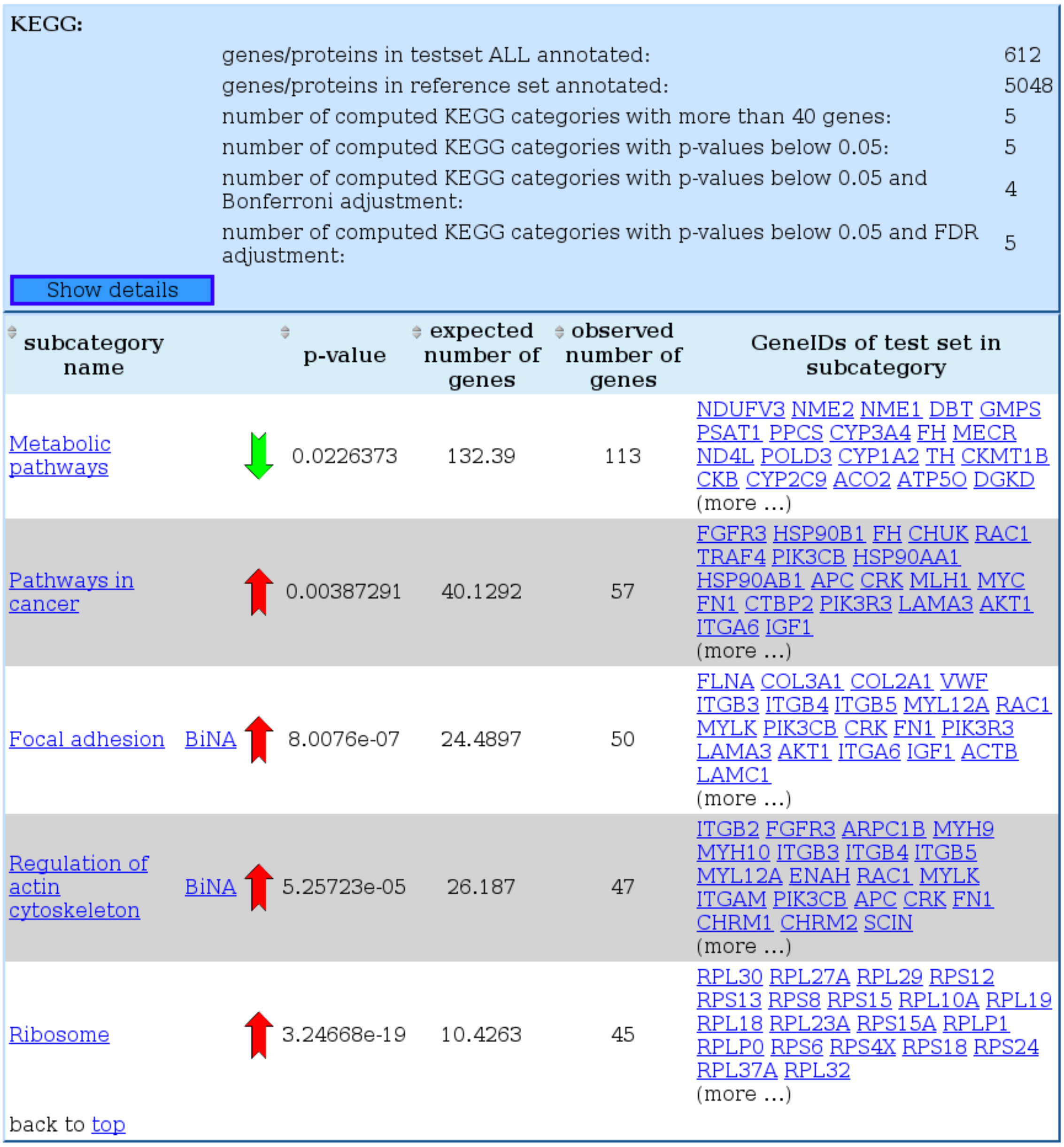
To make GeneTrail's analysis capabilities available to non-developers, we provide an easy-to-use web-application. The web-application is written in PHP and determines stepwise the parameters for the analysis to perform (see also Figure 2.1). The available identifier types and analyses for one organism are dynamically layouted in PHP. User queries are queued, so that the web-server is not overloaded with computation jobs. The results are created in HTML and we use JavaScript to enhance the capabilities of the otherwise static HTML output. The advantage of using HTML with JavaScript instead of PHP for the results page is that the users can download the results and view them offline while preserving the functionality. JavaScript provides the possibility to fade in/out large tables or additional information, or to sort the results according to p-value, subcategory name, expected and observed number of genes in a subcategory. An excerpt of a typical results page for a KEGG pathway analysis is presented in Figure 2.10.

Furthermore, we provide for each test set a link to a table that summarizes the genes that occur in the different significant subcategories. The genes are sorted by the number of significant subcategories they belong to (Figure 2.11). This view provides a quick overview of the genes that play a role in many biological categories. In addition, we provide this information as binary matrix that is suitable to be used in the generation of heatmaps.

Besides the HTML results page, we provide a link to a plain text version of the results that can be more easily used for the import into office applications, and a PDF version. To facilitate the download of the results, the HTML page has a link to a zip file that contains the results in all formats as well as the generated images etc.

2.1.5.1 Network visualization with BiNA

To further improve the visualization of the results, we provide the possibility to view the KEGG pathways in the Biological Network Analyzer (BiNA). BiNA is a visual analytics tool for biochemical networks written in Java that consists of two parts, the platform and a



KEGG:

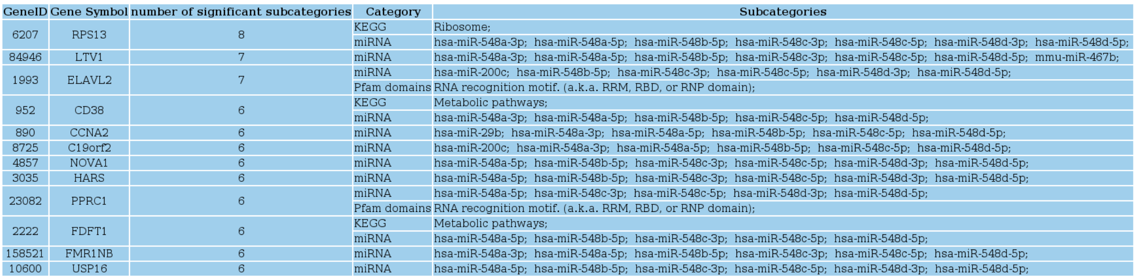
genes/proteins in testset ALL annotated:	612
genes/proteins in reference set annotated:	5048
number of computed KEGG categories with more than 40 genes:	5
number of computed KEGG categories with p-values below 0.05:	5
number of computed KEGG categories with p-values below 0.05 and Bonferroni adjustment:	4
number of computed KEGG categories with p-values below 0.05 and FDR adjustment:	5

Show details

subcategory name			p-value	expected number of genes	observed number of genes	GeneIDs of test set in subcategory
Metabolic pathways		↓	0.0226373	132.39	113	NDUFV3 NME2 NME1 DBT GMPS PSAT1 PPCS CYP3A4 FH MECR ND4L POLD3 CYP1A2 TH CKMT1B CKB CYP2C9 ACO2 ATP5O DGKD (more ...)
Pathways in cancer		↑	0.00387291	40.1292	57	FGFR3 HSP90B1 FH CHUK RAC1 TRAF4 PIK3CB HSP90AA1 HSP90AB1 APC CRK MLH1 MYC FN1 CTBP2 PIK3R3 LAMA3 AKT1 ITGA6 IGF1 (more ...)
Focal adhesion	BiNA	↑	8.0076e-07	24.4897	50	FLNA COL3A1 COL2A1 VWF ITGB3 ITGB4 ITGB5 MYL12A RAC1 MYLK PIK3CB CRK FN1 PIK3R3 LAMA3 AKT1 ITGA6 IGF1 ACTB LAMC1 (more ...)
Regulation of actin cytoskeleton	BiNA	↑	5.25723e-05	26.187	47	ITGB2 FGFR3 ARPC1B MYH9 MYH10 ITGB3 ITGB4 ITGB5 MYL12A ENAH RAC1 MYLK ITGAM PIK3CB APC CRK FN1 CHRM1 CHRM2 SCIN (more ...)
Ribosome		↑	3.24668e-19	10.4263	45	RPL30 RPL27A RPL29 RPS12 RPS13 RPS8 RPS15 RPL10A RPL19 RPL18 RPL23A RPS15A RPLP1 RPLP0 RPS6 RPS4X RPS18 RPS24 RPL37A RPL32 (more ...)

back to top

Figure 2.10: Excerpt of the HTML view for a KEGG analysis. The detailed results can be faded in by clicking "Show details". The columns of the significant subcategories' table can be sorted by subcategory name, p-value, expected and observed number of genes. If more than 20 genes belong to a subcategory, the view is limited to the first 20 genes in this subcategory. The remaining genes can be faded in and out by clicking "more" or "less" in the corresponding field. For the KEGG pathways available in the SQLite database version for BiNA, we provide additional links that directly open the pathway in the visualizer.



GeneID	Gene Symbol	number of significant subcategories	Category	Subcategories
6207	RPS13	8	KEGG	Ribosome;
			miRNA	hsa-miR-548a-3p; hsa-miR-548a-5p; hsa-miR-548b-5p; hsa-miR-548c-3p; hsa-miR-548c-5p; hsa-miR-548d-3p; hsa-miR-548d-5p;
84946	LTV1	7	miRNA	hsa-miR-548a-3p; hsa-miR-548a-5p; hsa-miR-548b-5p; hsa-miR-548c-3p; hsa-miR-548c-5p; hsa-miR-548d-5p; mmu-miR-467b;
1993	ELAVL2	7	miRNA	hsa-miR-200c; hsa-miR-548b-5p; hsa-miR-548c-3p; hsa-miR-548c-5p; hsa-miR-548d-3p; hsa-miR-548d-5p;
			Pfam domains	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain);
952	CD38	6	KEGG	Metabolic pathways;
			miRNA	hsa-miR-548a-3p; hsa-miR-548a-5p; hsa-miR-548b-5p; hsa-miR-548c-5p; hsa-miR-548d-5p;
890	CCNA2	6	miRNA	hsa-miR-29b; hsa-miR-548a-3p; hsa-miR-548a-5p; hsa-miR-548b-5p; hsa-miR-548c-5p; hsa-miR-548d-5p;
8725	C19orf2	6	miRNA	hsa-miR-200c; hsa-miR-548a-3p; hsa-miR-548a-5p; hsa-miR-548b-5p; hsa-miR-548c-5p; hsa-miR-548d-5p;
4857	NOVA1	6	miRNA	hsa-miR-548a-5p; hsa-miR-548b-5p; hsa-miR-548c-3p; hsa-miR-548c-5p; hsa-miR-548d-3p; hsa-miR-548d-5p;
3035	HARS	6	miRNA	hsa-miR-548a-5p; hsa-miR-548b-5p; hsa-miR-548c-3p; hsa-miR-548c-5p; hsa-miR-548d-3p; hsa-miR-548d-5p;
23082	PPRC1	6	miRNA	hsa-miR-548a-5p; hsa-miR-548c-3p; hsa-miR-548c-5p; hsa-miR-548d-3p; hsa-miR-548d-5p;
			Pfam domains	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain);
2222	FDFT1	6	KEGG	Metabolic pathways;
			miRNA	hsa-miR-548a-5p; hsa-miR-548b-5p; hsa-miR-548c-3p; hsa-miR-548c-5p; hsa-miR-548d-5p;
158521	FMR1NB	6	miRNA	hsa-miR-548a-3p; hsa-miR-548a-5p; hsa-miR-548b-5p; hsa-miR-548c-3p; hsa-miR-548c-5p; hsa-miR-548d-5p;
10600	USP16	6	miRNA	hsa-miR-548a-5p; hsa-miR-548b-5p; hsa-miR-548c-3p; hsa-miR-548c-5p; hsa-miR-548d-3p; hsa-miR-548d-5p;

Figure 2.11: The ”genes in significant subcategories“ view.

plugin system. While the platform contains the graphical user interface and many common utilities, it does not have any possibilities for displaying or analyzing networks. For this task, BiNA provides a powerful plugin structure, which can be used to extend BiNA for a variety of applications.

BiNA builds upon the integrative system BN++ [44] and the underlying comprehensive data warehouse BNDB [25]. This warehouse system ensures a full semantic integration of many databases, including KEGG [47] and TRANSPATH [48]. Since GeneTrail relies on the same data warehouse system, the usage of BiNA ensures that the user gets visual representations of exactly the data that are analyzed by our gene set analysis tool. For GeneTrail, we use a Java Webstart version of BiNA allowing the seamless integration into websites. On the HTML results page, GeneTrail adds for each significant KEGG pathway that is available in the database a link to a jnlp file (Figure 2.10). By following this link, the user directly generates a visualization of the respective network. To integrate the pathway data, BiNA provides an SQLite interface to the BN++ database BNDB. If a pathway visualization is started for the first time, BiNA and all available topological network information are downloaded (about 40 MB) and stored on the local hard drive. Whenever BiNA is used again, a version control is carried out ensuring that the newest version of BiNA and the pathway topology information are available on the local disk. Thereby, an efficient visualization is guaranteed, even if the respective networks are large.

A key feature of BiNA is the comprehensive set of available graph layout algorithms. It includes most standard graph layouts (e.g., organic, circular, and hierarchical), but, in addition, also provides biologically inspired graph layouts, implementing the drawing conventions common in textbooks and allowing for a dynamical visualization of the networks using the static KEGG layout information. The visualizer BiNA and the utilized plugins have been implemented by Andreas Gerasch from the Eberhard Karls University in Tübingen.

2.1.6 GeneTrail usage statistics

GeneTrail has been successfully applied in different projects of our own group [24–26, 66], in collaborations with groups of the Saarland University [67–70], and in external groups working in various research areas [71–74]. Our web-based application has been published in 2007 for the first time and gained many users throughout the world ever since. From 15.01.07 – 31.03.10 we had accesses from about 1500 different IP addresses and performed more than 12000 analyses. The numbers of accesses are still increasing as illustrated in Figure 2.12. In the meantime, GeneTrail has been cited in about 34 publications of external groups. Most analyses are performed for human, followed by mouse, thale cress, yeast, rat, and *S. aureus*. The most popular categories are KEGG and GO.

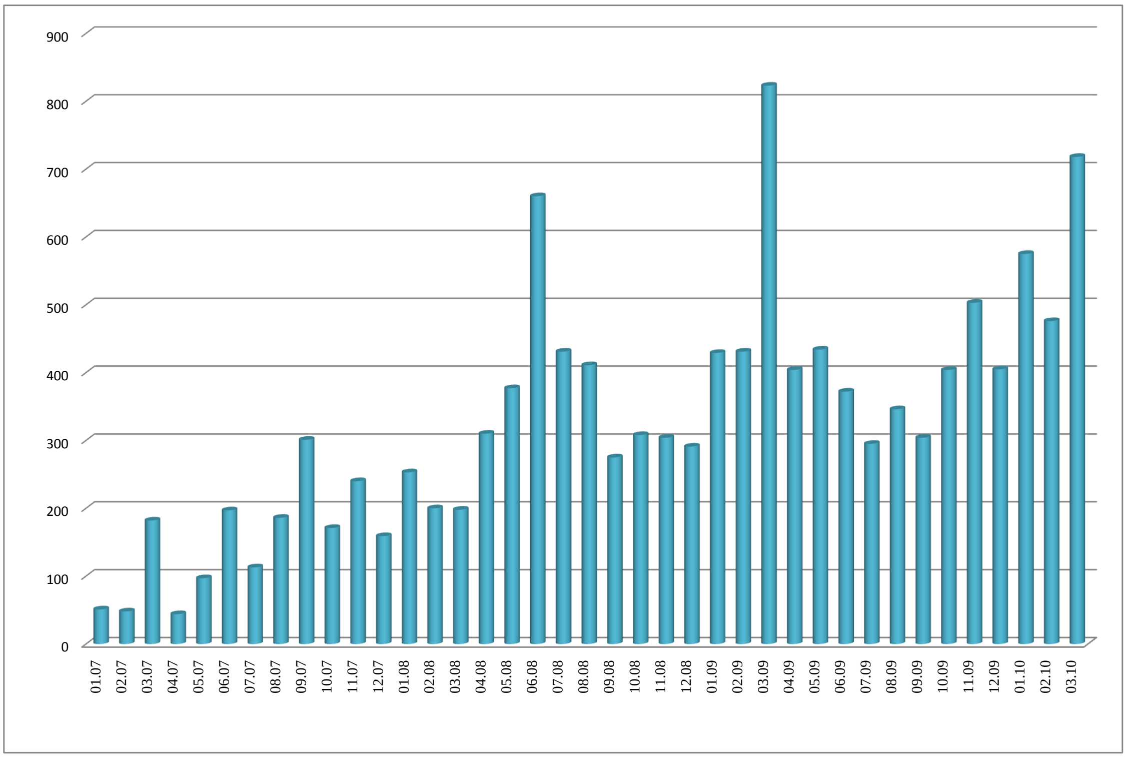


Figure 2.12: Usage statistics for the online version of GeneTrail from January 2007 until March 2010. The y-axis shows the number of program executions, the x-axis the month and year.

2.2 GeneTrailExpress

GeneTrail was developed to serve as an easy to use application for researchers working in different fields. Therefore, the input for GeneTrail consists solely of lists of ”interesting“ genes, which are independent of the type of performed experiments. However, this means that the pre-processing to create these gene lists is left to the researchers. To overcome this issue, we added a pre-processing pipeline, called GeneTrailExpress [24], for preparing raw expression data from microarray experiments.

Several approaches have been developed that focus on the pre-processing of microarray data and provide basic statistical analysis: PMmA [75] was one of the first tools for the detection of differentially expressed genes. The program NMPP [76] is tailored for the pre-processing of self-designed NimbleGen microarray data. Other tools, as AMDA [77] offer clustering methods and functional annotation of the differentially regulated genes. More examples of tools focusing on pre-processing and basic statistical evaluation are ArrayPipe [78], one of the first web-based application, or GEPAS [79], which provides clustering methods and can correlate its results to diverse clinical outcomes. Most recently, Morris et al. [80] described a comprehensive collection of perl modules for microarray management and analysis. However, none of these tools provide a dynamic graphical representation of the detected pathways. This has to be done manually using one of the existing network visualization tools. One of the most popular visualizers with a large user and developer base is Cytoscape [65], which also offers a plug-in architecture allowing to extend the functionality, e.g., for integrating data analysis methods. Other visualization tools for biological interaction data are VisANT [81], which has been designed specifically for the integrative visual data-mining of biological pathways, and OSPREY [82], which has been developed to explore large networks.

The usage of GeneTrailExpress comprises several steps. First, the expression data is uploaded or selected. Then, the user can select different normalization and gene scoring methods. The resulting list of interesting genes is directly subjected to GeneTrail's extensive gene set analysis methods and relevant findings are correspondingly visualized. The GeneTrailExpress pre-processing has been implemented by Maher Al-Awadhi during his bachelor thesis.

2.2.1 Input

GeneTrailExpress offers three options for uploading expression data. The user can either upload (1) their own expression matrix that must contain two groups of microarrays, e.g. control versus treatment, or (2) a list of genes with scores. A third possibility is the usage of a database connection to the NCBI Gene Expression Omnibus (GEO). In this case, the user can select two GEO GDS expression profiles of the same microarray platform. When uploading an expression matrix or using expressing profiles from GEO, GeneTrailExpress continues with normalization and scoring of the data.

2.2.2 Normalization

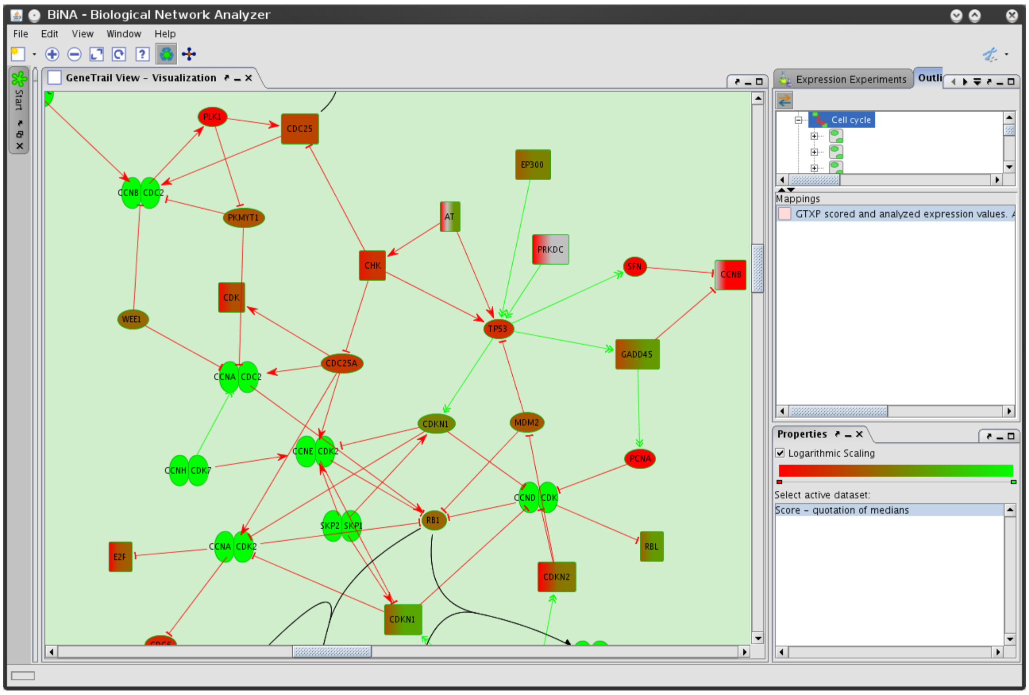
Microarray experiments can be influenced by many factors, as systematic and random biases. To overcome this issue, normalization techniques are applied to make different microarray experiments comparable to each other. GeneTrailExpress offers several standard statistical normalization techniques, including mean value normalization, median value normalization, or a normalization of mean and variance. The distributions of expression values before and after normalization are presented via bar charts to visualize the effects of the normalization on the expression values.

2.2.3 Scoring

Following the normalization, the next step is to identify differentially expressed transcripts. The following scoring functions for the computation of the differential expression are available in GeneTrailExpress: mean fold-change, median fold-change, unpaired t-test, paired t-test, Wilcoxon-Mann-Whitney test, ANOVA, and Wilcoxon Rank-Sum test. For facilitating the usage, scoring methods that are not suitable for the given input are disabled. The resulting scores of the genes can be manually inspected along with their distribution shown as a histogram. The final gene list is directly subjected to GeneTrail for a ORA or a GSEA as explained previously.

2.2.4 Network visualization with expression values

As we described in Section 2.1.5.1, we use a Java WebStart version of BiNA for a dynamical visualization of significant KEGG pathways in GeneTrail. Besides visualizing pathways, BiNA allows to map arbitrary scalar data, like expression data, onto the biological networks. When using GeneTrailExpress, the computed scores for the genes in an expression experiment can be used to color the nodes of the visualized significant pathway, which facilitates the interpretation of the statistical evaluation. Figure 2.13 shows BiNA's graphical user interface visualizing a real biological example. A GSEA of lung cancer expression data reveals overexpression of lung cancer genes in the Cell Cycle, indicated by the red-colored genes.



GTXP scored and analyzed expression values.

☑ Logarithmic Scaling

Select active dataset:
Score – quotation of medians

Figure 2.13: BiNA visualization of the cell cycle with mapped expression values. For the gene expression omnibus data set GDS1312, containing human lung cancer samples and normal controls, we used GeneTrailExpress to compute the quotient of medians for these expression experiments. The subsequent gene set enrichment analysis found the KEGG pathway cell cycle significantly enriched, which provides evidence for a clear up-regulation of the cell cycle in lung cancer. All genes are colored with respect to their quotient of median scores. The pale green complexes correspond to protein complexes.

2.3 GraBCas

In this section, we introduce GraBCas, a tool written in Java to predict granzyme B and caspase cleavage sites. Caspases are enzymes orchestrating the cellular pathways that lead to apoptosis and inflammatory signals. Besides these functions they are supposed to be involved in other cellular processes, such as development, cell cycle, cell proliferation, cell migration and receptor internalization [83, 84]. Caspases are cysteine proteases with specificity for an aspartic acid residue at position P_1 of the substrate. This primary specificity is shared by the serine protease granzyme B, which induces cytotoxic T lymphocyte-mediated target cell DNA fragmentation and apoptosis [85, 86]. Granzyme B-mediated cleavage also plays a role in the induction of autoimmunity [87].

A more comprehensive knowledge of caspase and granzyme B substrates is essential to understand the biological roles of these enzymes in more detail. The relatively high variability in their recognition sequence often complicates the identification of cleavage sites. At the time of publication in 2005, GraBCas was the first tool that allowed identification of caspase and/or granzyme cleavage sites differing from the consensus sequence. Other available tools at that time were the PeptidCutter program provided by the ExPasy Server[8] that considers only the preferred peptide substrate sites and 'PEPS', a tool of Lohmüller et al. [88], that is restricted to caspase 3 and cathepsin B and -L substrates. In the meantime, more recent applications make use of SVMs to predict the cleavage sites of caspases [89, 90] or combine sequence and structure information to predict substrates of endoproteases [91]. In the work of Wee and coworkers [92], GraBCas has been integrated to reduce efficiently the number of false positives when predicting caspase cleavage sites.

In the following, we briefly summarize the score-based prediction of potential cleavage sites integrated in GraBCas for the caspases 1-9 and granzyme B as presented in our NAR Web Server Issue publication [43].

2.3.1 Design of cleavage site scoring matrices

We developed position specific scoring matrices (PSSMs) for the endopeptidases granzyme B and caspase 1-9 based on experimentally determined substrate specificities for the cleavage site positions P_4, P_3, and P_2 of these proteases [93]. Thornbery et al. [93] determined the substrate specificities using positional scanning synthetic combinatorial libraries. Cleavage was fluorimetrically determined with maximum value annotated with 100 for the best cleavage site and the values for the remaining cleavage sites given as percentage of the observed maximum rate. These experimental values provided the basis for creating our PSSMs.

The values for each amino acid at position P_i are shown in Table D.1 in Appendix D. For

[8] `http://www.expasy.org/tools/peptidecutter`

a better readability we decided to set the maximum values to 1000 instead of 100 and adjusted the other values accordingly. For each endopeptidase the scores of the amino acids were entered in a 3 x 20 matrix. The rows of such a matrix correspond to positions P_4, P_3 or P_2 of a possible cleavage site. Each column represents one amino acid and contains the relative frequencies of the amino acid measured in the study of Thornbery et al. [93]. We are working with PSSM that can be interpreted as probability matrices. Since probabilities of value 0 should be avoided in such probability-based position scores, all entries of experimental relative frequencies with value 0 were set to 1. The amino acids cysteine and methionine were not part of the study of Thornbery et al. [93]. The entries for these amino acids were also set to 1 in Table D.1.

2.3.2 Computing the scores of endopeptidase cleavage sites

For computing the score, the GraBCas program screens for tetrapeptides with Asp (D) at their last position (P_1) in a given amino acid sequence. Given the tetrapeptide $A_4A_3A_2D$ ($\approx P_4P_3P_2P_1$) of a potential cleavage site, its score for a given endopeptidase is computed by the formula in Equation 2.5. The corresponding matrix entries of A_2 at position P_2, A_3 at position P_3, and A_4 at position P_4 are multiplied. The product is divided by the value of the product of the consensus recognition motif for normalization and multiplied by 100, yielding a total score between 0 and 100.

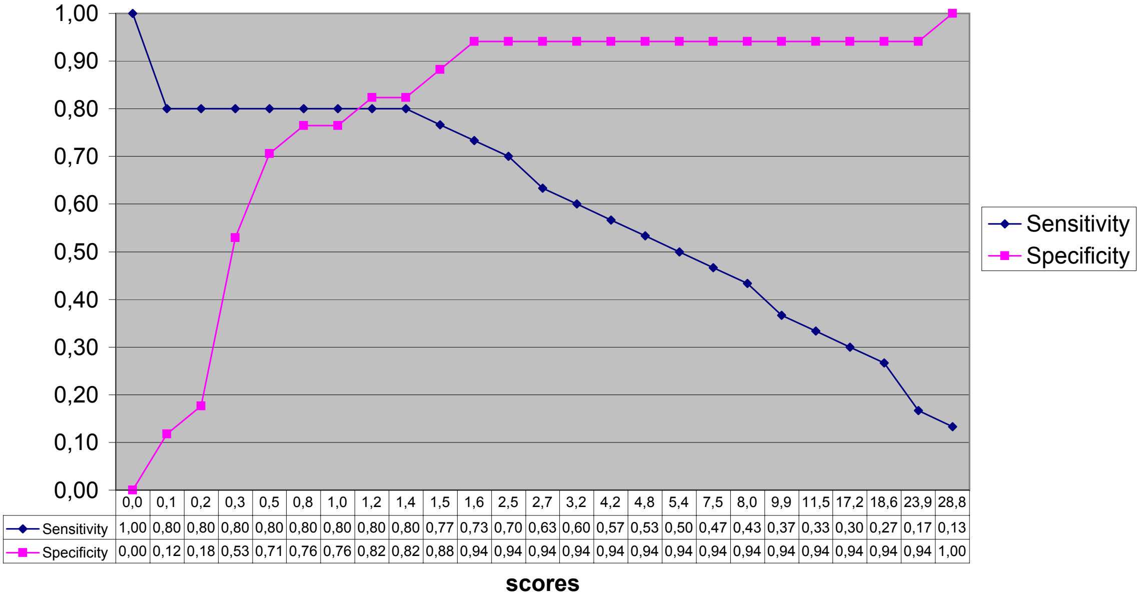
$$Score(A_4A_3A_2D) = 100 \cdot \frac{Score_{P_4}(A_4) \cdot Score_{P_3}(A_3) \cdot Score_{P_2}(A_2)}{1000^3} \tag{2.5}$$

2.3.3 Sensitivity-specificity plot for granzyme B

For determining the specificity and sensitivity of the GraBCas predictions and an optimal cutoff for the PSSM scores, we used the known cleavage sites of granzyme B [86, 87, 93–97] and the known non-substrates of granzyme B [87]. The x-axis of the plot in Figure 2.14 represents the cutoff values (with respect to the PSSM scores), while the y-axis represents the percentage of the specificity or sensitivity of the predictions made by GraBCas, respectively.

The specificity is computed as follows:

$$\text{Specificity} = \frac{\text{Number of true negatives}}{\text{Number of false positives} + \text{Number of true negatives}}$$



	0,0	0,1	0,2	0,3	0,5	0,8	1,0	1,2	1,4	1,5	1,6	2,5	2,7	3,2	4,2	4,8	5,4	7,5	8,0	9,9	11,5	17,2	18,6	23,9	28,8
Sensitivity	1,00	0,80	0,80	0,80	0,80	0,80	0,80	0,80	0,80	0,77	0,73	0,70	0,63	0,60	0,57	0,53	0,50	0,47	0,43	0,37	0,33	0,30	0,27	0,17	0,13
Specificity	0,00	0,12	0,18	0,53	0,71	0,76	0,76	0,82	0,82	0,88	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	0,94	1,00

Figure 2.14: Sensitivity-specificity plot for granzyme B cleavage sites according to GraBCas. x-axis: scores by the GraBCas program; y-axis: percentage of specificity or sensitivity.

The true negatives are the known non-substrates, where the maximal PSSM score of all tetrapeptides ending with a D is smaller than the chosen cutoff value. Analogously, the false positives correspond to the non-substrates that are falsely classified as substrates given the chosen cutoff value. A specificity of 1 means that all known non-substrates were below the cutoff, i.e. all known non-substrates were correctly classified as negatives.

The sensitivity is defined as:

$$\text{Sensitivity} = \frac{\text{Number of true positives}}{\text{Number of true positives} + \text{Number of false negatives}}$$

where true positives are the known cleavage sites with a score larger than the chosen cutoff value. Analogously, the false negatives correspond to the substrates that are falsely classified as non-substrates given the chosen cutoff value. A sensitivity of 1 means that all cleavage sites of our test set have a score higher than the chosen cutoff and that they have been correctly classified as positives.

In total, we collected 29 substrates with 30 cleavage sites for granzyme B and additionally 17 sequences which are non-substrates of this endopeptidase. We computed the scores of all putative cleavage sites in these sequences and extracted the best hit by GraBCas for each of these (non-)substrates. The sensitivity-specificity plot for granzyme B in Figure 2.14 shows that we obtain a sensitivity of 80% and a specificity of 82% when using a cutoff value of 1.2 in the GraBCas program. The cutoff value can be adjusted if a higher specificity or sensitivity is needed for the cleavage site prediction.

Given these findings for granzyme B, we integrated the GraBCas prediction for granzyme B in GeneTrail using a cutoff of 1.2.

2.4 Conclusion

In this chapter, we presented the comprehensive gene set analysis framework GeneTrail. Although the competition in this field is immense, we were able to establish GeneTrail in this area as demonstrated by the usage statistics and the citations of external groups. The continuous development and extensions lead to a powerful C++ framework that is not only useful for gene set analyses, but also builds the basis to answer more complex questions when using the network topology. Over the years, we integrated many ideas and suggestions of users and enhanced the functionality and the user-friendliness of the web-application. In summary, GeneTrail presents one of the most powerful non-commercial gene set analysis tools that are available for the research community.

In the next three chapters, we further demonstrate the usefulness of GeneTrail by performing comprehensive analyses for different fields of cancer research comprising the analysis of characteristics of tumor associated antigens, the putative target pathways of miRNAs, and differential network analyses concerning glioma versus normal and BRCA1 mutation carriers versus non-mutation carriers.

CHAPTER 3

TUMOR ASSOCIATED ANTIGENS

Tumor associated antigens (TAAs) are capable to elicit an immune response in cancer patients. Since these antigens stem for example from proteins which are also expressed under normal conditions in healthy tissues, there must exists reasons why they become immunogenic in cancer tissue. In cancer prognosis/diagnosis some TAAs are already applied as biological marker (e.g. PSA, prostate specific antigen) [30]. Although the detection and usage of TAAs is already widely possible, the mechanisms which lead to a humoral immune response against these antigens are for the most part elusive. In this work, we will try to shed some light into this topic by testing different hypotheses for immunogenicity. First, we give a short overview concerning immune response in general and the discussed mechanisms for eliciting an immune response in autoimmune diseases and cancer. Second, we describe the experimental methods available for detecting antigens and the data sets used in this work. Third, we apply bioinformatics approaches for verifying if the stated hypotheses can be generalized for TAAs.

3.1 Immune response and autoimmunity

The defense mechanisms of the immune system of higher multicellular organisms originate from the fact that their bodies provide an optimal environment for the reproduction of microorganisms such as bacteria, viruses, and parasites. In general, the immune system can be divided into two types of defense mechanisms: the innate and the adaptive immune system [98].

The innate immune system encompasses unchanging mechanisms that are continuously in force, as for example the skin as a physical barrier, which pathogens have to overcome. These non-specific mechanisms contribute to a basic resistance of an organism against foreign pathogens. In contrast to the innate immune system, the adaptive immune system is characterized by a high degree of specificity. The reaction of the adaptive immune system is elicited by the recognition of specific molecules called antigens, or rather by the recognition of some specific surface structures of the antigen, called epitopes. The recognition of pathogen specific antigens is a multi-step process, which starts with the digestion of pathogens by macrophages or immature dendritic cells. Subsequently, these cells become activated or mature to so-called antigen presenting cells (APCs). The digested antigens are fragmented to peptides of about 9-15 amino acids, which are presented to T-helper cells by means of major histocompatibility complexes (MHCs) on the surface of the APCs (Figure 3.1). T-helper cells specific for recognizing the peptide:MHC structure become activated and start to secrete cytokines, which activate in turn cytotoxic T-lymphocytes, antibody-secreting B-cells, macrophages, etc. resulting in the activation of the humoral and/or cellular immune response. The function of the humoral immunity is to recognize and to destroy extracellular pathogens and foreign substances. B-cells activated by their corresponding antigen and the cytokines of the CD4 T-helper cells will start to proliferate and differentiate into antibody secreting plasma cells. The antibodies secreted by the plasma cells bind to their specific epitope on the antigen, thereby disabling the antigen, and mark it for processes leading to its destruction. By contrast, the function of the cellular immunity is to detect intracellular pathogens. The main components of the cellular immunity are CD8 T-helper cells and cytotoxic T-lymphocytes (CTLs). To distinguish normal cells from modified cells, a mechanism is necessary that reports the cells' state. The proteins expressed in a cell are again decomposed to some extent. The protein fragments or peptides are presented on the cell surface by MHC class I molecules. CTLs can recognize cells presenting non-self peptides like virus-infected cells or tumor cells expressing modified proteins and induce cell death by secreting toxins.

The reasons for the loss of the so-called self-tolerance in autoimmune diseases or cancer, which results in the activation of the humoral immune response against self-antigens, are still elusive for the most part and can have many potential causes, some related to the immune system itself, and some related to the antigen targets. For some autoimmune diseases the loss of self-tolerance originates from the similarity of self proteins to pathogenic antigens, which is called molecular mimicry. This theory proposes that the immune reaction initially elicited by a foreign antigen, which is structurally similar to a human protein, can result in a cross-reaction against the human protein [99]. While the loss of self-tolerance often comes along with autoimmunity, the immune response in cancer patients may be initiated by alterations in the tumor itself. Such alterations comprise, e.g.,

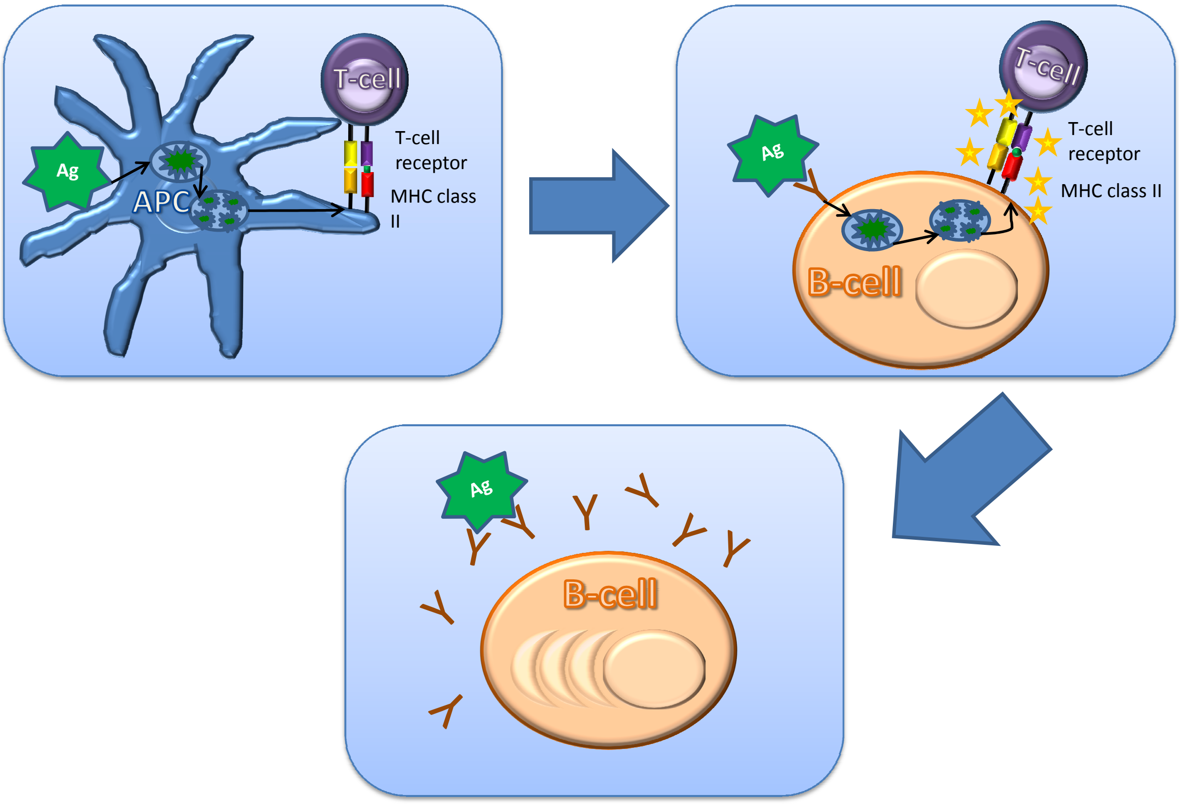


Figure 3.1: The humoral immune response. Antigen processing cells (APCs) ingest and process an antigen. The processed antigen is presented by MHC class II molecules to CD4 T-cells. The activation of antigen-specific CD4 T-cells leads to lymphoproliferation and cytokine secretion. The activation of a B-cell comprises several steps. Antigen-antibody complexes on the surface of the B-cell are internalized by receptor-mediated endocytosis and degraded to peptides. These are presented by MHC class II on the membrane to CD4 T-helper cells. Specific T-helper cells recognize the peptide:MHC structure and additional co-stimulatory signals, which lead to the activation of the T-helper cell. The activated T-helper cell secretes cytokines that help the B-cell to differentiate into an antibody secreting plasma cell.

mutated proteins or differential expression that may result in an increased immunogenicity of self-antigens [100].

3.2 Experimental techniques

In order to identify serum antibodies, several experimental techniques can be applied. The most commonly used techniques include Enzyme-Linked ImmunoSorbent Assays (ELISA) [101], SErological identification of antigens by Recombinant EXpression cloning (SEREX) [102–104], Protein Arrays [105], and Two-Dimensional Polyacrylamide Gel Electrophoresis (2D-PAGE) [106]. In the following, we describe the SEREX method and the protein arrays in more detail, since the data sets analyzed in this thesis have been generated with these methods. The corresponding experiments have been carried by the group of Prof. Eckart Meese.

3.2.1 SEREX

The SEREX (SErological identification of antigens by Recombinant EXpression cloning) method was developed by Sahin et al. [102] and serves to identify antigens eliciting an immune response in cancer patients. For the application of the SEREX method, first, a cDNA expression library is built by extracting mRNA of (tumor) tissue. Subsequently, *E. coli* cells are transfected with the cDNA library and plated on agar plates, where they express the recombinant proteins. The expressed proteins are incubated with the serum of a patient and if this serum contains antibodies against a certain protein of the cDNA library, this can be detected with a color reaction. The methods allows for the identification of the clone expressing the protein by sequencing the cDNA of the positive clone. The corresponding gene of the clone is determined by sequence alignment.

3.2.2 Protein arrays

Protein arrays present a further a high-throughput method for detecting autoantibodies. In general, proteins which can stem from different sources (purified or recombinant proteins, synthetic peptides, or fractioned proteins from tumor tissue or cell lines) are immobilized on the array and then incubated with specific sera. The antibody-antigen reaction can be detected via enzymatic labeling or fluorescent dyes.

For the primary screening, we used high-density protein arrays consisting of 38016 *E. coli* expressed proteins from the hex1 cDNA expression library [105], of which about 4000 represent known genes. These arrays were screened with sera from patients with various human diseases including cancer and inflammatory diseases, as well as blood sera from healthy controls. The screening was performed with minor variations as described in [107]. To lower the experimental costs, a second customized protein array was designed containing only those clones (about 1800) that showed reactivity in at least one of the pools of the primary screening.

Besides using cDNA expression libraries, protein arrays can also be spotted with peptides from Phage Display Libraries. A phage display library is constructed from tumor tissue or a cell line. The candidate antigen peptides are expressed and displayed on the surface of a phage. One advantage of these libraries is that peptides that are specifically recognized by patient serum can be enriched using a process called biopanning. On the other hand, this method has the limitation that the peptide sequences are short and the results may be difficult to interpret if the peptide stems from a non-coding sequence.

3.3 Data sets

The antigen sets used in this thesis stem either directly from experimental methods (SEREX, protein arrays) or from literature search. In the following we compose the name for the different data sets of: (1) their source ('Lit' for 'collected from literature', 'CIDB' for the database the antigens stem from, or 'Exp' if found in experiments performed by the Human Genetics Department of Prof. Meese), (2) the experimental method (e.g. 'Serex', 'Chip', and 'PhageDisplay' for the corresponding experimental method if available), and (3) the type of antigens contained in the data set (AAG for autoimmune antigens, HAG for antigens occurring in healthy persons, TAG for tumor antigens, INAAG for antigens occurring in non-tumor diseases comprising inflammatory, neural, and autoimmune diseases, AG for antigens containing mixtures of AAGs, HAGs, and TAGs). If we build subsets of an antigen data set, we add the extended criterion for building the subset to the name of the original set (e.g. 'dataset>5%Sera' means that we take all antigens from 'dataset' that were found in at least 5% of the screened sera).

The first data set we consider here was extracted from the "Cancer Immunome Database" [1] (CIDB). The antigen set *CIDB-Serex-AG* contains 1471 known genes for which antibodies

[1] http://ludwig-sun5.unil.ch/CancerImmunomeDB/

could be detected primarily in patients with different cancer types. The method applied for the detection of these antigens is SEREX. The CIDB data we use in this thesis was downloaded in February 2009. For excluding such antigens that may be in these sets because of experimental errors, e.g. the subjective optical evaluation of positive spots on the SEREX filters, we collected two data sets containing antigens that were found with at least two sera (*CIDB-Serex-AG>1Serum*) or were represented by at least two clones (*CIDB-Serex-AG>1Clone*).

The second and third data set were collected by the group of Prof. Meese using the SEREX method. For the set *Exp-Serex-HAG*, the screening was performed with sera of healthy donors to detect natural occurring autoantibodies yielding 86 known genes. The data set *Exp-Serex-TAG* contains 74 antigens detected when screening sera of glioma, meningioma, and lung cancer patiens.

Besides the antigen sets that have been isolated with the SEREX method, we also consider here antigens found when screening protein macroarrays. The set *Exp-Chip-AG* contains 298 antigens that were positive for at least one pool of sera in the primary screening (see Section 3.2.2), not distinguishing between cancer, healthy, or other diseases. In more detail, the primary screening was performed with pools of sera including prostate cancer, lung cancer, meningioma, glioma, morbus crohn, colitis, stroke, and healthy controls. From this base antigen set, we derived 5 more specific sets. The set *Exp-Chip-AG>1Pool* consists of antigens found in at least 2 pools of the primary screening with the original chip. Using the positive clones found in the primary screening of the chip, a second customized chip was designed. This customized chip was used in further disease specific and healthy control screenings with more than 500 sera. The results of these screenings are of course also subsets of the original set, therefore we keep the naming. The set *Exp-Chip-AG>5%* contains antigens that were positive in at least 5% of the tested sera in total. Analogously, the sets *Exp-Chip-HAG>5%* and *Exp-Chip-TAG>5%* contain antigens that were found in at least 5% of the healthy control sera or tumor sera, respectively. In addition, several autoimmunity associated, neural, and inflammatory diseases were screened (e.g. colitis, chronic obstructive pulmonary disease, morbus crohn, multiple sklerosis). These antigens that were found in at least 5% of patients are summarized in the set *Exp-Chip-INAAG>5%*.

Furthermore, we consider in our analysis two data sets that were collected by literature search. The *Lit-PhageDisplay-TAG* set contains 84 tumor antigens that were isolated with the Phage Display library method. The antigen set *Lit-AAG* consists only of genes as-

Table 3.1: Data sets for our analyses and their publication bias illustrated with the help of their GO annotations (GO version: January 2010). The average number of GO annotations and the percentage of high-quality annotations thereof are a indication for the publication bias underlying the genes in the set. The more interesting a gene is, the more annotations in total and the more high-quality GO annotations it is likely to have assigned. Most of our data sets show about the same average GO annotation, with exception of the Lit-AAG set showing the highest number of average GO annotations, and the ProteinCodingGenes with the lowest number of average GO annotations.

Data Set	Subset	Number of known Genes	GO annotations[a]
CIDB-Serex-AG		1471	10.16 (23.30%)
	CIDB-Serex-AG>1Clone	446	11.12 (25.16%)
	CIDB-Serex-AG>1Serum	306	11.04 (26.14%)
Exp-Serex-HAG		85	9.48 (17.95%)
Exp-Serex-TAG		74	11.86 (25.66%)
Exp-Chip-AG		298	10.10 (23.93%)
	Exp-Chip-AG>1Pool	130	9.77 (25.76%)
	Exp-Chip-AG>5%	217	9.58 (23.24%)
	Exp-Chip-HAG>5%	211	9.61 (23.12%)
	Exp-Chip-INAAG>5%	222	9.63 (23.57%)
	Exp-Chip-TAG>5%	241	9.60 (22.87%)
Lit-PhageDisplay-TAG		84	13.45 (26.86%)
Lit-AAG		348	15.88 (28.51%)
ALL		2079	10.99 (24.14%)
ProteinCodingGenes		23583	8.50 (18.71%)

[a] The average number of GO terms annotated per gene. The number in parentheses corresponds to the percentage of high-quality annotations (with evidence tags ’inferred from direct assay’ or ’traceable author statement’) amongst all annotations for genes in the group.

sociated with autoimmune diseases, is available online[2], and contains 348 genes. This set was initially collected to analyze the occurrences of SNPs (single nucleotide polymorphisms) in autoantigens [108]. SNPs are DNA sequence variations of single nucleotides in the genome. Such a variation must occur in at least 1 % of the population to be considered as SNP. Stadler et al. found that the occurrence of SNPs is significantly higher in these autoantigens than in remaining human genes [108].

The ALL set contains the union of all antigens of the data sets Lit-AAG, Exp-Serex-HAG, Exp-Chip-AG, Exp-Serex-TAG, CIDB-Serex-AG, and Lit-PhageDisplay-TAG. This set can help to find prevalent patterns of antigens if there exist common causes for eliciting the immune response in cancer patients, autoimmunity, and healthy controls.

As reference set, we used all human protein coding genes excluding the above mentioned antigens (human protein coding genes minus genes in the ALL set). The different data sets are summarized in Table 3.1.

We performed the analyses with GeneTrail for all antigen sets, if not mentioned otherwise, using the following parameters: significance level: 0.05; minimum number of genes in a subcategory: 2; p-value computation: FDR correction; reference set: ProteinCodingGenes. When performing an ORA (Section 2.1.3.1), we filtered the results afterwards for significantly enriched subcategories that contained at least 5% of the genes of the test set that had an annotation for the considered category. This way, we focused on subcategories that show a certain prevalence in our antigen sets.

3.4 Influence of genetic alterations and changed expression levels

Alterations on the molecular level in the cell can lead to the production of aberrant proteins. The production of these modified proteins can have many possible causes, e.g., a mutation on DNA basis can directly either influence the expression level or the amino acid sequence of the resulting protein. A changed expression in tissues where a protein is normally not expressed could lead to its increased presentation on MHC complexes. In general, all processes influencing the final protein expression may cause the presentation of aberrant self-peptides to the immune system, and therefore, may be responsible for eliciting an immune response in cancer patients. In the following, we test different hypotheses to analyze

[2] http://www.wiley-vch.de/contents/jc_2040/2005/25481_s.pdf

whether genetic alterations and/or changed expression levels are possible initiators for a humoral immune response.

3.4.1 Mutations

One of the characteristics of cancer is the accumulation of genetic alterations in the DNA. For the integrative analysis of heterogeneous data from several cancer-related sources we have developed the cancer-associated protein database (CAP) [109]. To study the connection between mutations and immunogenicity, CAP has been employed on data derived from SEREX experiments and data extracted from Cancer GeneticsWeb (CGW)[3], which provides general information and literature references about cancer-related genes. Out of 723 genes from SEREX experiments and the 606 genes contained in CGW, we found only 17 genes occurring in both data sets. Additionally, we analyzed if the genes in the overlap of both data sets have been found in the same cancer types. A total of seven genes were identified, where only two (TP53 and GSTT1) are known to carry specific mutations or polymorphisms, whereas the remaining five are over-expressed in the respective tumors. TP53 has been found to cause immune responses in primary colon carcinoma and in breast carcinoma, both known to carry TP53 mutations [110, 111]. Mutations in TP53 have also been found in a large number of other tumor types where the patients have no antibodies against TP53. The same holds for GSTT1, where antibody responses occur in patients with breast cancer that is associated with specific GSTT1 polymorphisms [112]. However, these types of polymorphisms also occur in other tumors including head and neck cancer without an antibody response [113]. From the data used in this analysis, it does not seem likely that the genetic alterations are primarily responsible for causing an immune response in cancer. However, we performed this analysis in the year 2004 with a limited amount of available data, so these findings represent only preliminary results.

As a more current data source, we used the “Roche Cancer Genome Database” (RCGDB)[4] [114] that combines different sources of human mutation databases including amongst others the Catalogue of Somatic Mutations in Cancer (COSMIC), the Cancer Genome Atlas, and Online Mendelian Inheritance in Man (OMIM). For our analysis we extracted for each gene in this database the different types of somatic mutations and the number of their occurrences in cancer. We performed a Wilcoxon-Mann-Whitney (WMW) test (described in Section 2.1.3.3) to check the hypothesis whether our antigen sets have a higher number

[3] `http://www.cancer-genetics.org`

[4] `http://rcgdb.bioinf.uni-sb.de/MutomeWeb/`

of somatic mutations compared to the reference. The results are illustrated in Figure 3.2.

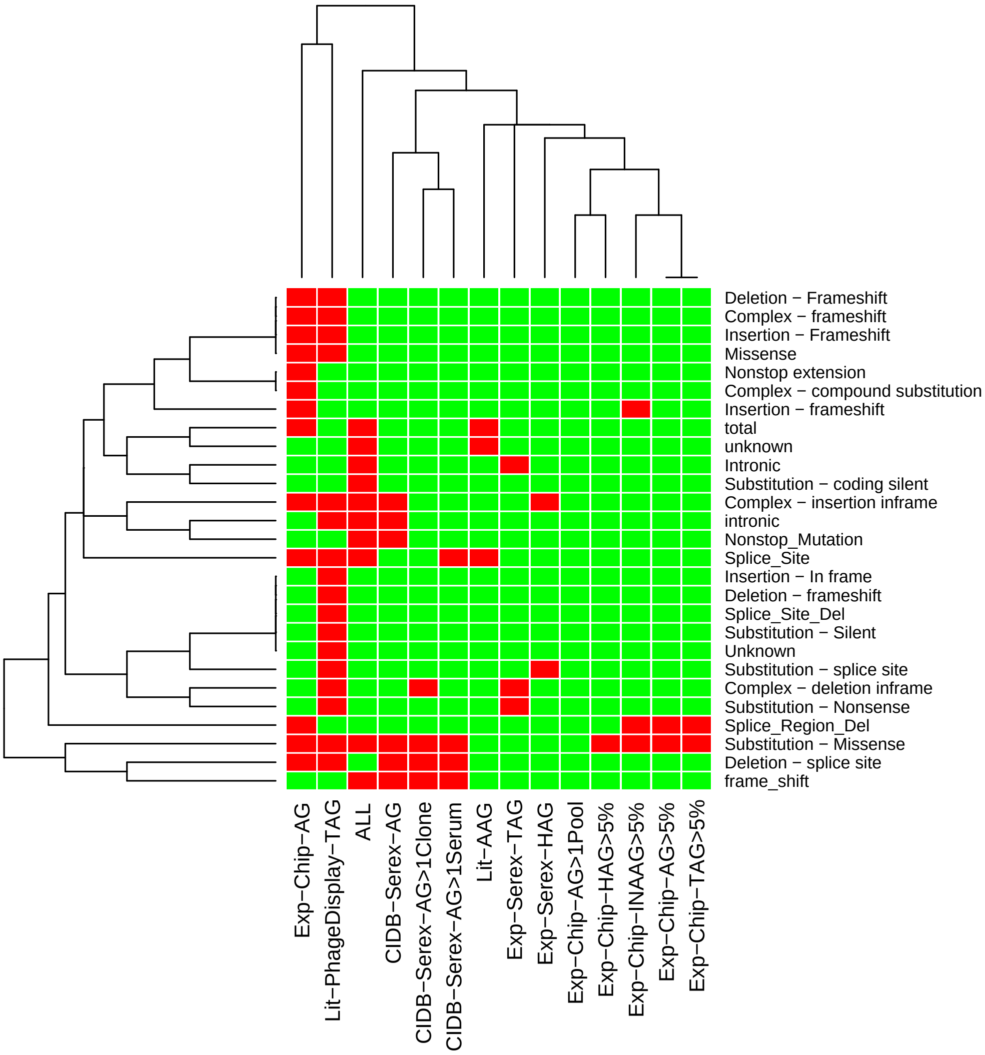


Figure 3.2: Heatmap of the results of the WMW test for comparing the distribution of somatic mutations in the antigen sets and the reference. Red = significantly enriched compared to the reference. Green = not significant.

The heatmap shows that there is an enrichment for almost all data sets for the subcategory “Substitution - Missense” except for the data sets Lit-AAG, Exp-Serex-HAG, Exp-Serex-TAG, and Exp-Chip-AG>1Pool. The latter set shows no enrichment for any somatic mutation subcategory. The data sets showing the most significantly enriched somatic mutations are the Exp-Chip-AG, the Lit-PhageDisplay-TAG, and the ALL set, however, there are only a few overlaps. Besides the “Substitution - Missense” subcategory, they have the categories “Splice_Site” and “Complex - insertion inframe” in common. Interestingly, the CIDB-Serex-AG set and its more specific sets (CIDB-Serex-AG>1Clone, CIDB-Serex-AG>1Serum) cluster together, as well as the specific Exp-Chip-AG sets (Exp-Chip-INAAG>5%, Exp-Chip-HAG>5%, Exp-Chip-TAG>5%, Exp-Chip-AG>5%). These results indicate that somatic mutations play a more important role than previously suggested. Especially the finding of the significantly enriched subcategory "Substitution - Missense" in almost all data sets that contain primarily tumor derived antigens in contrast to the data sets Lit-AAG, Exp-Serex-HAG (autoimmune diseases and healthy controls), and Exp-Chip-AG>1Pool (contains additionally antigens of inflammatory diseases) seems to be of major importance. However, the Exp-Serex-TAG set containing antigens from lung, glioma, and meningioma is an exception from this observation.

3.4.2 Expression levels

To test if changes in expression levels have an influence on the immune response in cancer patients, we correlated all cancer-related genes in the CAP database found by SEREX experiments with expression data from the NCI60 microarray project [115]. In this project, cDNA microarrays are used to explore the variation of gene expression in 8000 genes from 60 cancer cell lines. These 60 cell lines are also used by the National Cancer Institute for screening potential cancer drugs. The expression data provided by NCI include fluorescence ratios, normalized against a pool of 12 cancer cell lines. For our analysis we only considered genes that showed at least a 2-fold increase in expression levels and were measured in at least 4 of the 60 cell lines resulting in 319 genes of CAP having expression levels. In total, we found 277 (87%) of the genes to be over-expressed in at least one cell line. Out of the 277 genes, 69 were found to have an over-expression in at least 10% of all evaluated cell lines. In a more cancer-specific analysis, we extracted expression levels for genes that were found in the same cancer type in both SEREX experiments and the NCI60 data. A total of 13 genes meeting this restriction showed over-expression in at least 3 tumor-specific cell lines. These findings indicate that over-expression may contribute to the antibody responses against tumor antigens. The majority of the 319 genes are actually found to be over-expressed in the NCI60 data set. However, this result might be somewhat biased from the selection of genes tested for expression levels, since the NCI60 data set was designed to explore the variation in gene expression among different cancer types.

3.4.3 SNPs

Stadler et al. found an enrichment of SNPs in autoantigens and discussed these as the cause for the immunogenicity [108]. We verified whether this hypothesis also holds for TAAs or antigens in general. To this end, we extracted the different SNPs for every gene as deposited in dbSNP [116] from NCBI and performed a WMW test. The results of this statistical test are depicted in Figure 3.3.

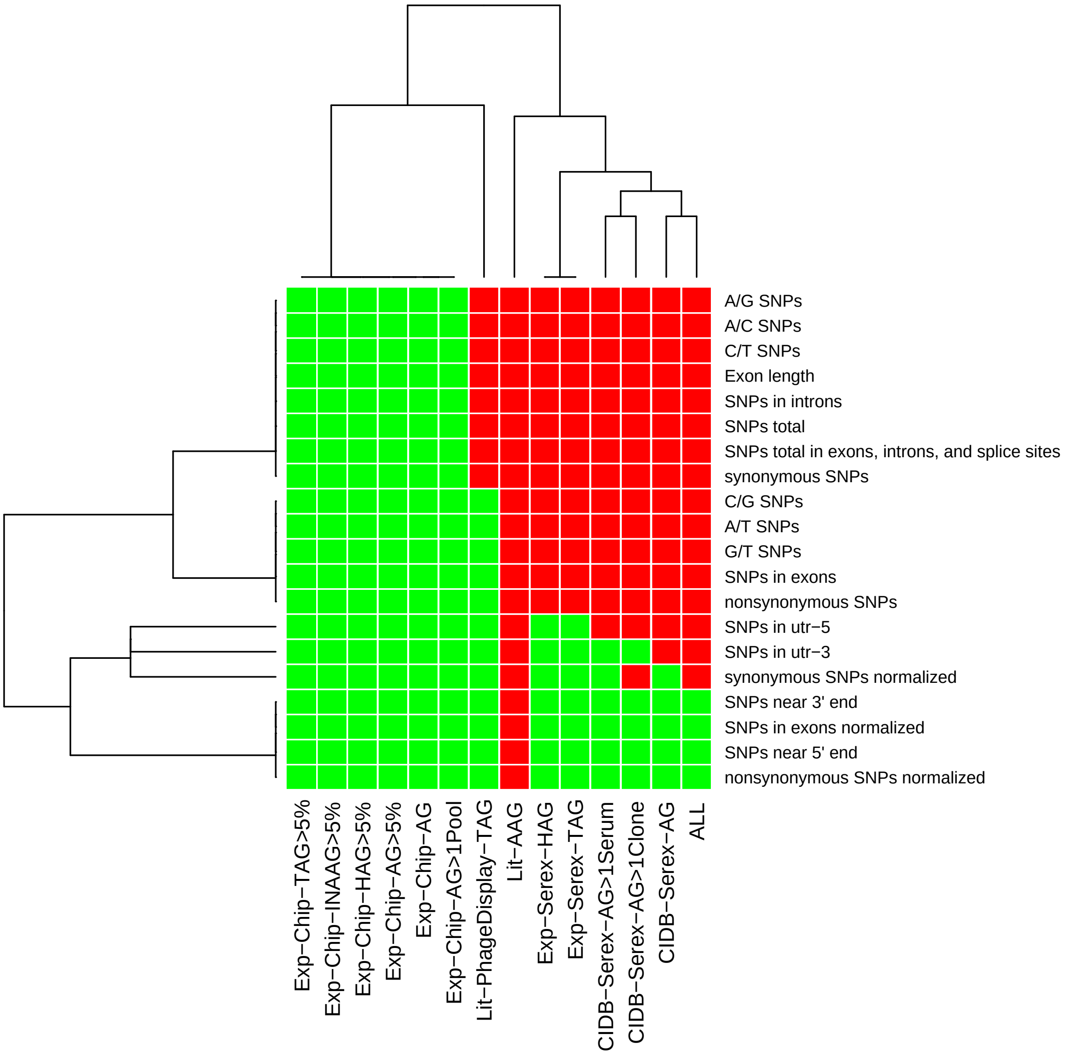


Figure 3.3: Heatmap of the results of the WMW test for comparing the distribution of SNPs in the antigen sets and the reference. Red = significantly enriched compared to the reference. Green = not significant or depleted.

As can be directly seen, there is a striking difference between the protein array derived sets and the remaining antigen sets. In the protein array set not any SNP type is significantly enriched. With our results we can also confirm the results of Stadler et al., since the Lit-AAG set contains the most significantly enriched SNP types including those already mentioned in their paper ((non-)synonymous SNPs, (non-)synonymous SNPs normalized). The SEREX derived sets show similar significantly enriched SNP types (e.g. SNPs total, synonymous SNPs, SNPs in exons, nonsynonymous SNPs). However, the exon length in these data sets is also significantly greater than in the reference, so these findings can be biased. The corresponding normalized SNP categories are not significant, with the exception of "synonymous SNPs normalized" in the ALL and CIDB-Serex-AG$>$1Clone data set. In summary, our findings confirm the results of Stadler et al. that SNPs are enriched in autoantigens. For the other antigen sets, the results may be biased by the higher-than-average exon length of the genes in the data sets and may therefore be not really relevant. Possible reasons for the separation of the protein array and the SEREX sets might be the limited selection of proteins on the chip or the bias of the SEREX method for detecting different proteins.

3.4.4 OMIM and cancer-related genes

Online Mendelian Inheritance in Man (OMIM)[5] is a comprehensive collection of human genes and genetic disorders. The focus of this database lies primarily on inherited, or heritable, genetic diseases. OMIM also contains about 900 genes that are related to the terms "oncogene" and "tumor suppressor gene" when querying the database. However, these genes may not directly serve as oncogenes or tumor suppressor genes themselves, but may be candidate tumor suppressor or oncogenes, or genes that interact with onco-/tumor suppressor genes. Unfortunately, the query cannot be more exactly specified when using the OMIM web-interface. The obtained genes were used as an additional subcategory in our analysis. We performed an ORA (described in Section 2.1.3.1) to test if there is an enrichment for disease-associated or "onco-/tumor suppressor-related genes" in our data sets.

The results of this analysis are summarized in Table 3.2. There are no significantly enriched OMIM subcategories in our antigen data sets that meet the 5% bound except the "onco-/tumor suppressor-related genes" subcategory that is significantly enriched in the antigen sets ALL, Exp-Chip-AG, Lit-PhageDisplay-TAG, and Lit-AAG. Interestingly, the autoantigen set contains about 9% of these cancer-related genes. The SEREX sets and the more specific Exp-Chip-AG sets show no enrichment for the "onco-/tumor suppressor-related

[5] http://www.ncbi.nlm.nih.gov/omim/

Table 3.2: Overview of the data sets with significantly enriched onco-/tumor suppressor-related genes. The coverage states how many genes are oncogenes or tumor suppressor-related genes in comparison to the genes in the data set that have an OMIM annotation.

Data Set	Coverage Onco-/Tumor suppressor-related genes	p-value (FDR)
ALL	128 / 1622 = 7.89%	0.037
Exp-Chip-AG	21 / 230 = 9.13%	0.037
Lit-PhageDisplay-TAG	10 / 73 = 13.70%	0.011
Lit-AAG	31 / 343 = 9.04%	0.019

genes" subcategory. On the basis of currently available data, this analysis indicates that onco-/tumor suppressor-related genes are not in general prevalent candidates for antigens.

3.5 Functional groups, processes, and subcellular location

Paul Plotz discussed in his paper "The autoantibody repertoire: searching for order" [117] several factors that might influence the selection of autoantigens, e.g., certain structural properties as the presence of coiled-coils [118] or Granzyme B cleavage sites [87]. Furthermore, certain amino acid motifs possess interesting properties that could make them possible targets for the immune system. The ELR motif is supposed to be a functional domain that bears chemotactic properties and plays a role in CXC chemokines [119]. CXC chemokines containing this motif are important for the activation of leukocytes that take part in phagocytosis of microbes and foreign antigens [120] and therefore have the ability to activate the immune system. RGD motif bearing peptides are able to directly induce apoptosis [121]. In addition to analyzing these sequence-based properties, we verified the hypothesis whether certain functional groups (GO terms), processes (GO terms, KEGG pathways) or subcellular locations play a central role in the antigen candidate selection process.

3.5.1 GrB cleavage sites, coiled-coils, amino acid motifs

As representatives for sequence-based properties of antigens, we analyzed our data sets for the presence of GranzymeB (grb) cleavage sites [43], coiled-coils [122], ELR and RGD motifs. We performed an ORA and tested if these amino acid properties are enriched in our antigen sets. The results are illustrated in Figure 3.4.

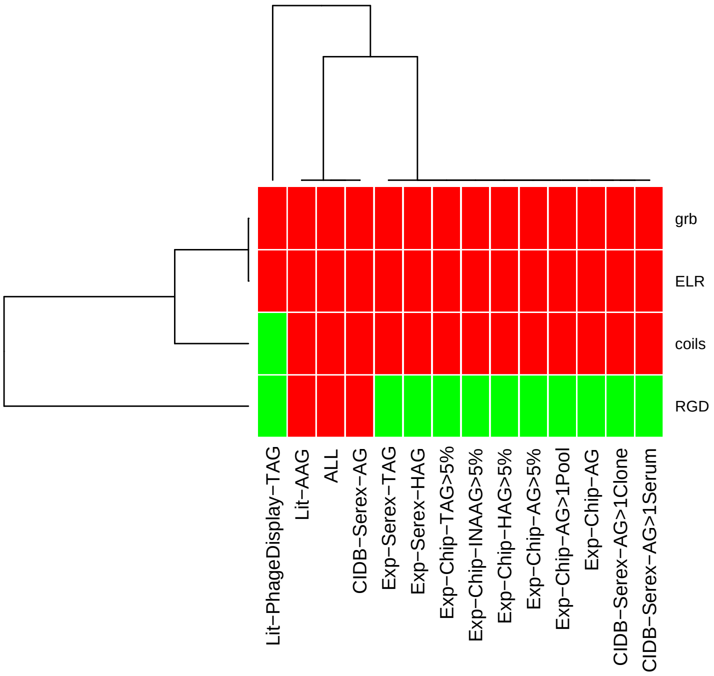


Figure 3.4: The heatmap illustrates the significantly enriched sequence features in our antigen sets. Red = significantly enriched compared to the reference. Green = not significant or depleted.

Evidently, the sequence-based properties Granzyme B cleavage sites and ELR seem to be strongly represented in the genes of our data sets. These two categories are significantly enriched for all our antigen sets. Furthermore, the coiled-coils category is enriched in all data sets except the Lit-PhageDisplay-TAG set. The RGD motif is the weakest property of these four considered, because this category is only enriched in the sets ALL, Lit-AAG, and CIDB-Serex-AG. However, these analyses are strongly dependent on which amino acid sequence is used for a gene if it has several splice-variants. In addition, the Granzyme B cleavage sites and the coiled-coils are predicted and not necessarily real cleavage sites or secondary structures, respectively.

3.5.2 KEGG

KEGG is a comprehensive database that contains regulatory as well as metabolic pathways [47, 123]. Here, we wanted to explore if our antigen sets have certain pathways in common and if these pathways are involved in immunogenic processes. We performed an ORA as previously explained and summarized the results in Figure 3.5.

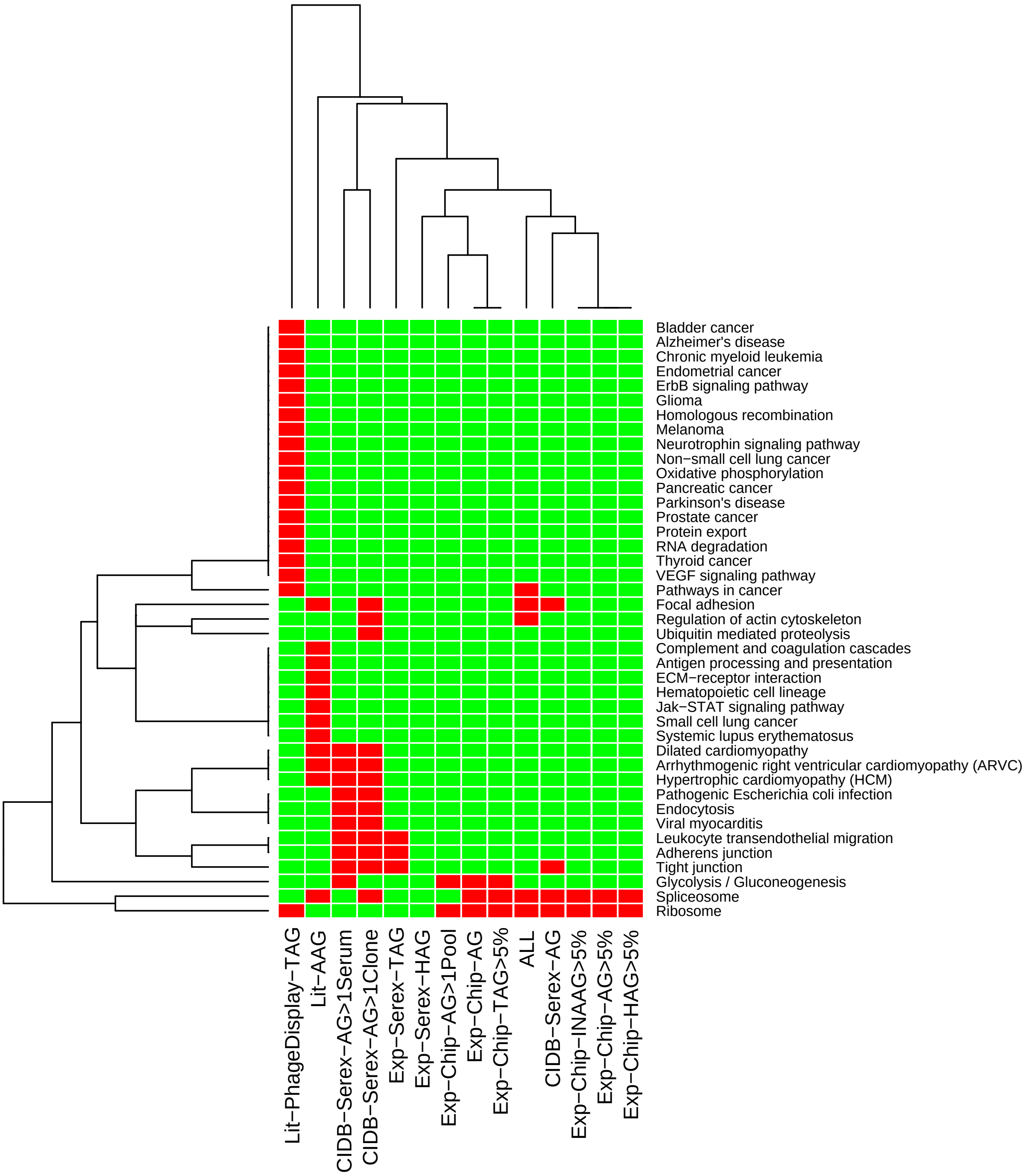


Figure 3.5: The heatmap illustrates the significantly enriched KEGG pathways in our antigen sets. Red = significantly enriched compared to the reference. Green = not significant or depleted.

When comparing the different antigen sets we find the subcategories Ribosome and Splicosome often covered throughout the different data sets. The only exceptions where none of these two subcategories is enriched is CIDB-Serex-AG>1Serum, Exp-Serex-TAG, and the Exp-Serex-HAG set, whereas the Exp-Serex-HAG set shows no enriched pathways at all. In general, we find only a few metabolic pathways enriched in our set, compared to many regulatory, signal-transduction and cancer pathways. Most of the cancer pathways that are enriched in our analysis are covered by the Lit-PhageDisplay-TAG set, which makes sense, because this set consists solely of antigens that were detected in cancer patients. The Exp-Chip-AG derived sets show almost the same enriched pathways with minor deviations with regard to the Glycolysis pathway. The CIDB-Serex-AG set and the more specific sets CIDB-Serex-AG>1Clone and CIDB-Serex-AG>1Serum show a quite different behaviour. The more specific sets show predominantly the same enriched pathways, however, the only pathway these three data sets have in common is “Tight junction”. Interestingly, the Lit-AAGs show enriched pathways that are not covered by any other antigen set. These comprise pathways of the immune system (“Complement and coagulation cascades”, “Antigen processing and presentation”, “Hematopoietic cell lineage”), the “ECM-receptor interaction”, the “Jak-STAT signaling pathway”, and the autoimmune disease “Systemic lupus erythematosus”. This supports the hypothesis that there are other/additional processes or failures of the immune system responsible for the occurrence of self-antigens in autoimmune diseases compared to cancer or healthy controls.

3.5.3 Gene Ontology

The Gene Ontology (GO) is a hierarchical collection of terms that aid to group genes or proteins in different functional groups [31]. The GO hierarchy is built of three main groups: molecular function, cellular component, biological process. For the ORA of GO terms, we used only the manually curated GO annotations, not the computationally assigned annotations (with “IEA” evidence code). Since this analysis yielded more than 450 subcategories that were significant in at least one antigen set, we decided to present the results for the three GO hierarchies separately.

The analysis of the molecular function hierarchy is illustrated in Figure 3.6. Obviously, “binding” and more specific variations of this term (especially “protein binding”, “nucleic acid binding”, “RNA binding”, “DNA binding”) seem to be a predominant property of the antigen sets. Other variations of binding can be found scattered throughout the different data sets. In contrast to the other antigen sets, the CIDB-Serex-AG>1Clone set has an

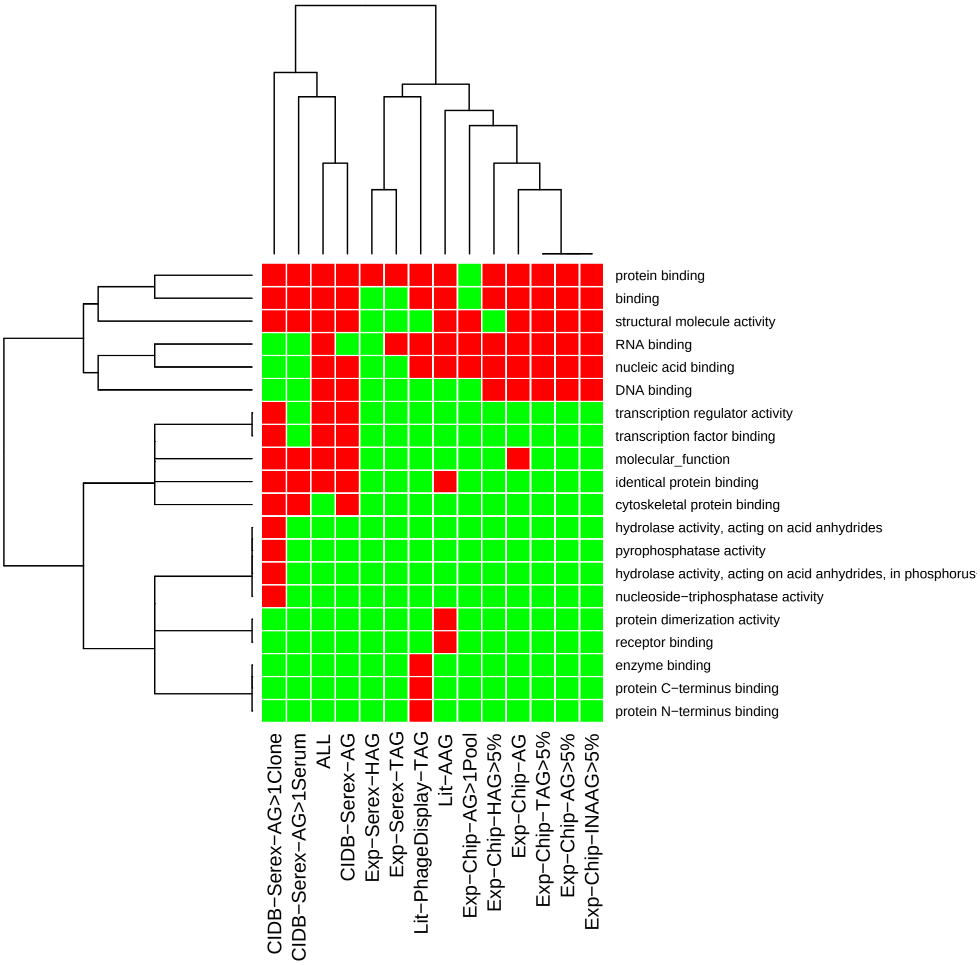


Figure 3.6: The heatmap illustrates the significantly enriched GO terms of the molecular function hierarchy in our antigen sets. Red = significantly enriched compared to the reference. Green = not significant or depleted.

enrichment in different enzymatic functions as e.g. hydrolase activity and pyrophosphatase activity. As observed previously, the CIDB-Serex-AG set and its derived sets, as well as the Exp-Chip-AG derived sets cluster together.

The analysis of the cellular component hierarchy is depicted in Figure 3.7. The Exp-Serex-HAG set shows no enrichment in any subcategory of this hierarchy meeting the 5% bound. There is a strong cluster of significantly enriched GO terms in almost all antigen set (except Exp-Serex-HAG, Exp-Serex-TAG) that comprises the components “nucleus”, “cytoplasm”, “organelle”, and derived terms. The Exp-Serex-TAG set joins the cluster of the other antigen sets with terms derived from “intracellular”. The CIDB-Serex-AG derived sets along with the ALL and Lit-AAG set show a difference to the Exp-Chip-AG derived sets and the

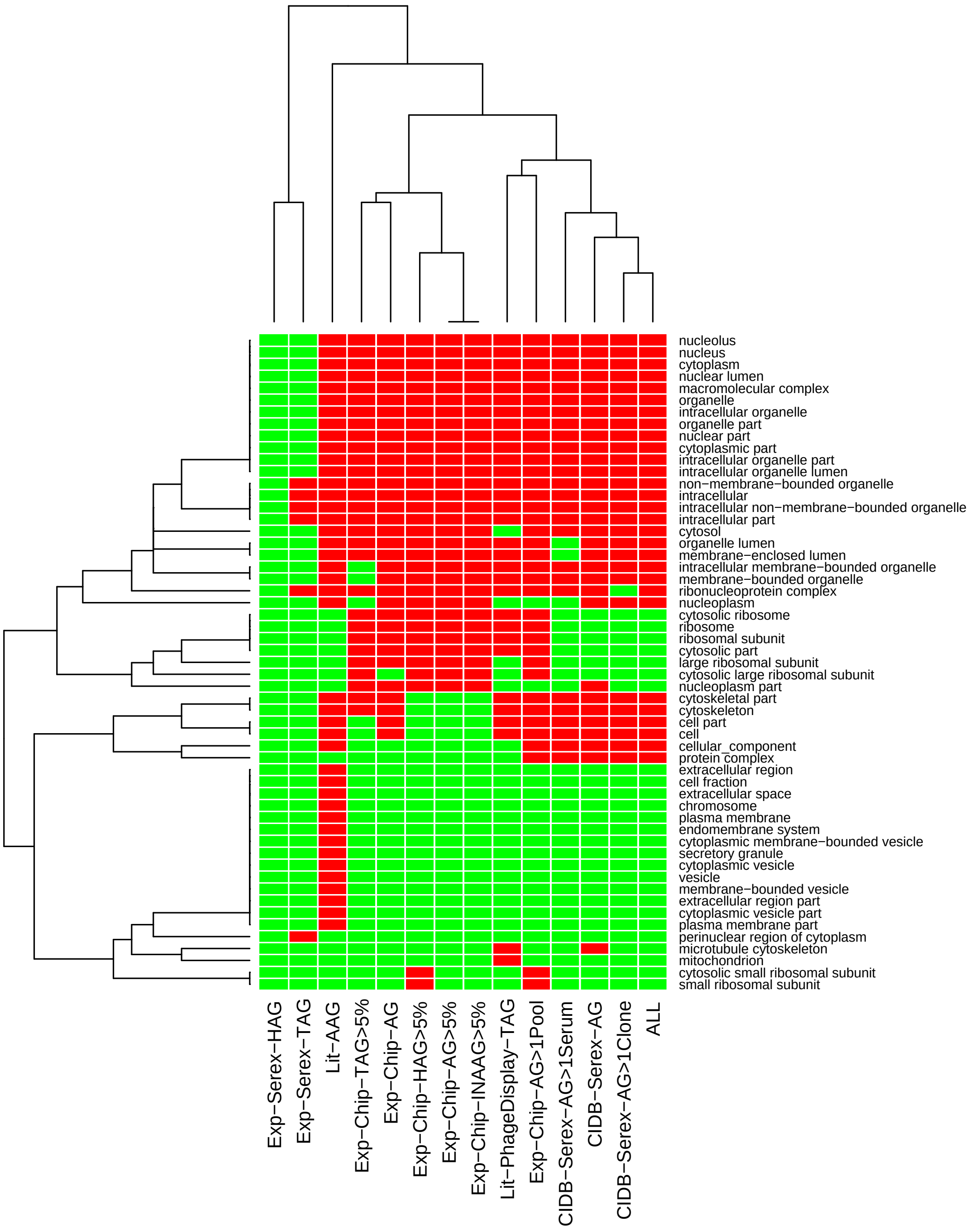


Figure 3.7: The heatmap illustrates the significantly enriched GO terms of the cellular component hierarchy in our antigen sets. Red = significantly enriched compared to the reference. Green = not significant or depleted.

Lit-PhageDisplay-TAG set in terms concerning “ribosome”, where the latter show an enrichment. Interestingly, the Lit-AAG set exhibits a collection of enriched GO terms that the other sets do not show. Amongst others, these terms belong to “extracellular region/space”, “plasma membrane part”, “vesicle”, and “secretory granule”. These findings indicate that the proteins in the Lit-AAG set have the tendency to get secreted or have an extracellular location where they have a potentially higher chance to stimulate an immune reaction.

As third analysis, we computed the significant GO terms of the biological process hierarchy, which yields the most diverse results. We find here more than 160 GO terms that are significantly enriched in at least one of our antigen sets. Since the number of GO terms is too comprehensive to be readable in the heatmap in a printed version of this thesis, we moved this illustration to the appendix for the sake of completeness (Figure D.1). The Exp-Serex-TAG and Exp-Serex-HAG set show no significantly enriched GO terms for this hierarchy. The Lit-AAG set is most different from all other antigen sets. There are more than 40 terms uniquely enriched in the Lit-AAG set comprising for example the terms “immune system process”, “immune response”, “cell communication”, “cell adhesion”, “cell differentiation”, and “response to stimulus”. Furthermore, the Lit-AAG set along with the Lit-PhageDisplay-TAG and ALL set show an enrichment of GO terms that are concerned with “apoptosis”: “cell death”, “regulation of cell death”, “programmed cell death”, “regulation of apoptosis”, etc. Some GO terms most of our antigen sets have in common are metabolic or synthetic processes like “(cellular) biosynthetic process”, “protein/cellular metabolic process”, and “RNA metabolic process”.

Taken together, the analysis of the different GO hierarchies showed that our antigen sets possess some similarities like the binding derived terms in the molecular function hierarchy or the predominant intracellular location in the cellular component hierarchy. Nevertheless, some differences between the autoantigen data set and the various tumor-/normal-antigen sets have emerged. The most striking difference seems to be that the genes of the Lit-AAG set have a direct association to the immune system. Furthermore, these genes take part in processes like cell communication and cell death, and show a tendency to get secreted or for an extracellular location. As we mentioned before, these findings indicate that the genes/proteins of the Lit-AAG set have a higher chance to interact with the cells of the immune system, which is presumably one major cause why these autoantigens are prone to elicit immune responses or at least partly explains the pathogenic effect the autoantibodies of the corresponding autoimmune diseases have.

3.5.4 Subcellular location

In addition to the GO annotations derived from the “cellular component” hierarchy, we decided to use the subcellular locations annotated from UniProt (see also Appendix C.4). To this end, we downloaded the UniProtKB/Swiss-Prot flatfile[6], parsed the necessary information and created a GeneTrail compatible flatfile. We performed an ORA of all antigen data sets against the ProteinCodingGenes as reference. The results are depicted in Figure 3.8.

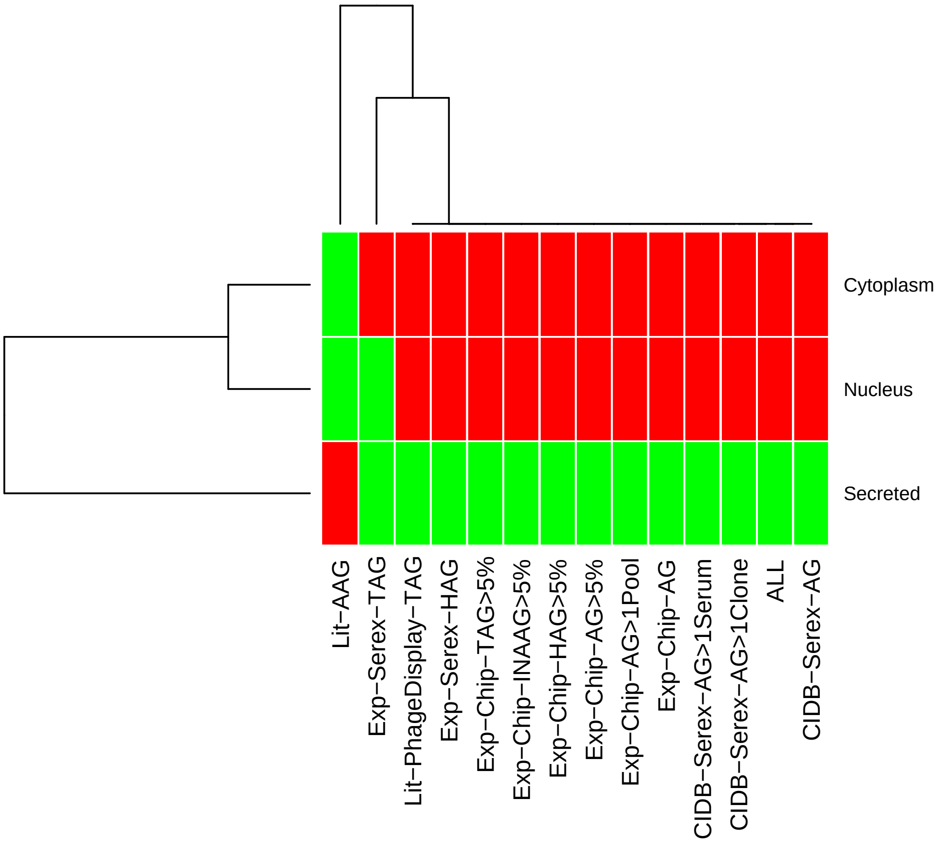


Figure 3.8: The heatmap illustrates the significantly enriched subcellular locations in our antigen sets. Red = significantly enriched compared to the reference. Green = not significant or depleted.

Interestingly, if we compare the Lit-AAG set to all other antigen sets, we almost get a complete negative image. The Lit-AAG set is only enriched for “Secreted”, whereas all other antigen sets are enriched for “Nucleus” (with exception of the Exp-Serex-TAG set) and “Cytoplasm”. On the other hand, this finding contradicts partly the results of the significant

[6] ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/uniprot_sprot.dat.gz

GO terms of the cellular component hierarchy, because we find here terms like “cytoplasm” and “nucleus” also enriched for the Lit-AAG set. However, we found GO terms – unique to the Lit-AAG set – that indicated an extracellular location or secretory property. Taken together, this confirms the assumption that autoantigens are often proteins that are more exposed to the immune system, because of their extracellular location. If the TAAs of our antigen sets become immunogenic although they are mostly intracellularly located, different processes and ways may be responsible for an immune reaction in that case.

3.6 Mimicry hypothesis

Molecular mimicry has been discussed for years in conjunction with autoimmunity [99,117,124]. The theory proposes that an infectious agent elicits an immune response and that a cross-reaction occurs because of a structural resemblance to a human protein. In this section we try to verify if the molecular mimicry hypothesis also holds for self-antigens in general. To this end, we analyze the relatedness of the proteins in our antigen sets to proteins in other organisms on different levels using miscellaneous data sources: homologs in eukaryotes, orthologs in the three kingdoms of life (Bacteria, Eukaryota, Archaea), prevalent and universal protein domains, and a BLAST analysis of human proteins against complete sequenced organisms. An overview of the terms and dependencies between homologs, orthologs, and paralogs is illustrated in Figure 3.9.

3.6.1 HomoloGene

HomoloGene[7] is a database of both curated and calculated orthologs and homologs for the organisms represented in NCBI’s UniGene database. Computed orthologs and homologs are identified from BLAST nucleotide sequence comparisons between all UniGene clusters for each pair of organisms. HomoloGene provides homologs of several completely sequenced eukaryotic genomes of which we consider here: *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Caenorhabditis elegans*, *Danio rerio*, *Arabidopsis thaliana*, *Drosophila melanogaster*, and *Saccharomyces cerevisiae*. For comparing the distribution of homologs in our antigen sets and the reference, we collected for each gene its number of homologs in the different organisms and performed a WMW test for each of these organisms. The results are summarized in Figure 3.10.

[7] `http://www.ncbi.nlm.nih.gov/homologene`

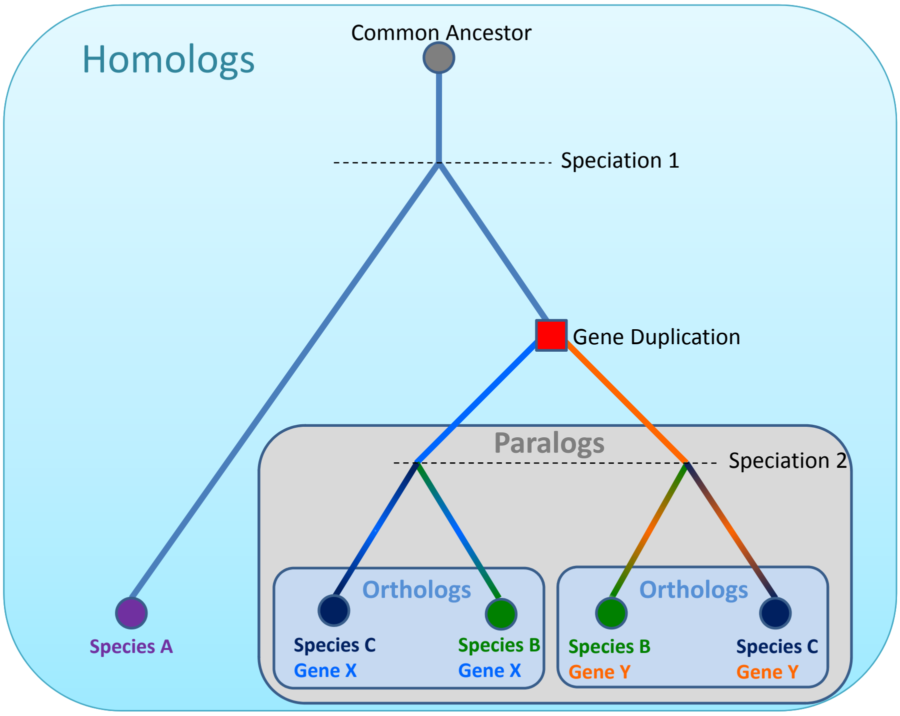


Figure 3.9: Formation of orthologs and paralogs. The evolutionary tree shows five homologous genes from three species designated A, B and C. The gene duplication event (red box) produced paralogs X and Y in the ancestor of B and C. The genes X in species B and X in species C are orthologs.

We can observe that there is in general an enrichment for homologs in different eukaryotes for the genes of our antigen sets. The Exp-Serex-HAG set shows only an enrichment for the homologs in *Danio rerio*. The sets Exp-Chip-AG, ALL, CIDB-Serex-AG, Lit-AAG, and Exp-Chip-HAG$>$5% are enriched for all tested organisms. Interestingly, we find *Rattus norvegicus* and *Mus musculus* – the two organisms which are the most closely related species to *Homo sapiens* in this analysis – enriched in the least number of antigen sets. Unfortunately, there is no real tendency recognizable when considering the taxonomic distance of the different species to *Homo sapiens*, since e.g. *Danio rerio*, *Drosophila melanogaster*, and *Caenorhabditis elegans* are less distant to *Homo sapiens* than *Saccharomyces cerevisiae* and *Arabidopsis thaliana*, but the former are enriched by more antigen sets. Since HomoloGene covers only eukaryotes, we extended our analysis to species comprising also the other parts of the taxonomic tree of life in the following subsections.

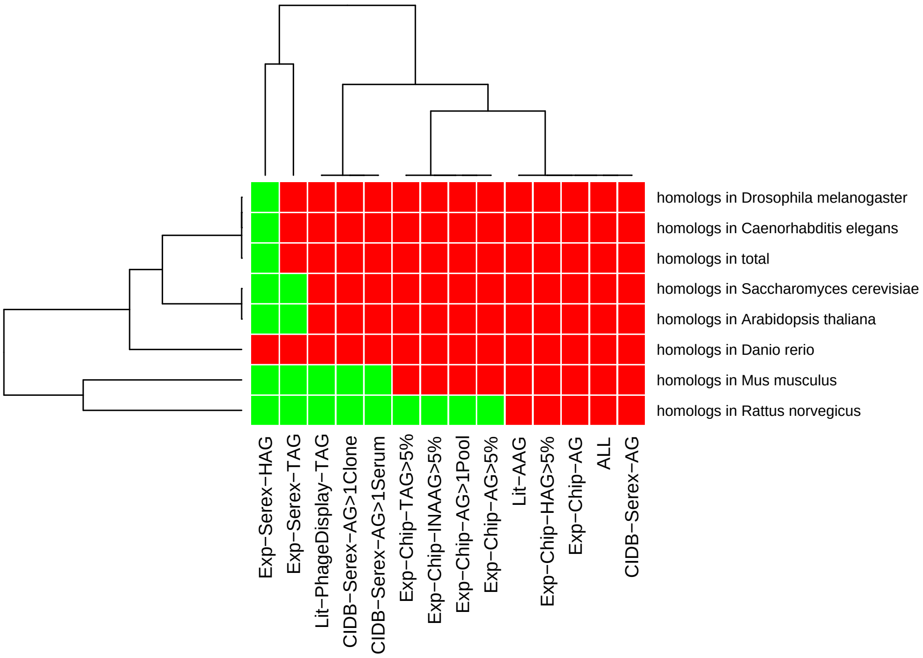


Figure 3.10: Heatmap of the results of the WMW test for comparing the distribution of homologs in the antigen sets and the reference. Red = significantly enriched compared to the reference. Green = not significant.

3.6.2 Protein families

The following analyses concern predefined “protein families” that we extracted from different data sources. We examine so called orthologous groups and protein domains that were derived from protein sequences or structures. First, we explore if there are prevalent protein families in our antigen sets. Second, we analyze if there is an accumulation of universal or ancient protein families.

3.6.2.1 Orthologous groups

Orthologous Groups (OGs) consist of orthologous proteins from different organisms. The Clusters of Orthologous Groups of proteins (COGs) are based on protein sequence comparisons of complete sequenced genomes comprising prokaryotes and unicellular eukaryotes [125]. An extension of this system was applied to construct clusters of predicted orthologs of different eukaryotic genomes, named KOGs after eukaryotic orthologous groups [126]. By contrast, NOGs (non-supervised orthologous groups) are orthologous groups which are assembled automatically by computer-aided inference of functional categories taken from the original COG/KOG databases. The extended groups are deposited in the eggNOG database [127, 128]. We used the extended OGs from eggNOG since they provide a more current and comprehensive coverage of organisms than the original COGs/KOGs.

3.6.2.2 Protein domains

We analysed structure-based and sequence-based protein domains from CATH [129] or Pfam [130, 131], respectively, for enrichment in our data sets. The Pfam database consists of conserved protein families and domains. The protein families are derived from sequences deposited in UniProt with each family represented by multiple sequence alignments and profile hidden Markov models (HMMs). For our analyses we used Pfam-A, which consists of high quality, manually curated families. CATH is a database of manually derived structural domains from the Protein Data Bank (PDB) [132] that are placed within a hierarchy including topology, homology, and conservation. The CATH database contains only crystal structures from the PDB with a better resolution than 4.0 angstroms, together with NMR structures. Unfortunately, the CATH domains are only available for about 2000 human proteins. Therefore, we decided to use additionally the CATH domain annotation generated by Gene3D [133, 134]. Gene3D provides comprehensive structural and functional annotation of most available protein sequences, including the UniProt, RefSeq and Integr8 resources. The main structural annotation is generated through scanning these sequences against the CATH structural domain database profile-HMM library. Hence, Gene3D transfers the structural annotation to thousands of sequences resulting in an annotation of about 9500 human proteins. The advantage of using a structure-based domain database is the capability to infer more ancient and divergent homology relationships than with using solely sequence-based approaches. For the following analyses of structure-based protein domains we extracted the CATH domains deposited in the Gene3D database v5.2.0 (updated August 2007).

3.6.2.3 Prevalent enriched protein families

To analyze if there is a prevalence of protein families consisting of a certain structure or sequence, we subjected our data sets to an ORA for CATH, Gene3D and Pfam domains, and the orthologous groups from eggNOG. To this end, we extracted for each gene/protein, of which protein families it is composed and saved the information for each data source separately in a GeneTrail compatible file format. Since we are especially interested in protein families that occur frequently among the antigens, we only present here significantly enriched domains that appear in at least 5% of the annotated proteins of one data set. The results are summarized in Figure 3.11.

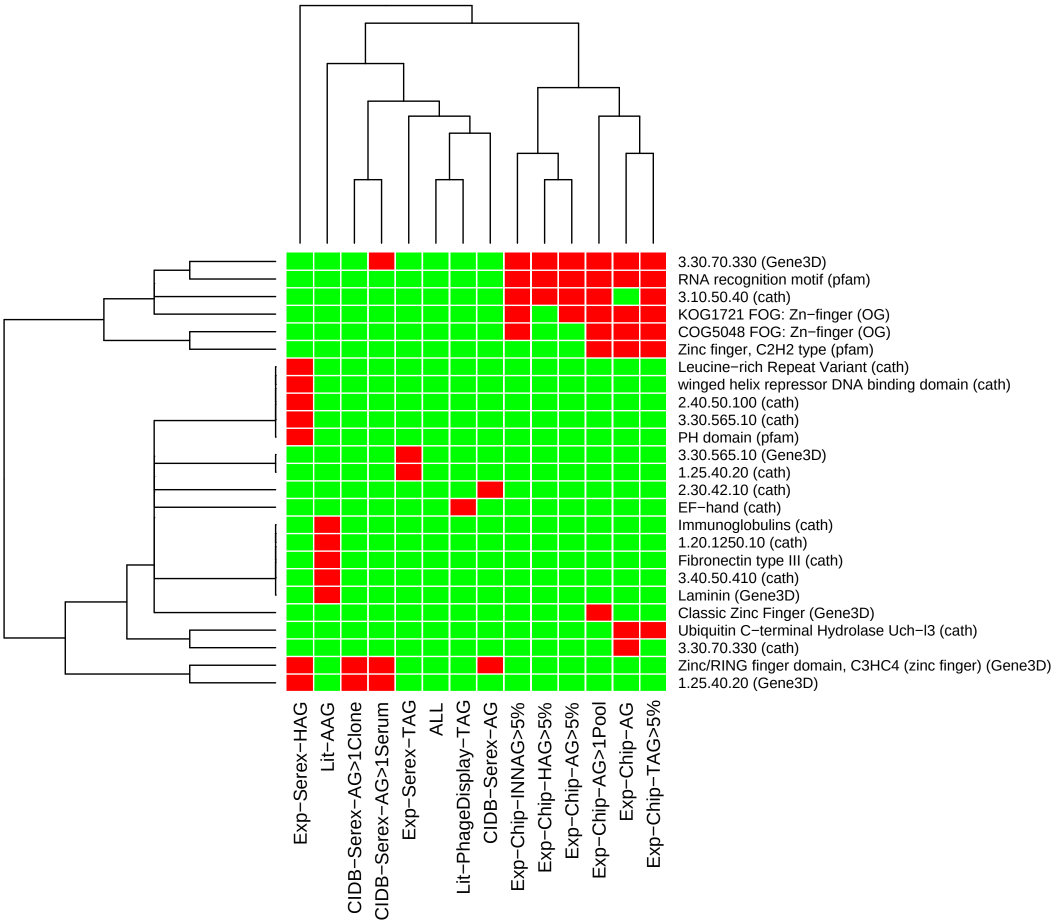


Figure 3.11: The heatmap illustrates the significantly enriched protein families (from CATH, Gene3D, Pfam, eggNOG) in the antigen data sets. Red = significantly enriched compared to the reference. Green = not significant.

There are only a few domains of several thousand that meet the 5% threshold in our data sets. This analysis separates our antigen sets into two clusters, the first consisting of the Exp-Chip-AG derived sets and the second consisting of the remaining sets where the SEREX method prevails. The Exp-Chip-AG derived sets show an enrichment for protein families with a Zinc finger motif or an RNA recognition motif. Furthermore, the Exp-Serex-HAG and CIDB-Serex-AG derived sets are enriched in a CATH domain named “Zinc/RING finger domain”. Interestingly, Zinc finger motifs are in general DNA-binding motifs that are often found in transcription factors. These findings also confirm our previous GO results where “binding” and its variations were predominantly present in our antigen sets.

3.6.2.4 Analysis of universal protein families

Following the mimicry hypothesis, we investigated whether our antigen data sets contain predominantly “ancient” protein families. Lee et al. analyzed the distribution of different domain architectures in completed genomes from all kingdoms of life [135]. In 2005, they extracted 219 domain families that were found in at least 70% of the genomes from each of the three kingdoms of life and hypothesized that these domains may correspond to universal families with essential functions.

For our analysis, we extracted not only for each gene/protein such protein families occurring in at least 70% of the genomes from each of the three kingdoms of life (UNIVERSAL_INTERSECT), but also families occurring in at least 70% of the genomes of one kingdom (UNIVERSAL_BACTERIA, UNIVERSAL_ARCHAEA, UNIVERSAL_EUKARYOTA), or of at least one kingdom (UNIVERSAL_UNION). Additionally, we had to exclude such organisms that were not completely sequenced or those having only a very small number of protein family annotations. The first problem was solved using the list of completely sequenced and published genomes from the Genomes OnLine Database (GOLD) v3.0[8] comprising 742 organisms. The second problem predominantly affected the Pfam domains. When including too many organisms with a low domain annotation, we had only a few universal domains meeting the 70% restriction. Excluding too many organisms led to a bad distribution of organisms for the different kingdoms of life. Therefore, we decided to use a threshold of at least 150 protein families per organism. This way, we had a similar distribution of organisms for CATH and Pfam domains (Table 3.3) and an average number of domains per taxon of about 450. The orthologous groups from eggNOG were not affected by this threshold. These are widely spread through the different organisms, but often contain only a very low number of genes or proteins, respectively.

The distribution of the universal protein families in the different data sources for the three kingdoms of life separately, unified, and intersected is depicted in Table 3.4. The UNIVERSAL_INTERSECT class is the most restrictive of these groups, the UNIVERSAL_UNION class the least restrictive. As denoted previously, the NOGs have almost a 1:1 relation of genes to protein families, i.e. an OG often consists only of one gene per organism. Therefore, these OGs are very organism and gene specific. In addition, this table shows that

[8] http://genomesonline.org/index2.htm

Table 3.3: Number of complete sequenced organisms in the three kingdoms of life having at least 150 protein family annotations for Pfam, CATH (Gene3D), or NOG.

	Bacteria	**Eukaryota**	**Archaea**
Pfam	313	22	15
CATH	202	21	20
NOG	374	20	36

the structure-based CATH domains from Gene3D cover much more genes in the different "universal" classes than the sequence-based domains from Pfam. This supports the theory that these structure-based domains are more suitable to infer divergent homologies than sequence-based approaches.

Table 3.4: Overview of the distribution of universal protein families from CATH (Gene3D), Pfam, and eggNOG in human genes. We considered only such organisms that were completely sequenced and had at least an annotation for 150 protein families for each data source. The protein families had to be present in at least 70% of: one kingdom of life (UNIVERSAL_BACTERIA, UNIVERSAL_ARCHAEA, UNIVERSAL_EUKARYOTA), at least one kingdom of life (UNIVERSAL_UNION), each kingdom of life (UNIVERSAL_INTERSECT).

	CATH (Gene3D)	**Pfam**	**eggNOG**
UNIVERSAL_INTERSECT	139 families in 2796 genes	4 families in 8 genes	185 families in 638 genes
UNIVERSAL_UNION	609 families in 7545 genes	337 families in 3201 genes	12792 families in 10644 genes
UNIVERSAL_EUKARYOTA	476 families in 7479 genes	42 families in 2597 genes	11903 families in 10634 genes
UNIVERSAL_BACTERIA	290 families in 3202 genes	223 families in 508 genes	600 families in 1399 genes
UNIVERSAL_ARCHAEA	257 families in 3381 genes	137 families in 315 genes	1061 families in 1130 genes

Performing an ORA, we tested the hypothesis if there is an enrichment for universal protein families in our data sets using the above collected UNIVERSAL sets as special subcategories of the corresponding protein family category (Pfam, Gene3D, eggNOG). The heatmap in Figure 3.12 summarizes the results of this analysis.

In general, we notice an enrichment for the different universal classes throughout our tested antigen sets. Each of our data sets shows an enrichment in at least three universal classes. However, the data sets are again clustered in two major groups. The first consists of the Exp-Serex-HAG, Lit-PhageDisplay-TAG, Lit-AAG, and Exp-Serex-TAG set, the second of the remaining ALL and CIDB-Serex-AG/Exp-Chip-AG derived sets. The latter group shows an enrichment for almost all tested universal classes with some minor exceptions, whereas the enrichment of the universal classes in the former group is scattered and restricted to a few cases. The UNIVERSAL_INTERSECT class for Pfam and NOG families was

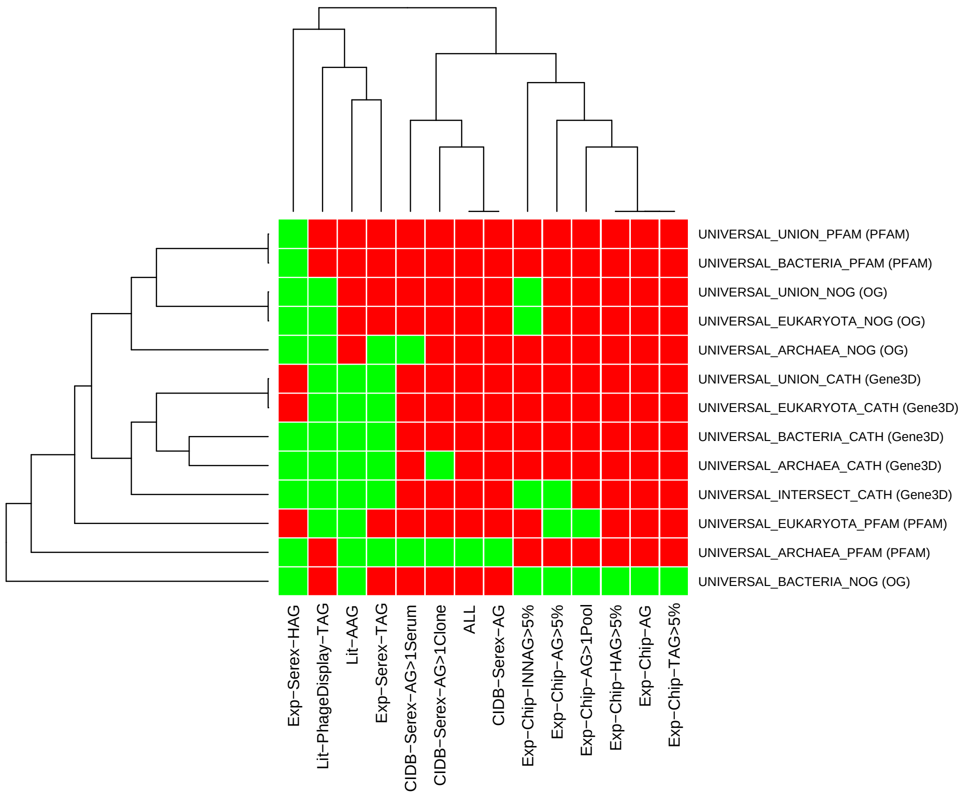


Figure 3.12: Heatmap summarizing the enriched universal protein families in the considered data sets. Red = significantly enriched compared to the reference. Green = not significant.

not enriched for any data set. By contrast, the corresponding class for CATH domains is enriched in most of the antigen sets of the second group as are the remaining CATH universal classes. Interestingly, the CATH universal classes are not enriched in the first group with the exception of UNION and EUKARYOTA for the Exp-Serex-HAG set.

3.6.3 BLink and BLAST

In the following, we explored whether our antigen sets have more similar sequences in other organisms than the reference using the Basic Local Alignment Search Tool (BLAST) [136]. BLAST is a well-established method for finding local sequence similarities of a search pattern against a database of sequences. BLink (”BLAST Link”) is available online on the NCBI homepage and displays the results of BLAST searches that have been done for every protein sequence in the Entrez Proteins data domain. Some features of BLink are the graphical presentation of pre-computed “blastp” results against the protein non-redundant (nr) database, and the display of the number of organism hits, the number of protein hits, and the number of overall hits. We downloaded the information for 21441 proteins in total. For each of these proteins, we collected the information of how many hits in total the sequence of this protein has (all hits are counted, even several hits to the same protein), the number of proteins it is similar to (only one hit per protein is counted), and the total number of hits in different organisms (each organism is only counted once), each for a similarity score threshold ≥ 100.

When we analyzed our antigen sets with the Wilcoxon-Mann-Whitney test against the ProteinCodingGenes reference for these numbers, we found that we have a significant enrichment in our data sets concerning number of hits, proteins, and organisms. In addition, we analyzed the number of splice variants of the genes of our antigen sets, which were also enriched in all of our antigen sets compared to the reference. Since we cannot distinguish with the downloaded information which organisms were completely sequenced and to which taxonomy they belong, we performed a BLAST analysis of the 21441 proteins against the protein sequences from RefSeq release 30 (including sequences from 5395 different organisms). Furthermore, the BLAST analysis has the advantage that we can also find sequence similarities that do not belong to a pre-defined functional domain, but may also be candidates for eliciting immune responses via molecular mimicry. In brief, we extracted for each of the 21441 proteins the BLAST hits that had at least a similarity score of 100 and at most an E-value of 0.001. To retrieve the information, to which kingdom of life these hits belong, we mapped the hits to their corresponding organisms. Additionally, we investigated, if there are certain taxa (families, classes, etc.) that show a significant accumulation of hits. To this end, we built up the taxonomy tree as follows and considered each node as a subcategory for an ORA: The organisms represent the leaves of the taxonomy tree, internal nodes represent taxa that group the organisms hierarchical into classes, orders, families, etc. In a first step, we added the human proteins – we performed the BLAST analysis with – to the organism nodes they had a hit. In an additional step, we traversed the taxonomy tree from the leaves to the root and assigned each internal node the union of the proteins of its children. Hence, we are able to analyze if there is a significant accumulation of sequence similarities for the different taxonomic lineage levels when using the protein sets of the internal nodes as subcategories. For excluding hits to not completely sequenced organisms, we filtered the taxonomy tree using the list of completely sequenced and published genomes from GOLD.

Since the ORA performed for all nodes of the complete taxonomy tree yielded too many significantly enriched taxa, we decided to analyze the kingdoms Bacteria, Archaea, and

Eukaryota separately to get a better overview. The number of nodes and leaves (organisms) of the three disjoint subtrees are summarized in Table 3.5. The table shows that we had the most BLAST hits in Bacteria, followed by Eukaryota, and Archaea. However, this mirrors primarily the distribution of the kingdoms in the list of completed genomes.

Table 3.5: Number of nodes and leaves that had BLAST hits for the disjoint subtrees of the three kingdoms of life.

	Bacteria	**Eukaryota**	**Archaea**
leaves (organisms)	447	82	39
nodes	1272	509	154

Interestingly, the ORA performed for the three trees showed that almost all possible taxa were significantly enriched in at least one of our data sets. In detail, we found 1248 of 1272 possible taxa significantly enriched in the kingdom Bacteria for at least one of our antigen sets. We obtained similar numbers for Archaea (150/154) and Eukaryota (408/509). When we restricted the ORA to the leaf nodes, we found 61/82 for Eukaryota, 441/447 for Bacteria, 39/39 for Archaea significantly enriched in at least one tested set. However, these findings may strongly depend on the thresholds we have chosen for the BLAST analysis. We used a lower bound for the BLAST score of 100 to be able to compare the results to BLink and because we also wanted to include alignments of shorter lengths in our analysis. The upper bound of 0.001 for the E-value should be sufficient to exclude coincidental findings on a large scale. In the following, we will briefly discuss the results for the enriched eukaryotes (Figure 3.13).

Based on the findings for the protein families and the assumption that the results are not biased, we observe that the eukaryotic organisms are predominantly enriched for our data sets. The Exp-Serex-HAG set shows the lowest number of enriched organisms, followed by the Lit-AAG, Lit-PhageDisplay-TAG, and Exp-Serex-TAG set. The remaining sets present almost a uniform image of enriched organisms. Taking a closer look at the types of organisms included in Figure 3.13, we find well-known representatives of parasites. In addition, these parasites had most often the lowest p-values for our different data sets, e.g., *Theileria parva strain Muguga*, *Theileria annulata strain Ankara*, *Plasmodium falciparum 3D7*, *Plasmodium yoelii yoelii str. 17XNL*, *Cryptosporidium parvum Iowa II*, *Entamoeba histolytica HM-1:IMSS*, *Cryptosporidium hominis*, and *Brugia malayi* to mention the most important of these parasites.

The results for the different Archaea and Bacteria species showed a more differential pic-

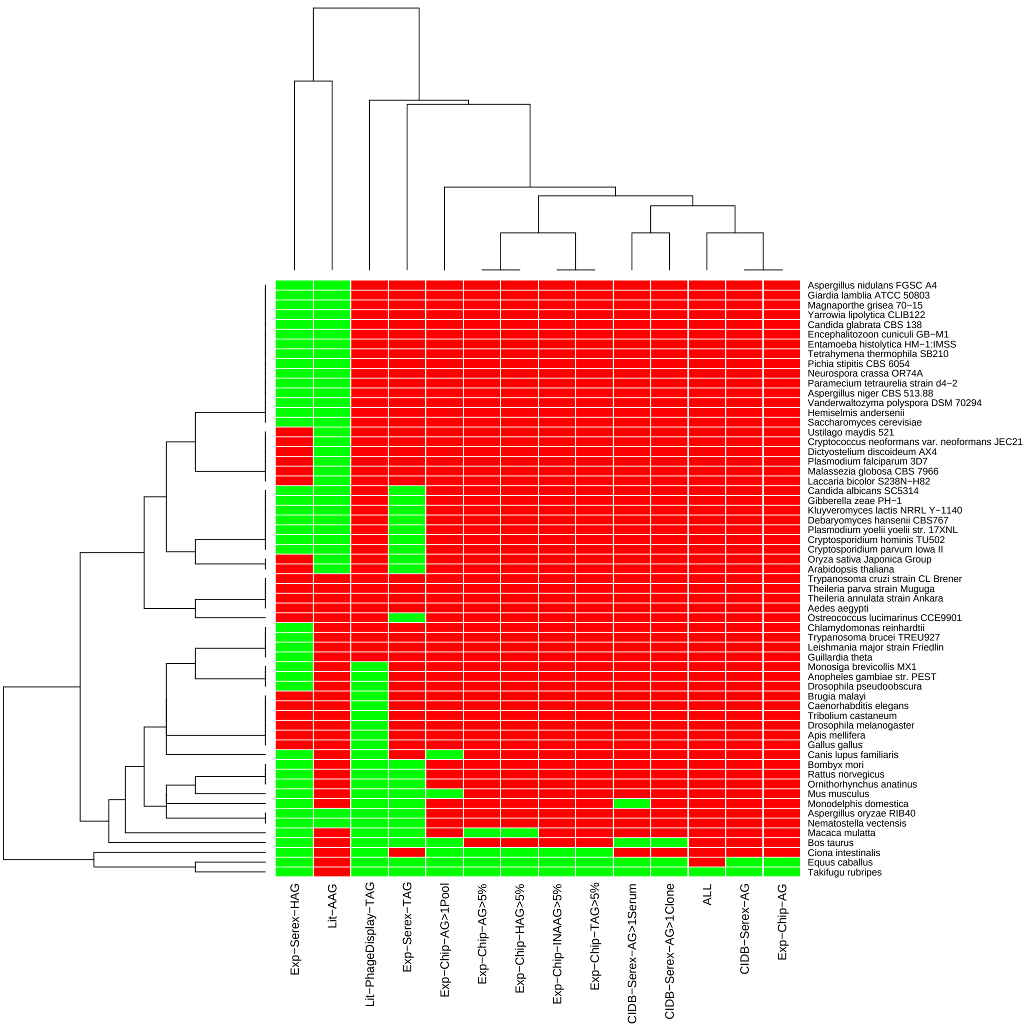


Figure 3.13: Heatmap summarizing the enriched eukaryotic organisms in the considered data sets. Red = significantly enriched compared to the reference. Green = not significant.

ture. The Exp-Serex-HAG and Lit-PhageDisplay-TAG set showed no enrichment at all for Bacteria or Archaea. The ALL, CIDB-Serex-AG and its derived sets had the most enriched organisms for Bacteria, whereas ALL, CIDB-Serex-AG, Exp-Chip-AG and its derived sets had the most significantly enriched organisms for Archaea. Taken together, these results confirm in general our previous findings that there is a certain similarity of the proteins in our antigen sets especially to proteins in other eukaryotic species, in which parasites seem to play an important role.

3.7 Feature subset selection: differences between TAAs and AAGs

Besides testing different hypotheses for the possible causes of humoral immune responses in autoimmune diseases and cancer, we wanted to verify if the analyzed properties from above are suitable to discriminate between tumor associated antigens and autoantigens. To this end, we applied a classification based on a so called “feature subset selection” (FSS), a method that is widely employed in bioinformatics to reduce the number of features available to the ones that are relevant and sufficient to discriminate between different classes. In our case, these classes correspond to the labelling of a gene as “autoantigen” or “tumor antigen”.

The classification was performed with a Naïve-Bayes Classifier. For the FSS, we apply here a method developed by Andreas Keller, Alexander Rurainski, and Matthias Hein [137,138]. In their approach, they use the mutual information as measure for the computation of the statistical dependency not only between features and class labels, but also between the features themselves. The problem of finding the subset maximizing the statistical dependency can be formulated as a quadratic 0-1-program. As solver for the quadratic 0-1-program, the commercial software CPLEX[9] (version 11.1.1) was used. The quality of the selected features was estimated by computing the classification accuracy with the Naïve-Bayes Classifier performing a 10-fold cross-validation. For the technical details of the computation and more detailed background information we kindly refer the interested reader to the PhD thesis of Alexander Rurainski [138].

As input for the algorithm, we assembled a feature matrix containing the following properties for each gene: SNP counts (dbSNP), somatic mutation counts (RCGDB), number of homologs in different eukaryotes (HomoloGene), BLink hits, number of splice variants, oncogene / tumor suppressor-related gene (OMIM), and locations (UniProt). Because not all of these features are categorical, a pre-processing step to bin the data was performed where necessary. In total, we tested three different classification problems. For the first problem, we assembled a feature matrix containing the autoantigens from the Lit-AAG set that had no overlaps with any of our other antigen sets, and the remaining genes from the ALL set that had no overlap with the Lit-AAG set. Since the ALL set also contains the Exp-Serex-HAG set consisting of genes found in healthy persons, we assembled a second and a third matrix where we added the Exp-Serex-HAG genes to the autoantigens or

[9] `http://www-01.ibm.com/software/integration/optimization/cplex/`

removed them completely from the input. Hence, we were able to verify if these natural occurring antigens are more similar to autoantigens or tumor antigens. The classification accuracy and the corresponding most commonly selected features for these three cases are summarized in Table 3.6.

Table 3.6: Overview of the classification accuracy and the selected features for the three different autoantigen versus tumor antigen matrices.

	AAGs_vs_TAAs_plusHAGs	**AAGs_plusHAGs_vs_TAAs**	**AAGs_vs_TAAs**
Classification Accuracy	88.76%	86.44%	88.22%
Selected Features	somatic mutation "unknown" and location "Secreted"; somatic mutation "unknown" and location "Cytoplasm"	somatic mutation "unknown" and location "Secreted"	somatic mutation "unknown" and location "Secreted" and sequence feature "coiled coils" and "onco-/ts related gene"

Interestingly, we find that the classification accuracy is the highest when we count the genes of the Exp-Serex-HAG set to the TAAs. On the basis of this data, we can discriminate autoantigens from tumor antigens with a classification accuracy of about 88% using only two features: either somatic mutation "unknown" and location "Secreted" or somatic mutation "unknown" and location "Cytoplasm", which had the same number of occurrences in the 10-fold cross-validation. The location difference has already been obvious in our previous analysis. By contrast, the meaning of the somatic mutation "unknown" is elusive, but it may indicate that somatic mutations play a more important role than previously supposed. However, the data source for the somatic mutations is the "Roche Cancer Genome Database", from which we extracted the number of occurrences of different somatic mutations in cancer, which may lead to a certain bias when comparing antigens from cancer patients and autoimmune diseases, where we do not have informations about the somatic mutation state. Unfortunately, we could not include all of the above analyzed properties or perform the feature selection for TAAs versus non-TAAs, because of memory restrictions of the FSS algorithm.

3.8 Discussion and Conclusion

In this chapter, we presented a comprehensive analysis of a collection of putative properties of antigens. In recent years, different hypotheses and models have been discussed for the immunogenicity of self-antigens. The molecular mimicry hypothesis has been the target of diverse arguments in the context of autoimmune diseases. On the one hand, evidence exists for single cases of autoimmunity that the infection with a pathogen and the following cross-reactivity with a self-antigen was the autoimmune response eliciting factor, e.g., the gastric autoimmunity that is associated with *Helicobacter pylori* antigens [139]. On the other hand, this theory cannot account for the seemingly limited repertoire of autoantibodies associated with human diseases [117].

Considering the results of this work, we could argue that because of the general increased similarity of the proteins in the antigen sets to proteins in other species, these have a higher probability to become immunogenic than proteins that are more specific for human. The adaptive immune system must be flexible enough to detect a wide range of possible pathogenic targets, even those that are similar to self-antigens. However, this flexibility comes with the risk of autoimmune diseases [140]. One way, in which proteins of cancer cells can stimulate an immune reaction is by necrotic processes or a defective apoptosis. While apoptosis is normally an anti-inflammatory process [141], where cell debris is removed by phagocytic cells, an abnormal apoptosis could lead to APC activation and presentation of self-antigens. Furthermore, necrosis is in general a pro-inflammatory process [142] that can occur during tumor growth and exposes the contents of the cell to the immune system. Hence, those proteins that possess a high similarity to foreign proteins may be more susceptible to elicit immune responses against self-antigens than others.

By contrast, recent findings suggest that not the structural differences define self and foreign antigens, but the strength of avidity during the thymocyte activation steps [143–145]. Furthermore, the diversity of the immunoglobuline repertoire is of the order of at least 10^6 possible combinations and is supposed to be “sufficiently large to recognize, with moderate affinity, essentially any molecular shape” [146]. If this is the case, sequence similarities between self and foreign antigens should not be the main decisive factor, but are probably a secondary side effect of other influences.

In this work, we analyzed several of these factors in question. Interestingly, sequence-based properties, such as Granzyme B cleavage sites, coiled-coils, and ELR motifs seem to be dominant in our antigen sets. However, we should keep in mind that the cleavage sites and the coiled-coils are based on predictions that may be more or less reliable [43, 122]. When considering pathways, GO terms, and locations, we highlighted several differences and similarities between autoantigens and tumor antigens. Autoantibodies in autoimmune diseases can be directed against intra- and extracellular targets, whereas extracellular targets can often be directly linked to the pathogenesis of the disease [147]. In contrast, the antigens contained in our TAA sets are predominantly located intracellular. Furthermore, we showed for the first time that tumor antigens and autoantigens can be discriminated by a few features, such as the locations “Secreted” and “Cytoplasm” and an “unknown” somatic mutation. Using these features we can correctly classify a gene as autoantigen or tumor antigen with an accuracy of about 88%. Naturally occurring antigens seem to be more similar to TAAs than to autoantigens when comparing the classification accuracies and the number of selected features. However, we will have to confirm these preliminary findings, because we extracted the somatic mutations from cancerous diseases and may have introduced a bias when using these informations as features. In addition, the sets may give some sort of biased impression depending on their sources, e.g., from literature or from different experimental methods, or depending on the types of features selected as input for this analysis.

Considering tumor associated antigens in detail, we hypothesize that the genes found immunogenic in cancer underlie a certain selection pressure that makes them more susceptible for genetic alterations or altered expressions. Such genes must either be key players influencing directly crucial cellular processes like the apoptosis or the cellular proliferation or contribute indirectly to these processes, e.g., by regulating expression and translation of other genes. We found that the TAAs in our data sets are enriched for molecular functions such as binding, protein binding, DNA and RNA binding. Furthermore, at least a significant part of our antigen sets showed an enrichment for Zinc-finger motifs that are often found in transcription factors. These have a high chance to be over-expressed themselves in cancer driving the proliferation in these cells. Another interesting result is that we find many proteins involved in ribosomes in our antigen sets. Ribosomes are discussed to play an active role in tumorigenesis [148, 149]. Antibodies against famous key players as p53 are frequently found in different cancer types [100]. Some antibodies even occur frequently and specifically in certain cancer types and are potential diagnostic biomarkers [150]. Following our theory, the corresponding antigens must have a certain crucial function in theses cancer types. However, since tumors are in general quite heterogeneous entities as is the immune system unique in each individual, we will not always find the same antibodies for the same cancer types in different persons. Tan et al. also reported that TAAs are often proteins that play a crucial role in carcinogenesis and presented in their paper a detailed overview [151].

A crucial point that influences the results of all performed analyses is the selection of antigen sets that were used in this thesis. As we have seen, the properties of the antigen sets seem to be at least in part dependent on their experimental isolation technique, since the antigens derived from the SEREX method and the protein chip often built separate clusters in our analyses. Furthermore, most of the considered antigens were detected with few sera and the mode of detecting positive antigen-antibody reactions during isolation is commonly error-prone. Taking these factors into consideration, we were still able to gain significant insights in a highly complex field of research that will probably improve with the increase of data in the future.

Taken together, we provided further indications for differences and similarities in tumor antigens and autoantigens. However, the picture that emerged is by far not complete. More effort and research will be necessary to deepen our understanding in this area of research and to reveal the processes of antigen candidate selection.

CHAPTER 4

miRNAS AND CANCER

Despite recent advances in sequencing methodology, microarray expression profiling is still a commonly applied technique for studying natural and pathogenic biochemical processes. While in the past decade the analysis of coding RNA molecules, mostly messenger RNAs (mRNAs) were in the focus of research, the relevance of non-coding RNAs has not been realized as of recent years. Especially microRNAs (miRNAs) are of increased interest. These endogenous non-coding small RNAs usually of length 19 to 23 nucleotides are known to regulate the translation of the coding mRNAs in a sequence-specific manner, e.g. through binding and enabling the degradation or silencing of their target mRNAs [27]. miRNAs seem to be involved in almost all biological processes, including cellular development, differentiation, proliferation or apoptosis [152, 153]. Evidently, these molecules also play an important role in cancer, as recently reviewed by Drakaki et al. [28]. A variety of studies describe that miRNAs can function either on tumor suppressor genes or on oncogenes and thus acting as major regulators of gene expression. While they were so far considered to be negative regulators, recent studies impressively demonstrate that miRNAs can also have positive effects on gene expression [29].

In addition to experimental approaches for the identification of miRNA targets, a variety of computer-aided target prediction algorithms have been developed [154–157]. These algorithms are trained by well-known miRNA-mRNA interaction rules gained from microarray data in order to identify novel miRNA targets. One of the most comprehensive resources for miRNA targets is MicroCosm, a web resource developed by the Wellcome Trust Sanger Institute and now hosted by the European Bioinformatics Institute (EBI) containing computationally predicted targets for miRNAs across many species. The targets of MicroCosm have been predicted with the miRanda algorithm [154]. As recently reviewed by Bartel [158], several other methods exist that either use conservation information including TargetScan [159], PicTar [160], or PITA [161] or do not rely on this conservation information as RNA22 [162]. The analyses in this work rely on MicroCosm, because (1) this algorithm acknowledges complementarity at the $5'$ end of the microRNA, where a rather strict complementarity is required, (2) excludes non stable conformations by using the Vienna RNA folding approach, and (3) in addition checks whether the site is conserved in orthologous transcripts from other species.

To further improve our understanding of the mode of action of miRNAs and their function, gene set analysis based approaches can be used. Most recently, the group of Hatzigeorgiou proposed two approaches, DIANA-microT [163] and DIANA-mirPath [164]. The DIANA-mirPath software performs an enrichment analysis of multiple miRNA target genes comparing each set of miRNA targets to all known KEGG pathways [47, 123] and thus is a valuable tool for elucidating targets that are affected by deregulated miRNAs. Given the increasing amount of mRNA and miRNA data measured from the same disease or even the same individuals, more and more computer-aided tools for the integrative analysis of these data are developed and published. Among the most popular tools, developed for this purpose is ”microRNA and mRNA Integrated Analysis” (MMIA) developed by Nam and co-workers [165] that interprets miRNA and mRNA data in the context of gene ontologies and biochemical pathways.

In this chapter, we aim at an improved understanding of miRNA and mRNA relations by addressing three issues. First, as a sequel of the study by Hatzigeorgiou and coworkers, we carry out a comprehensive gene set analysis of the miRNA target sets by considering not only KEGG Pathways but also TRANSPATH networks [48], TRANSFAC [49] transcription factors, and Gene Ontology (GO) terms [31]. Second, we perform a network analysis of all target genes of all miRNAs. Third, we screen differentially expressed mRNAs for enrichment of specific miRNA targets. With the help of this analysis we exemplify once more the extensive capabilities of our comprehensive gene set analysis pipeline GeneTrail [10, 24].

4.1 miRNA target enrichment analysis

In order to detect target pathways of miRNAs, we carried out a standard Over-Representation Analysis (ORA) as described in Section 2.1.3.1. In brief, for each of the human miRNAs in the Sanger miRBase (“MicroCosm Targets Version v5“)[1] [54–56] we extracted their target genes along with their significance value. The lower this value, the higher the chance that the respective gene is actually targeted by the respective miRNA. For

[1] http://www.ebi.ac.uk/enright-srv/microcosm/htdocs/targets/v5/

the following analysis, we extracted the target genes for each human miRNA to a significance threshold of 0.0001. The resulting approximately 800 gene sets for *Homo sapiens* were separately evaluated with GeneTrail analyzing about 13000 biological pathways and categories including KEGG Pathways, TRANSPATH pathways, Gene Ontology terms and others using all human genes as reference set. The significance level for the ORA was set to 0.05 and the computed p-values were adjusted by the FDR correction method as proposed by Benjamini-Hochberg [61].

Of 13160 screened biological categories, 1766 are significant for at least a single miRNA. The highest number of hits are achieved by the categories ”Metabolic Pathways” (30), ”Cell Cycle” (23), and ”Pathways in cancer” (20) followed by a long list of disease relevant pathways including TGF-beta and MAPK signaling cascade (see also Table 4.1, categories which are significant for more than 10 miRNA target sets).

Table 4.1: Categories that are most frequently enriched with miRNA target gene sets

Category	**Number of significant miRNA target gene sets**
Metabolic pathways	30
Cell cycle	23
Pathways in cancer	22
Focal adhesion	15
TGF-beta signaling pathway	13
Fatty acid metabolism	13
catalytic activity	12
cellular ketone metabolic process	12
ECM-receptor interaction	11
Fc Epsilon RI signaling pathway	11
Organic acid metabolic process	11
Carboxylic acid metabolic process	11
MAPK signaling pathway	11
substrate-specific transporter activity	11
substrate-specific transmembrane transporter activity	11
oxoacid metabolic process	11
transporter activity	10
E2F network	10
Valine,leucine and isoleucine degradation	10
p53 signaling pathway	10
Colorectal cancer	10
Toll-like receptor signaling pathway	10

For target sets of 254 miRNAs, at least one significant category has been found. On average each miRNA has 5 significant categories. The miRNAs with the highest number of significant categories was miR-202 (90) followed by miR-101 (65). A list of miRNAs whose targets are enriched in more than 40 significant categories is provided in Table 4.2.

To improve our understanding of the putative pathways or biological categories that miRNAs may regulate or influence, we carried out a clustering approach. First, we removed

Table 4.2: miRNAs with highest number of significant categories

	Number of significant categories				
miRNA	**Gene Ontology**	**KEGG**	**TRANSFAC**	**TRANSPATH**	**total**
hsa-miR-202	89	1	0	0	90
hsa-miR-101	64	0	0	1	65
hsa-miR-613	55	6	0	0	61
hsa-miR-936	58	0	0	0	58
hsa-miR-196a	54	0	2	0	56
hsa-miR-1	53	1	1	0	55
hsa-let-7f	49	0	1	0	50
hsa-miR-302b*	48	1	0	0	49
hsa-miR-23b	47	0	1	0	48
hsa-miR-212	43	4	0	0	47
hsa-miR-23a	47	0	0	0	47
hsa-miR-196b	44	0	2	0	46
hsa-miR-29c	40	5	1	0	46
hsa-miR-191	45	1	0	0	46
hsa-miR-181c*	45	0	0	0	45
hsa-let-7a	44	0	1	0	45
hsa-miR-801	43	0	0	0	43
hsa-miR-29a	37	3	1	0	41
hsa-miR-199b-5p	39	1	0	0	40
hsa-miR-29b	36	3	1	0	40

miRNAs with less than 5 significant categories and categories that are enriched for less than 5 miRNA target sets. The clustering is based on a binary matrix that describes which categories (rows) are enriched with respect to the corresponding miRNA target sets (columns), i.e., the matrix contains a 1 at position (i, j) if the targets of miRNA j are enriched in category i and a 0 otherwise. Based on this matrix we carried out a hierarchical clustering of miRNAs and categories separately. In more detail, we applied bottom-up hierarchical clustering using the Euclidian distance for measuring the distances between pairs of column and row vectors. The result of this clustering is shown in Figure 4.1. In the lower left corner of the heatmap, a cluster containing the let-7 family can be detected. These miRNAs seem to control, among others, categories as ”transporter activity”, ”RNA interference”, ”macrolide binding” or ”drug binding”. The second cluster in the lower left corner contains miRNAs hsa-miR-525-3p, hsa-miR-524-3p, hsa-miR-506, hsa-miR-614, hsa-miR-920, hsa-miR-124, hsa-miR-376a, and hsa-miR-376b that control metabolic pathways.

We also addressed the question how specific the detected pathways or categories are and whether there are pathways or categories that are triggered by miRNAs in general. To this end, we set up three lists, containing genes that are targets of at least one miRNA at a threshold level for the probability of the predicted targets of 0.01, 0.001 and 0.0001, respectively. These lists containing 16217, 13168 and 8508 genes have been processed using GeneTrail performing an ORA. For the most unspecific miRNA target threshold of 0.01 no significant KEGG pathways have been detected. The target threshold values of

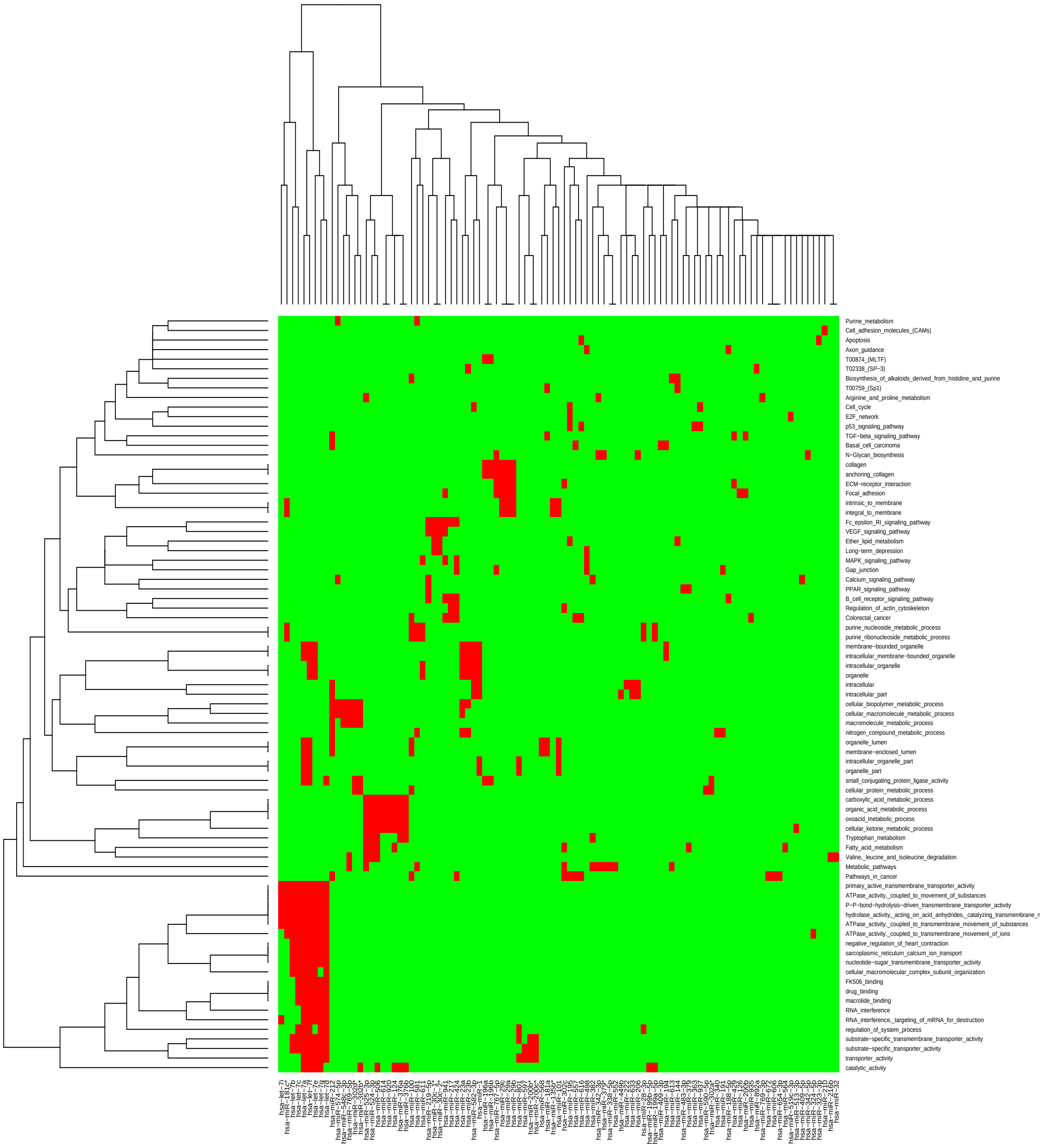


Figure 4.1: This heatmap presents significant miRNA to putative pathway or category correspondences. The heatmap has a red spot at position (i, j) if the targets of a miRNA j are significantly enriched in category i. In the bottom left corner, a cluster containing the let-7 family can be detected.

Table 4.3: KEGG pathways targeted by all miRNAs for different thresholds. The values in the cells of the table correspond to the FDR adjusted p-values computed for the pathway. – = not significant

Pathway	0.01	0.001	0.0001
ABC transporters	–	0.0362	–
Aminoacyl-tRNA biosynthesis	–	–	0.0050
Basal cell carcinoma	–	0.0154	0.0250
Complement and coagulation cascades	–	0.0447	–
ECM-receptor interaction	–	0.0447	0.0056
Epithelial cell signaling in Helicobacter pylori infection	–	0.0495	–
Focal adhesion	–	–	0.0090
Glycine, serine and threonine metabolism	–	0.0154	–
Lysosome	–	0.0362	–
MAPK signaling pathway	–	0.0018	0.0103
Metabolic pathways	–	0.0119	0.0173
p53 signaling pathway	–	–	0.0420
Pathways in cancer	–	0.0236	0.0269
Purine metabolism	–	–	0.0003
Steroid biosynthesis	–	0.0447	–
Toll-like receptor signaling pathway	–	0.0109	–
TGF-beta signaling pathway	–	–	0.0239

0.001 and 0.0001 showed increased numbers of pathways and additionally entailed a significant overlap between both sets. For 0.001 (0.0001), we detected 12 (10) putative target pathways. Of these, 5 pathways were significant for both sets including ”Basal cell carcinoma”, ”ECM-receptor interaction”, ”MAPK signaling pathway”, ”Metabolic pathways”, and ”Pathways in cancer”. A summary of all pathways and all threshold values is presented in Table 4.3.

4.2 miRNA target network analysis

For the network analysis of miRNAs we retrieved the KEGG regulatory network for *Homo sapiens* from our Biochemical Network Database (BNDB) [25] as described in Section 2.1.2.1. The resulting graph contains 1679 nodes and 2509 edges in total. Since not all predicted targets of the available 851 human miRNAs could be mapped onto the regulatory network, we removed those miRNAs where less than 10% of the targets could be mapped or the overall number of mapped targets was < 3, resulting in 695 remaining miRNAs. In the following analyses we used the threshold value of 0.001 for the miRNA targets.

For the considered miRNAs, we wanted to investigate if the average distance between pairs of targets for the different miRNAs is significantly lower in comparison to randomly selected nodes from the complete network. To this end, we computed for each pair of targets or randomly selected nodes v_i and v_j their distance. Since our considered regulatory network is directed, the distance of two nodes $dist(v_i, v_j)$ is not necessarily equal to $dist(v_j, v_i)$. Therefore, we chose for each pair (v_i, v_j) the minimum of these distances for the computation of the average distance. If there exists no path between two nodes, the distance was set to the diameter of the complete regulatory network to penalize the absence of a path. The sum of the pair distances is finally divided by the number of pairs considered. To estimate if the average distance of the m targets of a given miRNA is significant, we carried out 1000 permutation tests for each target set size m. To this end, we randomly selected m nodes from the complete network and calculated the average distance for the random node set. The distribution of the average distances of randomly selected nodes against the average distances of the miRNA targets is shown in Figure 4.2. For testing the significance, we performed an unpaired two-tailed t-test, which yielded a p-value $< 10^{-9}$ confirming that miRNA target pairs have a lower average distance than randomly selected nodes.

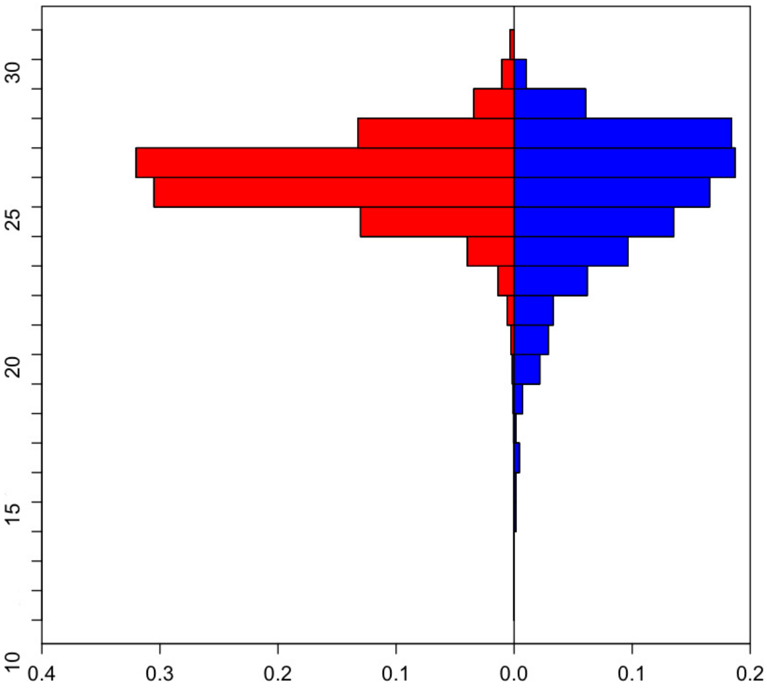


Figure 4.2: Comparison of the distributions of the average distances between randomly selected nodes on the left hand side and the miRNA targets on the right hand side. The y-axis of this back-to-back histogram presents the distance between nodes and the x-axis shows how many percent of random node pairs and of miRNA targets have this distance. The distribution of the miRNA targets is slightly shifted towards smaller distances.

Furthermore, we analyzed the coverage of all miRNA targets and the complete regulatory network considering only such nodes that are proteins (not protein families or complexes).

When regarding the union of the targets of each of the 695 miRNAs that can be mapped to proteins in the network, we reach a coverage of the regulatory network of 640 / 825 (78%). If we take the number of all human genes having an amino acid sequence as reference set (25673), we would expect to find about 414 proteins mapped on the network instead of 640, if we choose 12885 miRNA targets from the reference set coding for proteins. The hypergeometric distribution test yields a p-value of $< 10^{-60}$ for obtaining such a coverage per chance. This finding significantly points out the crucial role these miRNAs play in the regulation of biochemical processes and indicates that the regulation takes place on basis of balance and interplay of concentrations of miRNAs rather than by regulating some few important targets or hubs in the network.

4.3 Deregulated cancer mRNAs as potential miRNA targets

In this section, we analyze whether the deregulation of genes in cancer could be caused by miRNAs. More exactly, we investigated if genes that are deregulated in cancer are statistically significant enriched with targets of certain miRNAs. This hypothesis has been tested on two independent cancer entities, lung cancer and glioma and both comparisons are directly compared to each other.

Lung cancer

We extracted expression profiles of squamous lung cancer biopsy specimens and paired normal specimens from 5 different patients (GDS1312, [166]) from the Gene Expression Omnibus [167]. For this data set a standard Gene Set Enrichment Analysis (GSEA) has already revealed a manifold of deregulated pathways, including core regulatory pathways as the cell-cycle [24]. The GDS1312 data set contains 10 samples, five normal lung tissue expression profiles and 5 profiles of cancer patients. Using GeneTrailExpress [24], we computed for each gene on the microarray the fold quotient of medians in the control and diseased group. The resulting list of genes sorted by the fold quotient serves as input for GeneTrail. On the basis of this list, we carried out analyses for detecting miRNAs whose targets are significantly up- or down-regulated using standard GSEA. Here again, we considered targets with thresholds of 0.01, 0.001 and 0.0001 separately.

For the threshold value of 0.01 we detected 44 miRNAs to be significant. For 42 of these miRNAs, the targets were significantly up-regulated in tumor tissue and for two down-regulated. Most of these miRNAs can be related to cancer in the literature, e.g., the most significant miRNA of these, hsa-miR-146b is known to be down-regulated in lung cancer [168]. For the miRNA target threshold of 0.001, we detected no significant miRNAs, while for the threshold of 0.0001 we detected the three miRNAs miR-29a, miR-29b and miR-29c as significant. Notably, all these miRNAs are also known from the literature to be down-regulated in lung cancer (miRNAs miR-29a [169, 170], miR-29b [168–170], miR-29c [169, 170]). In addition, we carried out a blood screening of healthy individuals and lung cancer patients as described by Keller et al. [171] using the Geniom RT Analyzer (febit biomed gmbh, Heidelberg, Germany) and found these miRNAs at least 4 times down-regulated compared to the control. For the most down-regulated miRNA, miR-29c, the target network is presented in Figure 4.3 and the significant categories for its target genes are listed in Table 4.4.

If we now go back to our primary analysis of target pathways presented in Section 4.1, we detected for miRNAs miR-29b and miR-29c the KEGG pathway ”Small cell lung cancer” to be significantly regulated by these miRNAs. This means that we can find the predicted target pathway directly in the expression data providing evidence for the performance of the target pathway prediction.

High grade glioma

For high grade gliomas (WHO grade III and IV astrocytomas) we considered two data sets of the Gene Expression Omnibus, GDS1975 and GDS1815 that have been analyzed separately. As control we used for both expression profiles the data set GDS596 [172] containing 158 profiles from 79 physiologically normal tissues obtained from various sources. As described above, we pre-processed the data with GeneTrailExpress, first normalizing the data sets using median normalization, then computing the sorted list of fold quotients of medians in the control and diseased group, and finally submitting the resulting list as input for GeneTrail to perform a GSEA.

We extracted the data set GDS1975 containing 85 tumors for comparison against the control set GDS596. Here, we found 115 miRNAs, 74 enriched and 41 depleted. The most significant miRNAs were hsa-miR-101, hsa-miR-200b, and hsa-miR-200c. For the data set GDS1815 that contains 100 samples, we carried out the same analysis. Here, we de-

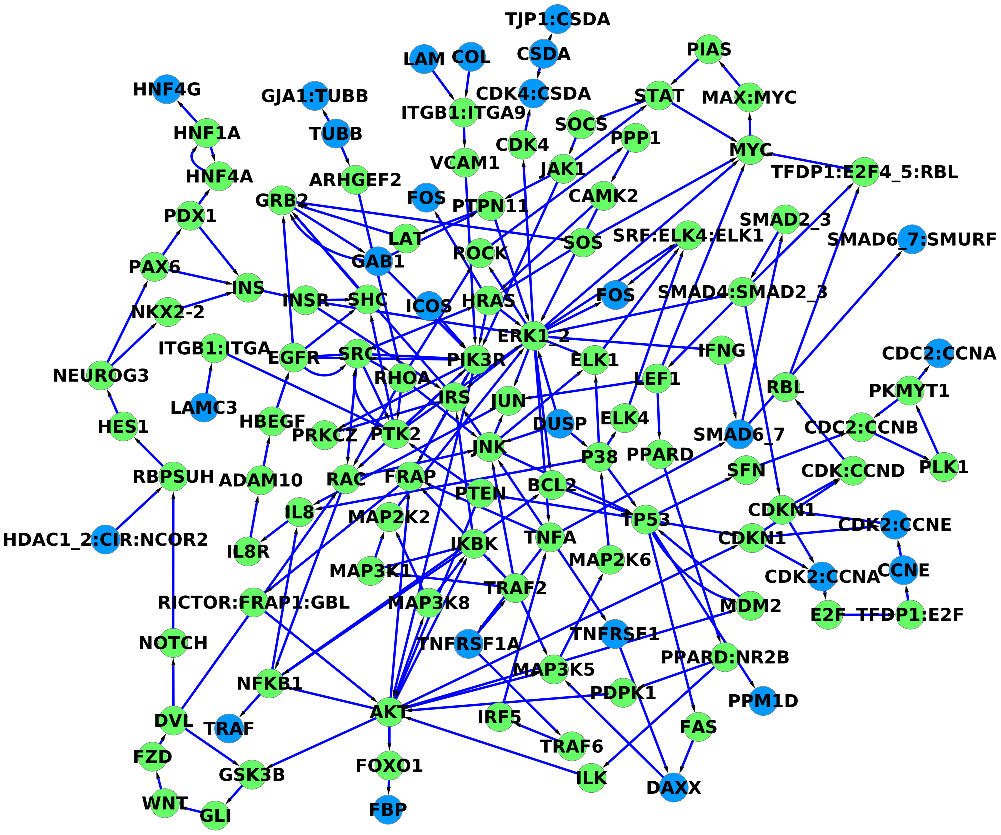


Figure 4.3: This figure presents the target network of the miRNA hsa-miR-29c. The subgraph consists of the nodes of the shortest paths between the miRNA targets. The targets of the miRNA are colored in blue.

Table 4.4: Overview of the significant categories for the target genes of miR-29c for a threshold value of 0.0001

Gene Ontology	KEGG	TRANSFAC
collagen	ECM-receptor interaction	T09836 (hsa-miR-29c)
extracellular matrix part	Focal adhesion	
proteinaceous extracellular matrix	Primary immunodeficiency	
extracellular matrix	Small cell lung cancer	
extracellular matrix structural constituent	Lysine degradation	
structural molecule activity		
anchoring collagen		
extracellular region part		
basement membrane		
collagen type IV		
sheet-forming collagen		
fibrillar collagen		
extracellular region		
extracellular matrix organization		
membrane part		
intrinsic to membrane		
membrane		
integral to membrane		
chromatin		
microfibril		
protein binding, bridging		
localization		
FACIT collagen		
collagen fibril organization		
androgen receptor binding		
cell adhesion		
biological adhesion		
fibril		
lysine N-methyltransferase activity		
protein-lysine N-methyltransferase activity		
histone-lysine N-methyltransferase activity		
extracellular structure organization		
S-adenosylmethionine-dependent methyltransferase activity		
nuclear chromatin		
nuclear hormone receptor binding		
androgen receptor signaling pathway		
steroid hormone receptor binding		
histone methyltransferase activity		
hormone receptor binding		
protein methyltransferase activity		

tected by far more significant miRNAs, 168 of which 108 are enriched, and 60 depleted. In addition, we compared the two sets of significant miRNAs. The first set contained 115 miRNAs, the second set 168 miRNAs. The overlap between both sets was 103, i.e., of the 115 miRNAs detected for the smaller set, 90% were also significant for the independent second set.

For the larger data set containing 100 samples, we also investigated the influence of the miRNA-mRNA target threshold. For the target gene thresholds of 0.01, 0.001, and 0.0001, 388 (205 up, 183 down), 168 (108 up, 60 down) and 62 (53 up, 9 down) have been identified. To reveal the similarity between the three target gene threshold sets, we produced a three-way Venn Diagram, which is shown in Figure 4.4. This diagram outlines that, e.g, the 62 significant miRNAs for threshold 0.0001 split in the following four groups: 6 are significant only for this threshold, 3 are also contained in the set for threshold 0.001, 6 are also contained in the set for threshold 0.01. However, the majority of 47 miRNAs is significant for all three thresholds.

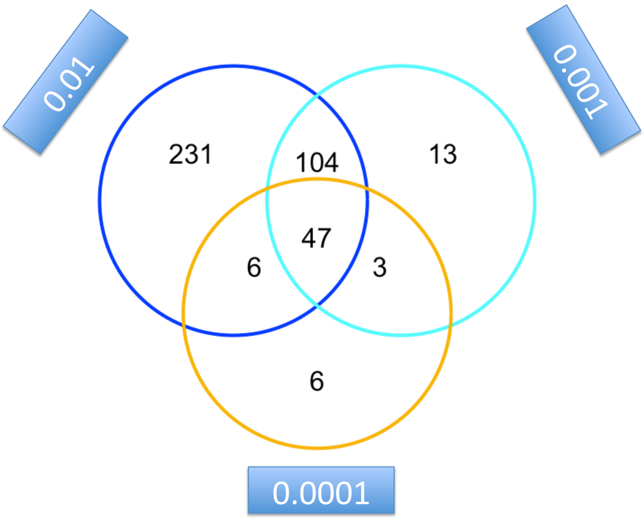


Figure 4.4: Three-way Venn diagram for the three glioma data sets computed for the miRNA target thresholds 0.01, 0.001 and 0.0001, respectively.

The three miRNAs with highest significance values included hsa-miR-1, miR-200b, and miR-144. These miRNAs are known to be deregulated in various human neoplasms [173]. Looking specifically at miRNAs known to be related to glioma tumors, we find several occurrences among the significant miRNAs in the analyzed data sets, including hsa-miR-

181a and hsa-miR-181b. However, some other popular miRNAs connected to Glioma are not detected to be significant in our study, including hsa-miR-221 and hsa-miR-222.

Overlap between lung cancer and glioma

As a final comparison, we computed for the miRNA target threshold level of 0.01 the overlap between 44 lung cancer miRNAs and 388 glioma miRNAs of the GDS1815 data set. The result of this comparison is shown as Venn diagram in Figure 4.5. In detail, 22 of the 44 lung cancer miRNAs have also been detected with glioma. On top of this list, miR-146b can be found (the complete list is provided in the supplemental material). This miRNA is known to be related to a manifold of human cancers from literature including the two cancer entities whose expression profiles are the basis of this analysis, i.e., lung cancer [168] and glioma [174]. Besides these two tumor types it has also been found to be deregulated in breast cancer [175], Leukemia [176], Pancreatic cancer [175], Prostate cancer [175] and Thyroid neoplasms [177]. These results provide further evidence for the common deregulation of some miRNAs in cancer, which can very accurately be re-detected in cancer expression profiles.

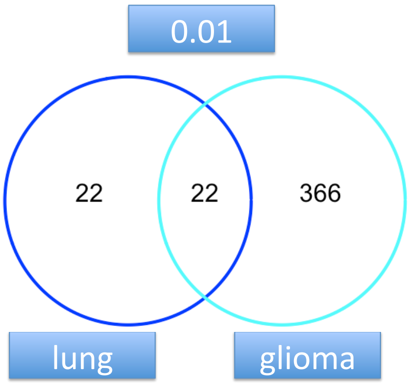


Figure 4.5: Venn diagram for the glioma and lung cancer data sets computed for the miRNA target threshold of 0.01

4.4 Conclusion

Our computational analysis deepens the understanding of miRNAs and their putative targets in biochemical networks. We provide a comprehensive ”dictionary” of miRNAs to possible target pathways that may be regulated by this miRNAs. This dictionary enables researchers to look up the target pathways of differentially regulated miRNAs that can be used, e.g., for functional studies. As an additional key result, the study also provides further evidence that miRNAs are key-players in the regulation of oncogenetic processes. Thus, our results demonstrate that an integrative screening of miRNAs and mRNAs can contribute to an improved understanding of human diseases, finally providing new starting points for disease diagnosis, prognosis and monitoring.

CHAPTER 5

IDENTIFYING DEREGULATED REGULATORY SUBGRAPHS

In the last decade, microarray-based gene expression profiles have become a central data resource to study deregulated molecular processes of diseases. Initially, microarray studies focused on single differentially expressed genes. Later, Gene Set Analysis (GSA) and related approaches were taking into account that genes do not act individually but in a coordinated fashion [4, 6, 9]. The disadvantage of this type of methods is that they can only reveal the enrichment of genes in predefined gene sets, e.g., canonical biological pathways. In recent years, the research focus has shifted towards analysis methods that integrate topological data mirroring the biological dependencies and interactions between the involved genes or proteins. In general, these graph-based approaches use scoring functions that assign scores or weights to the nodes and/or edges and make strong efforts to identify high-scoring pathways or subgraphs.

A seminal work in this area is the paper of Ideker et al. who proposed a method for the detection of active subgraphs by devising a scoring function and a heuristic approach for detecting these subgraphs [14]. Other groups reported similar methods, which are all based on scoring networks given experimental data [16, 17, 22]. In 2008, Ulitsky and coworkers presented an algorithm for detecting disease-specific deregulated pathways by using clinical expression profiles [20]. However, the abovementioned approaches focused on protein-protein interaction (PPI) networks (undirected graphs) and used heuristics to find the subgraphs. Dittrich et al. devised the first approach to solve the maximal-scoring subgraph problem optimally by Integer Linear Programming (ILP) in the context of undirected PPI networks [21].

In this chapter, we present two novel approaches for detecting deregulated components in regulatory networks using expression profiles. The first approach, called FiDePa (Finding Deregulated Paths), is a dynamic programming algorithm that identifies deregulated paths of a certain length [26] relying on standard Gene Set Enrichment Analysis (GSEA) [4,5,57]. Our second approach is an ILP algorithm which reveals the most deregulated connected subnetwork of a certain size. FiDePa has been developed in collaboration with Andreas Keller, who developed and implemented the dynamic programming algorithm. The ILP approach emerged from the collaboration with Alexander Rurainski, who developed and implemented the ILP.

5.1 FiDePa

In this section, we present a novel algorithm for detecting differentially regulated paths in a regulatory network that is based on the unweighted GSEA as described in Chapter 2.1.3.2. The input of FiDePa is a list of genes that are sorted with respect to their expression differences between two investigated states, e.g. cancer and normal tissue, and a regulatory network. As data source for the network information, we imported the KEGG [47] and the TRANSPATH [48] database into the Biochemical Network Database (BNDB) [25]. We extracted the complete human regulatory network from the BNDB and projected the ranks of the genes in the list onto the corresponding nodes. The algorithm does not consider nodes that are not contained in the sorted list.

For the computation of the deregulated paths, the algorithm interprets each path p of a certain length l in the given network as a biological category C_p that consists of the l genes represented by the nodes of the path p. Using a Kolmogorov-Smirnov-like test that computes whether the set of genes C_p belonging to the path p are equally distributed in the expression list or accumulate on the top or bottom of the list, we determine if the given path p is deregulated (contains a large number of up- or downregulated genes) or not. The applied Kolmogorov-Smirnov-like test is a standard test of GSEA [5] that computes the running sum of all genes in the sorted list. Hereby, the sorted list consisting of n genes is processed from top to bottom. Whenever a gene belonging to C_p is detected, the running sum is increased by $n-l$, otherwise it is decreased by l. The value of interest is the running sum's maximal deviation from zero, for which a p-value can be computed [57]. Since the number of paths is growing exponentially with the length l, the brute-force approach that enumerates all paths of length l and computes the running sum for each path separately is applicable only for very small values of l.

To identify the most significant paths efficiently, FiDePa computes the paths of length l with the smallest p-value. In order to facilitate the interpretation of the findings, the resulting paths have to be visualized in a well-arranged manner. To this end, we added respective functionality to our Biological Network Analyzer (BiNA) [44] that enables the user to visualize and compare significant paths. An overview of the workflow of the whole analysis procedure is summarized in Figure 5.1.

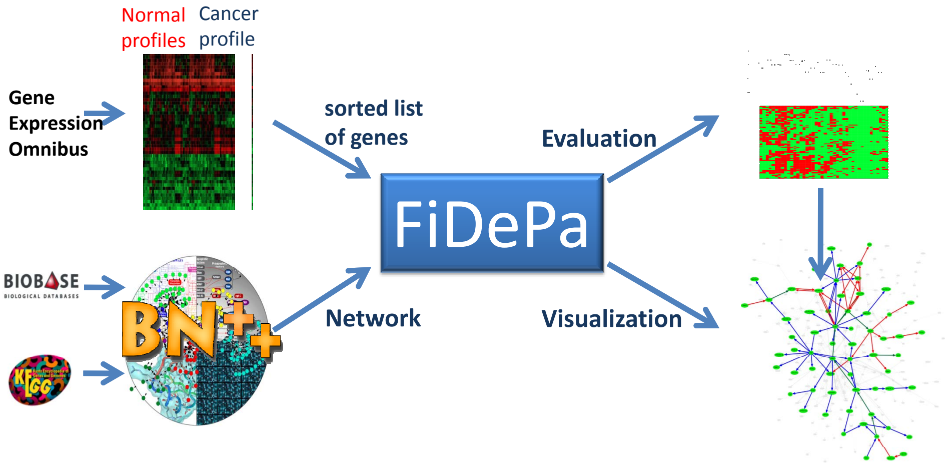


Figure 5.1: Workflow of the FiDePa algorithm

In the following, we describe the dynamic programming algorithm in more detail as presented in our Bioinformatics publication [26]. Afterwards, we present the results of the application of FiDePa to expression profiles of 100 glioma patients (WHO grades III and IV, extracted from the Gene Expression Omnibus (GEO) [167] dataset GDS1815 [178]) against a control group of 158 expression profiles (GDS596 [179]) of physiologically unaffected tissues.

5.1.1 The dynamic programming algorithm

Biological networks are often represented as directed graphs $G = (V, E)$, where the vertices (nodes) $V = \{v_1, ..., v_q\}$ represent genes, proteins or other compounds and the directed edges $e(v_i, v_j) \in E$ represent interactions or reactions between the respective compounds. A path of length l in G is a sequence $v_{p_1}, .., v_{p_i}, v_{p_{i+1}}, .., v_{p_l}$ of l nodes, where each pair v_{p_i}, $v_{p_{i+1}}$ of consecutive nodes is connected by a directed edge, which starts in v_{p_i} and ends in $v_{p_{i+1}}$. We denote the set of all paths of length l by P_l and the subset of paths in P_l that end in the node v_k by $P_l(v_k)$. The set $N(v_k)$ of predecessors of node v_k is defined as

$$N(v_k) = \{v_s \in G | \exists e(v_s, v_k) \in E\}$$

Besides the graph G, the input of the algorithm consists of a gene list S of length n. The genes in the list S are sorted with respect to an arbitrary criterion, e.g. their fold changes of expression values between two investigated states. Given a gene represented by a node v in the graph G, we denote the rank of the gene in the sorted list S as $r(v)$.

To compute significance values of a path p of length l, we carried out an unweighted GSEA as described in 2.1.3.2. In brief, the sorted list is processed from top to bottom to compute a running sum statistic RS. Whenever a gene belonging to C_p is detected, the running sum is increased by $n - l$, otherwise it is decreased by l. The value of interest is the running sum's maximal deviation from zero, denoted as RS_p. The significance value of the score RS_p can be calculated by our dynamic programming algorithm introduced in Chapter 2.1.3.2.

We define the number of nodes on the path p that have a rank smaller equal i in S as:

$$b_p[i] = |\{v \in C_p | r(v) \leq i\}|. \tag{5.1}$$

Our algorithm relies on the fact that the running sum value at position i can be computed as:

$$RS_p[i] = b_p[i] \cdot (n - l) - l \cdot (i - b_p[i]) \tag{5.2}$$

In order to compute the most significant paths $p \in P_l$ of length l, where l ranges from 1 to a user-defined upper bound m, we will first focus on the subset $P_l(v_k)$ of paths that end in a certain node v_k and have a fixed length of l. Hereby, we will derive a recurrence scheme for filling the 3-dimensional matrix $M[l, k, i]$ of size $m \cdot |V| \cdot n$ that allows to solve the problem for all nodes and the considered range of path lengths in an efficient manner.

Equation 5.2 implies that the best score of any path p of length l ending in v_k can be computed as:

$$M^*[l, k] = \max_{i=1,\dots,n} (M[l, k, i] \cdot (n - l) - l \cdot (i - M[l, k, i]), \tag{5.3}$$

where

$$M[l,k,i] = \max_{p \in P_l(v_k)} b_p[i]. \tag{5.4}$$

If M has been filled, we can easily calculate the best running sum score for any length l and any node v_k and the corresponding paths can be determined by a simple standard backtracking procedure. If no path of length l ending in node v_k exists, we set $M[l,k,i] = -1$ for all indices $i = 1, ..., n$.

Since the path of length 1 ending in a node v_k consists only of the node v_k itself, the computation of the first matrix layer $M[1,k,i]$ is straightforward:

$$M[1,k,i] = \begin{cases} 1 & : \quad r(v_k) \le i \\ 0 & : \quad r(v_k) > i \end{cases}$$

In the following, we derive the recurrence formula that allows for computing all values $M[l,*,*]$ of layer l from the values $M[l-1,*,*]$ of layer $l-1$. The idea behind the approach is similar to the principle used in shortest/longest path calculations. In order to compute the best path of length l leading to v_k, we determine the optimal paths of length $l-1$ ending in one of the predecessor nodes $v_s \in N(v_k)$ and add the path of length 1 consisting of the node v_k:

$$M(l,k,i) = \begin{cases} \max_{v_s \in N(v_k)} M[l-1,s,i] + 1 & : \quad r(v_k) \le i \\ \max_{v_s \in N(v_k)} M[l-1,s,i] & : \quad r(v_k) > i \end{cases} \tag{5.5}$$

The pseudocode in Algorithm 5.1 is applied to fill the remaining layers $2, ..., m$.

Algorithm 5.1 The computation of the dynamic programming matrix in the FiDePa algorithm

```
DYNAMIC PROGRAMMING:
for $l \in 2..m$ do  // for all layers (path lengths)
  for $i \in 1..n$ do  // for all genes in the sorted list
    for $k \in 1..|V|$ do  // for all nodes
      if $(N(v_k) == \emptyset || \max_{v_s \in N(v_k)} M[l-1,s,i] == -1)$ then
        $M[l,k,i] = -1$
      else
        if $(r(v_k) \le i)$ then
          $M[l,k,i] = \max_{v_s \in N(v_k)} M[l-1,s,i] + 1$
        else
          $M[l,k,i] = \max_{v_s \in N(v_k)} M[l-1,s,i]$
```

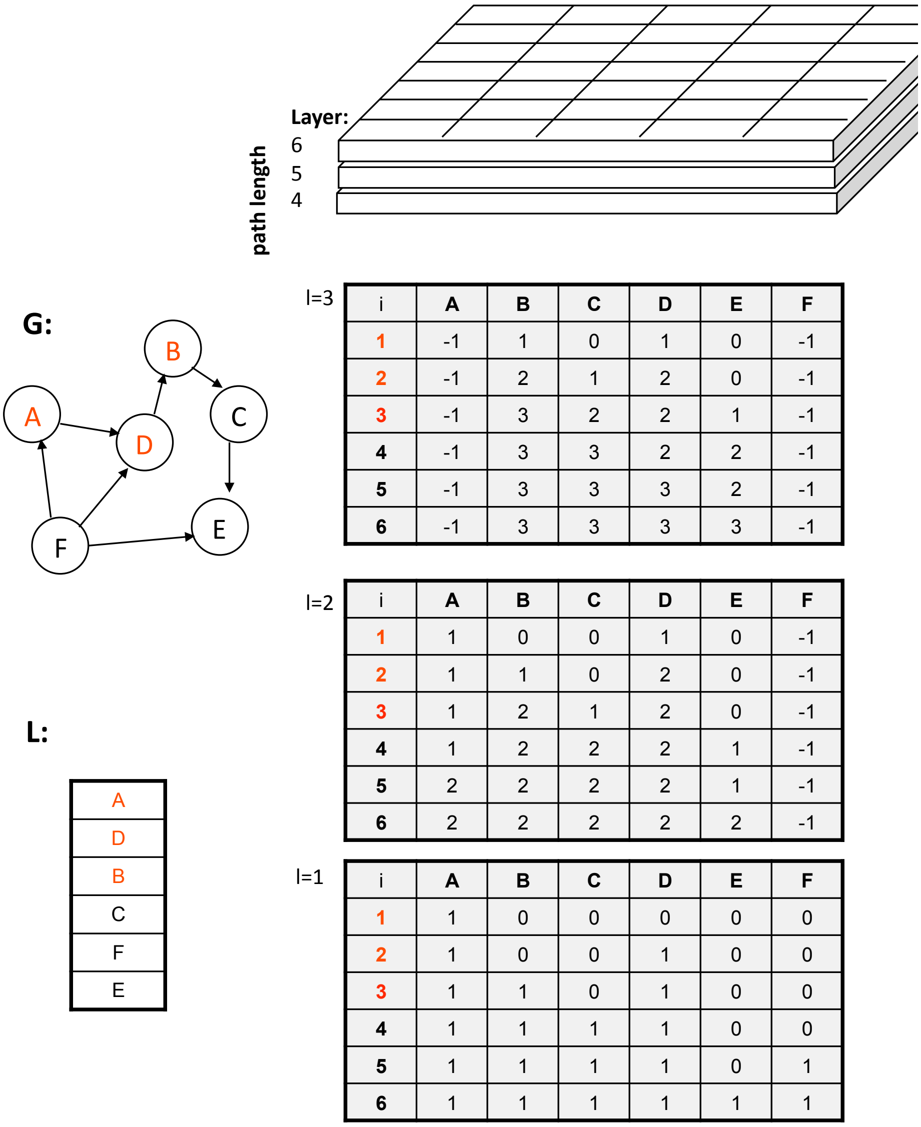
Here, the first if statement evaluates whether any path of length l ends up in v_k. If this condition does not hold, $M[l, k, i]$ is set to -1. Otherwise, the value $M[l, k, i]$ is calculated via the previously described recurrence Equation 5.5. Since we had to avoid cycles, we added a further condition which is not listed in the pseudocode described above: our algorithm searches for the best path ending in one of the predecessor nodes $v_s \in N(v_k)$ that does not contain node v_k. An example of the dynamic programming approach is provided in Figure 5.2 for a small network with 6 nodes. In this example, the layers 1-3 of M are computed.

5.1.2 Deregulated glioma paths

We applied our dynamic programming algorithm to study deregulated signaling cascades in glioma tumors. To this end, we analyzed 100 glioma expression profiles of WHO grades III and IV [178]. As background distribution, we used 158 expression profiles (GDS596) [179] of physiologically unaffected tissues. Control and cancer expression profiles were downloaded from GEO and all profiles were quantile normalized. Then, for each transcript t, the mean value μ_t and the standard deviation σ_t of the transcript in the control profiles were computed. For a given cancer profile, we computed the z-score z_t for transcript t with expression value x_t as follows:

$$z_t = \frac{x_t - \mu_t}{\sigma_t}$$

The corresponding genes were sorted in decreasing order with respect to the absolute value of their z-scores, resulting in one sorted gene list for each cancer profile. The input of the FiDePa algorithm consisted of the sorted z-score lists and the union of the KEGG and TRANSPATH networks that was imported from our BNDB database [25]. For each cancer profile, we carried out the following computation steps: the z-scores of the genes present in the network were assigned to the corresponding nodes, the ranks of the nodes were calculated, the dynamic programming algorithm was carried out and the resulting paths plus their p-values were computed. Hereby, the considered path lengths ranged from 2 to 8 edges. Afterwards, the union graph unifying all detected paths was constructed and stored. Finally, we analyzed the obtained results by carrying out comprehensive statistical tests that will be described below.



A
D
B
C
F
E

l=3

i	A	B	C	D	E	F
1	-1	1	0	1	0	-1
2	-1	2	1	2	0	-1
3	-1	3	2	2	1	-1
4	-1	3	3	2	2	-1
5	-1	3	3	3	2	-1
6	-1	3	3	3	3	-1

l=2

i	A	B	C	D	E	F
1	1	0	0	1	0	-1
2	1	1	0	2	0	-1
3	1	2	1	2	0	-1
4	1	2	2	2	1	-1
5	2	2	2	2	1	-1
6	2	2	2	2	2	-1

l=1

i	A	B	C	D	E	F
1	1	0	0	0	0	0
2	1	0	0	1	0	0
3	1	1	0	1	0	0
4	1	1	1	1	0	0
5	1	1	1	1	0	1
6	1	1	1	1	1	1

Figure 5.2: Dynamic Programming Algorithm of FiDePa. Example of the dynamic programming algorithm for a network G with $|V| = 6$ nodes, the path length $m = 6$, and a sorted list S of $n = 6$ genes. The initialization and computation for the first three layers are presented. The last column in the second layer reveals that no pathways of length 2 ends in node F. The entry $(4, E) = 1$ in the second layer means that a pathway of length 2 exists that ends in node E and has one gene in the sorted list with position in the sorted list smaller or equal than 4. Source: PhD thesis Andreas Keller [137]

5.1.2.1 Enrichment analysis of genes of the union graph

First, we studied the glioma union graph, consisting of the union of edges that proved to be significant in at least one of the analyzed glioma expression profiles. The union graph consisted of a total of 192 nodes and 549 edges. The genes that occurred most frequently in the deregulated subnetworks are ATF4 (45), ELK1 (43), DDIT3 (39), MAP2K2 (38), MAPKAPK5 (37), ATF2 (36), MOS (36), TP53 (36), JUN (34), MAP2K7 (34), CDC25B (32), MAP2K3 (32), MAP3K10 (32), MYC (32), ELK4 (31), MAP2K1 (31) and MAPT (31). Here, the numbers in brackets denote the number of cancer profiles where FiDePa detected paths containing the respective genes. A literature inquiry revealed that all these genes are closely connected to cancer development or progression, most of them are also directly connected to glioma. As the gene list indicates, many of the above genes belong to the MAPK (mitogen-activated protein kinase) signaling pathway or to the Apoptosis pathway. To detect the significantly enriched biochemical pathways, we used GeneTrail [10] (see also Chapter 2.1).

We carried out an over-representation analysis, comparing the genes of the union network to all human genes using GeneTrail's standard parameters. Our analysis revealed a total of 26 significantly enriched KEGG pathways (Table 5.1). On top of the results list appeared the MAPK signaling pathway, with an expected number of 18 genes and an observed number of 69 genes. The pathway with the second best significance value was the Natural killer cell-mediated cytotoxicity, with an expected number of 9 genes and an observed number of 34 genes, followed by the Apoptosis with 25 observed genes and 5 expected genes. The list, of course, entailed several cancer pathways, including colorectal cancer, pancreatic cancer and glioma.

To compare the results of our FiDePa algorithm with an analysis, which does not consider the network topology, we carried out a standard GSEA by using GeneTrail. The KEGG pathway analysis identified 16 enriched pathways, including several pathways that were also identified by FiDePa, e.g. the Natural killer cell-mediated cytotoxicity or the T-cell receptor signaling pathway. However, some clearly cancer-related pathways including Apoptosis, Glioma cancer, Pancreatic cancer, MAPK signaling pathway and others were only identified by the FiDePa analysis, while they were missed using the standard GSEA.

Table 5.1: Significant KEGG Pathways of the union network

pathway	# expected genes	# observed genes	sig. value
MAPK signaling pathway	18	69	1.6^{-25}
Natural killer cell mediated cytotoxicity	9	34	1.39^{-11}
Apoptosis	5	25	9.29^{-11}
Epithelial cell sig. in H. pylori infection	3	17	5.01^{-10}
Focal adhesion	9	26	2.52^{-06}
Adherens junction	5	18	1.38^{-05}
T cell receptor signaling pathway	5	18	1.46^{-05}
Chronic myeloid leukemia	4	16	1.65^{-05}
Fc epsilon RI signaling pathway	4	16	1.85^{-05}
Toll-like receptor signaling pathway	5	17	5.53^{-05}
Colorectal cancer	2	10	9.40^{-05}
Pancreatic cancer	4	14	0.0001
Cytokine-cytokine receptor interaction	14	30	0.0001
Adipocytokine signaling pathway	4	14	0.0001
mTOR signaling pathway	2	10	0.0002
GnRH signaling pathway	6	17	0.0003
B cell receptor signaling pathway	4	11	0.0013
Insulin signaling pathway	9	20	0.0013
Cell cycle	5	13	0.0014
Glioma	4	11	0.0030
Type II diabetes mellitus	3	9	0.0033
VEGF signaling pathway	5	12	0.0059
Type I diabetes mellitus	0	3	0.0059
Leukocyte transendothelial migration	5	11	0.0135
Axon guidance	8	15	0.0173
Maturity onset diabetes of the young	2	5	0.0386

5.1.2.2 Single patient analysis

To demonstrate the applicability of the FiDePa algorithm, we compared two WHO grade III gliomas, both without microvascular proliferation and necrosis. Both patients were of similar age with 39 and 35 years. However, the respective survival time were quite different with 41 weeks and 477 weeks. The comparison of the deregulated networks of both tumors revealed a very small overlap, as shown in Figure 5.3.

5.1.3 Discussion and Conclusion

In this section, we presented a dynamic programming algorithm that aims at detecting the significantly deregulated signaling cascades in tumor cells. The FiDePa algorithm interprets expression differences between tumor and normal tissue and relies on GSEA. Since FiDePa enables the comparison of a single tumor expression profile with the control group, it provides information on regulatory features that are specific for the corresponding tumor

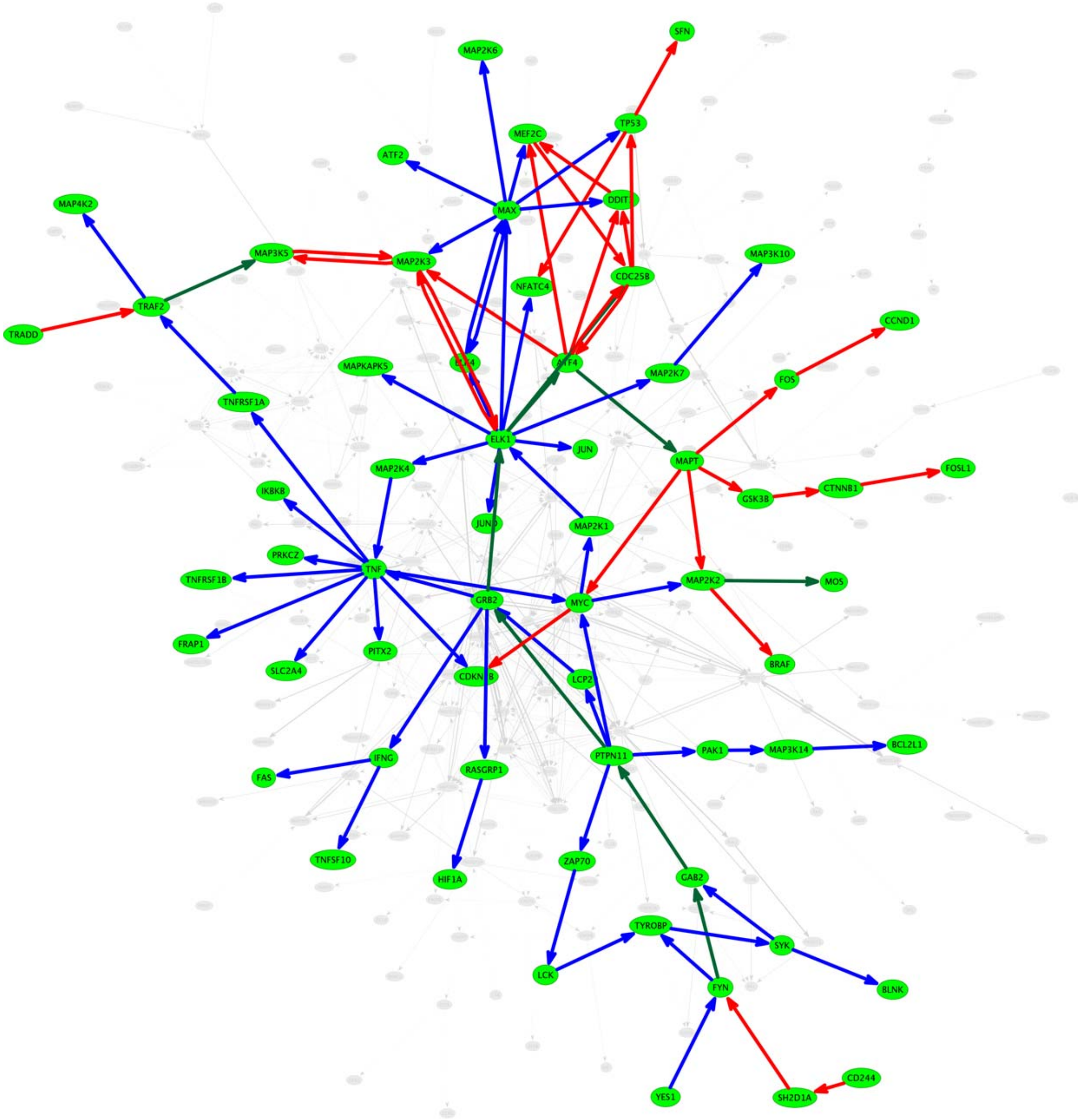


Figure 5.3: BiNA visualization of the two patients A (blue, survival time 477 weeks) and B (red, survival time 41 weeks). Edges on significant paths in both tumors are colored green. The network shows the relevant part of the complete consensus network, whereas the non-relevant part of the consensus network is presented by gray edges and nodes in the background.

and that can contribute to a personalized medicine by tailoring the tumor therapy to the specific regulatory tumor features identified by FiDePa.

The application of FiDePa to a glioma dataset showed that the algorithm is able to detect the relevant signaling cascades that are known to be glioma and/or cancer related. The most significant pathway was the MAPK signaling cascade, followed by the Natural killer cell-mediated cytotoxicity and the Apoptosis. It is well known that both pathways are deregulated in glioma: the MAPK signaling cascade, e.g. is described to be deregulated in glioma in various studies [180–183]. An upregulation of the MAPK signaling cascade in cultured glioma cells mediated by fibroblast growth factors indicated that MAPK pathway participates in the FGF-dependent glioma development [181]. As for the MAPK signaling pathway, we were able to retrieve all other significant pathways in the glioma literature, providing evidence for the effectiveness of the FiDePa algorithm. However, besides the results coherent with the findings in the literature some additional glioma-related pathways are cited in the literature. For example, Phillips et al. [178] suggest that Akt and Notch signaling are hallmarks of a poor prognosis of gliomas, while these pathways were not ostentatious in our work. This, however, might be explained by the fact that our FiDePa analysis did not focus on prognosis, but on the comparison of glioma and normal tissue. In 2008, two large-scale integrated studies on glioblastoma have been published by McLendon et al. [184] and Parsons et al. [185]. McLendon and co-workers identified ERBB2, NF1 and TP53 as key players in glioblastoma together with the RTK signaling, the p53 and RB tumor suppressor pathways. In our expression pattern-based study, we identified the TP53 component, while the other pathways play a less important role. In contrast to our results, the MAPK signaling cascade is non-significant in the study of McLendon et al., which is based on genetic alterations including validated somatic nucleotide substitutions, homozygous deletions and focal amplifications. Parson and co-workers identified the gene IDH1 as CAN-gene (candidate cancer gene) by integrating of sequencing, copy number and expression data. However, this gene does not show significant deregulation in our data and thus is not included in our union network. Other CAN genes identified by Parson et al. were included in our network, e.g. TP53, RB1 or EGFR.

In this work, the FiDePa algorithm has been applied for studying regulatory networks, which play an essential role for cancer development and progression. However, our algorithm can, of course, be applied to arbitrary networks, including protein-protein interaction networks. Here, an additional preprocessing step is necessary for matching the proteins in the network to the genes in the sorted list.

In the light of the ongoing discussion on the quality and effectiveness of gene set analysis methods [8, 186, 187], we would like to underline that our dynamic programming approach can be easily adapted to other gene set analysis method or gene scoring approaches, e.g. Wilcoxon rank-sum test, median, mean, SAM-GS, and some other approaches discussed in Ackermann and Strimmer [8]. Actually, the dynamic programming algorithm can be simplified for most of the other gene set analysis methods. However, the direct p-value computation usually has to be replaced by more laborious permutation tests.

5.2 ILP based approach

In contrast to the FiDePa algorithm, our branch-and-cut based ILP approach computes subgraphs or subnetworks instead of simple paths. The input of our algorithm consists of a regulatory network and a list of genes that are scored according to their deregulation. In this work, the underlying regulatory network was taken from the KEGG database [47, 123]. Since KEGG pathways also contain nodes for protein families, we transformed the original KEGG pathways by splitting the nodes of protein families into their components as described in Section 2.1.4.5.

The second necessary input for our algorithm is a list of scored genes. These scores can be derived, e.g., from expression experiments. In brief, if we want to compare the differences in expression of two conditions, we compute for each transcript on the microarray a score that mirrors the difference between the considered states. In general, we can use any measure that is also applied for finding differentially expressed genes as, e.g., the fold change. In an additional step, the transcript IDs are converted to gene identifiers. The resulting list contains for each gene on the microarray a score that mirrors the deregulation of the gene under the considered conditions, i.e., the higher the expression difference between the two considered states, the larger the score of a gene.

Before the computation, the genes of the list have to be mapped to the network nodes. Since not all nodes or gene identifier of the network are also available on the microarray, we cannot assign a calculated score to every node of the regulatory network. Missing scores are assumed to be zero. In our tests, about an eighth of all nodes had a zero score.

Given this input, our ILP-based algorithm computes the heaviest connected subnetwork of size k, i.e. the most deregulated subnetwork with the highest sum of node scores. Here, we define a subgraph G as connected if it contains at least one root node v_r from which all other nodes in G are reachable, i.e., for each node v in G, a path from v_r to v consisting only of nodes in G exists. The vision implicated by the proposed connectivity model is to identify – besides the most deregulated components – the root node that may represent a key player in the pathogenic process. This key player may be responsible for the observed differences between the investigated conditions and may serve as a potential target for therapy purposes.

The results of the computation can be visualized in the Biological Network Analyzer (BiNA) [44], which is a Java application suited for the visualization of metabolic and regulatory networks. An overview of the different steps of our approach is presented in Figure 5.4.

In the following, we describe the formulation of the ILP in more detail. Afterwards, we present the results of the application of our algorithm to gene expression profiles of nonmalignant mammary epithelial cells from BRCA1 mutation carriers and non BRCA1 mutation carriers [188]. We explore the effect of the mutations on the regulatory processes to gain new insights how these mutations may contribute to the development of breast cancer.

5.2.1 Integer linear program

The problem of finding a connected subgraph of size k which maximizes the sum of the scores is formulated as an Integer Linear Program (ILP) and then solved by a branch-and-cut approach. Here, we define a subgraph G as connected if it contains at least one root node v_r from which all other nodes in G are reachable, i.e., for each node v in G, a path from v_r to v consisting only of nodes in G exists. We assign a score (absolute value of the corresponding real data if available) to every node in the network. Since not all nodes or gene identifier of the network are also available on the microarray chip, we cannot assign a calculated score to every node of the regulatory network. Missing scores are set to zero.

Our ILP formulation uses two variables for each node i: x_i and y_i. The variable $x_i \in \{0,1\}$ determines whether its corresponding node is contained in the subgraph ($x_i = 1$) or not ($x_i = 0$). The variable $y_i \in \{0,1\}$ indicates that its corresponding node i is the root node ($y_i = 1$) or not ($y_i = 0$). Let s_i be the score of node i then the optimization problem can be formulated as

$$\max_{\mathbf{x}} \sum_i s_i x_i.$$

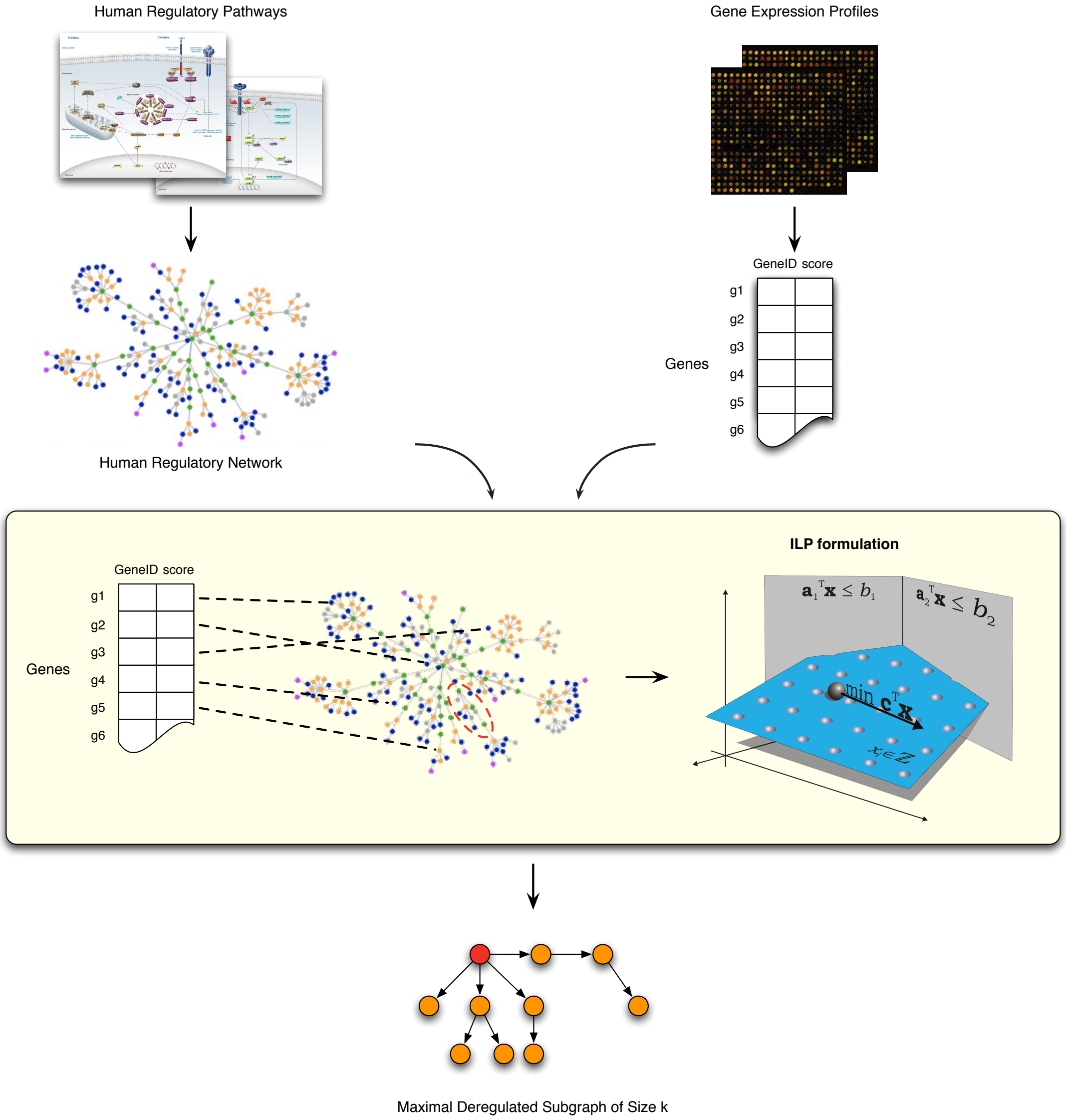

Human Regulatory Pathways
Gene Expression Profiles
GeneID score
Genes
g1
g2
g3
g4
g5
g6
Human Regulatory Network
ILP formulation
$\mathbf{a}_1^T\mathbf{x} \leq b_1$
$\mathbf{a}_2^T\mathbf{x} \leq b_2$
$\min \mathbf{c}^T\mathbf{x}$
$x_i \in \mathbb{Z}$
Maximal Deregulated Subgraph of Size k

Figure 5.4: Workflow of our ILP-based algorithm for the computation of deregulated subgraphs. Our algorithm requires as input a biological network and a list of genes with scores that have been derived from expression data and express the degree of deregulation. After the scores of the genes have been mapped to the corresponding nodes of the network, our ILP-based branch-and-cut approach calculates the most deregulated subgraph.

The constraint that the subgraph has a predefined size of k nodes, is given by

$$\sum_i x_i = k.$$

We ensure that we obtain one root node by

$$\sum_i y_i = 1.$$

The inequalities

$$y_i \leq x_i \quad \text{for all } i$$

guarantee that a designated root node is also chosen. All remaining constraints concern the connectivity of the desired subgraph. Let $\mathsf{In}(i)$ be the set of indices of the predecessors of node i, i.e. there exists an in-edge into node i, then we ensure that a chosen node has either a predecessor or it is the designated root node by

$$x_i - y_i - \sum_{j \in \mathsf{In}(i)} x_j \leq 0 \quad \text{for all } i.$$

Unfortunately, this kind of constraints is also fulfilled in cycles since every node in a cycle has a predecessor. Hence, a subgraph generated by the above constraints alone may have disconnected cycles. Let $\mathcal{C}$ be the indices of a cycle and analogously $\mathsf{In}(\mathcal{C})$ the set of indices of nodes which share an in-edge into this cycle, then the extension of the above constraint to the cycle $\mathcal{C}$ is given by

$$x_i - \sum_{j \in \mathcal{C}} y_j - \sum_{j \in \mathsf{In}(\mathcal{C})} x_j \leq 0 \quad \text{for all } i \in \mathcal{C}. \tag{5.6}$$

In theory, the complete description of our optimization problem as given above requires a constraint for every cycle, resulting in millions of inequalities of type (5.6) for the considered problem instances. In practice, branch-and-cut algorithms start with a basic set of constraints, solve the relaxed underlying LP problem, and check if the result violates constraints. If so, the violated constraints are added and the solver is restarted. As our set of basic cycle constraints we only consider cycles with two or three nodes. In order to identify violated constraints, we implemented an efficient algorithm that searches in given LP solutions for cycles which do not satisfy inequalities of type (5.6). These inequalities will be added to the constraint set. This procedure is iterated until either we obtain an optimal subgraph, i.e, an integer solution without violated constraints, or we have a non-integral

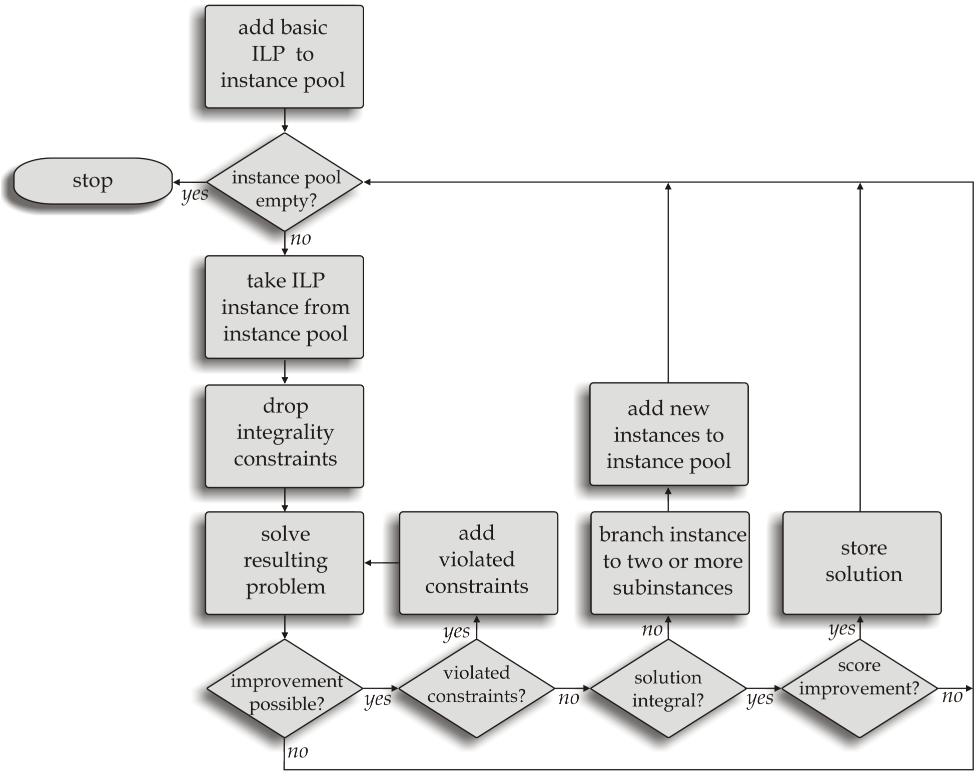


Figure 5.5: Branch-and-Cut workflow for solving the ILP.

solution, but we cannot identify further violated constraints. In the latter case we perform a branching step. In this study, we used the branch-and-cut framework of CPLEX[1], version 11.110, with the “traditional mixed integer search method”. This commercial library provides the possibility to branch using automatically detected favorable strategies. We used CPLEX’s default settings. A general workflow of such branch-and-cut algorithms is presented in Figure 5.5. For a detailed survey of branch-and-cut algorithms the interested reader is referred to Nemhauser [189] and Schrijver [190].

Our reference implementation is a single thread application, i.e. we could further speed up the solution process by parallelization techniques. However, all calculations finished within a few minutes on an Intel Xeon CPU, 2.5GHz. Thus, we did not incorporate advanced programming methods.

[1] `http://www-01.ibm.com/software/integration/optimization/cplex/`

5.2.2 Nonmalignant primary mammary epithelial cells

For the evaluation of our method, we downloaded and analyzed the GSE13671 data set from GEO. The GSE13671 set contains expression data from nonmalignant primary mammary epithelial cells with and without BRCA1 mutations and was published in a study of Burga et al. [188]. We computed the fold difference for the mean of the BRCA1 mutation carriers against the mean of non-mutation carriers given the normalized and log transformed expression values. The Affymetrix chip IDs were mapped to NCBI Gene IDs and the resulting list containing genes and corresponding expression values served as input for our algorithm. To explore the stability of the core components in this case, we computed the most deregulated subgraphs for different subgraph sizes ranging from 10 to 25 nodes. We denote the union of all nodes and edges that appear in at least one of the 16 calculated optimal subgraphs as the union graph. The less nodes this union graph consists of, the more stable are the core components of the subgraphs in the total regulatory network.

Figure 5.6 shows the best subgraph for 25 nodes and, additionally, the remaining nodes of the union graph as isolated vertices. The corresponding genes along with their number of occurrence in the different 16 subgraphs are also listed in Table 5.2. Figure 5.6 reveals that the complete union graph is very compact (only 34 vertices for the most deregulated subgraphs consisting of 10-25 nodes), which means that the most deregulated part of the network seems to be stable. The core components occurring in all of these subgraphs are the path EGLN3 (PHD3) $\rightarrow$ EPAS1 (HIF-2α) $\rightarrow$ VEGF $\rightarrow$ KDR (VEGFR2) with the designated root node EGLN3 and, more downstream located, the subgraph rooted in MAPK13 consisting of the genes TP53, DDIT3, RRM2, and GADD45B. It is interesting to note that the root node is very stable, i.e., independent of the size of the subgraph, EGLN3 is always the designated root node.

For testing the significance of the computed subgraph of size 25 and root node EGLN3, we carried out 1000 permutation tests, where we permuted the scores of the network nodes and computed the best subgraph of the same size with this root. The p-value was calculated as the number of permutations reaching an equal or better score than our original subgraph rooted in EGLN3 divided by the number of permutations. No other subgraph of this size with this root node reached a better score in 1000 permutation tests (p-value $<$ 0.001).

When performing an ORA for the genes of the subgraph of size 25 as test set and the genes of the regulatory network as reference set, we find many pathways significantly en-

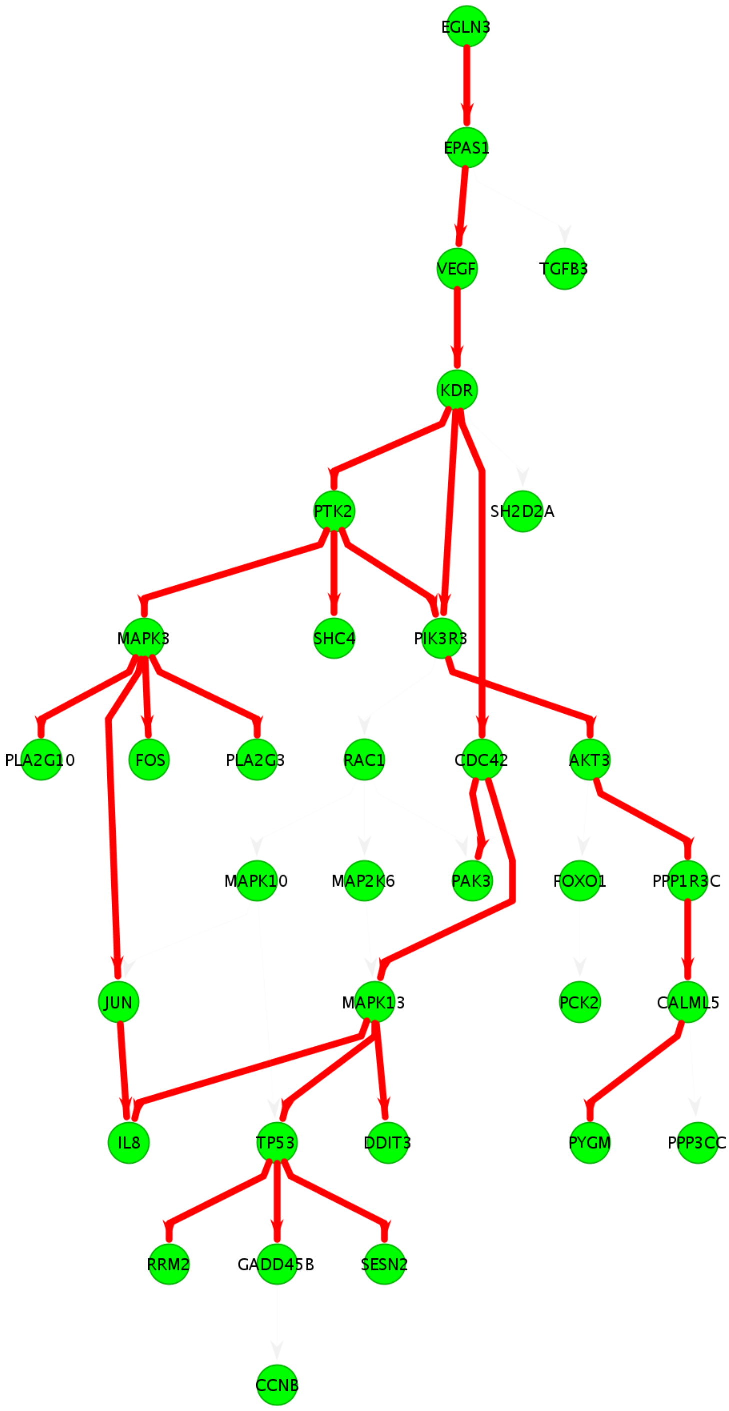


Figure 5.6: The most deregulated subgraph for BRCA1 mutation carriers against non mutation carriers for a network size of 25 with root node EGLN3 (p-value $<$ 0.001). The isolated nodes are part of the union network of the deregulated subgraphs of size 10-25.

riched that are associated with cancer, e.g., the KEGG pathways: “VEGF signaling pathway”, “MAPK signaling pathway”, “Focal adhesion”, “ErbB signaling pathway”, and the “p53 signaling pathway”. These pathways have in common that they influence crucial cell processes as proliferation, differentiation, cell motility, and survival. Furthermore, we can confirm the results of Burga et al. [188], since the genes of the detected subgraph are also enriched in the EGF pathway (MSigDB), as well as in the GO terms cell cycle and cell cycle arrest. Interestingly, we also find pathways or categories significantly enriched that are associated with hypoxia and oxidative stress, as e.g. “Hypoxia review”, “Hypoxia normal up”, and “Oxstress breastca up” from MSigDB. An overview of significantly enriched pathways from KEGG or MSigDB which cover at least 4 genes of the deregulated subgraph is summarized in Table 5.3.

To compare the results of our algorithm to a standard gene set enrichment analysis, we subjected the input list containing the genes sorted by the absolute values of their fold differences to the GSEA variant implemented in GeneTrail. The analysis revealed many significantly deregulated pathways (p-value $<$ 0.05, FDR adjusted), amongst others the KEGG pathways “cell cycle”, “DNA replication”, and “missmatch repair”. When regarding the MSigDB gene sets, we find the breast cancer related categories “BRCA ER neg”, “BRCA ER pos”, “Breast cancer estrogen signaling”, and “Breast ductal carcinoma genes”, as well as the hypoxia related category “Hypoxia reg up” significantly deregulated. Interestingly, in this analysis neither the p53 signaling pathway nor the EGF signaling pathway was significantly deregulated.

Table 5.2: List of genes found in the 16 computed deregulated subgraphs of sizes 10-25 and number of occurrences.

Gene ID	Gene Symbol	Gene Description	Number of deregulated subgraphs
7157	TP53	tumor protein p53	16
6241	RRM2	ribonucleotide reductase M2	16
5603	MAPK13	mitogen-activated protein kinase 13	16
4616	GADD45B	growth arrest and DNA-damage-inducible, beta	16
1649	DDIT3	DNA-damage-inducible transcript 3	16
7422	VEGFA	vascular endothelial growth factor A	16
3791	KDR	kinase insert domain receptor (a type III receptor tyrosine kinase)	16
2034	EPAS1	endothelial PAS domain protein 1	16
112399	EGLN3	egl nine homolog 3 (C. elegans)	16
83667	SESN2	sestrin 2	15
998	CDC42	cell division cycle 42 (GTP binding protein, 25kDa)	15
8503	PIK3R3	phosphoinositide-3-kinase, regulatory subunit 3 (gamma)	14
5063	PAK3	p21 protein (Cdc42/Rac)-activated kinase 3	13
3576	IL8	interleukin 8	11
5837	PYGM	phosphorylase, glycogen, muscle	9
51806	CALML5	calmodulin-like 5	9
5507	PPP1R3C	protein phosphatase 1, regulatory (inhibitor) subunit 3C	9
10000	AKT3	v-akt murine thymoma viral oncogene homolog 3 (protein kinase B, gamma)	9
891	CCNB1	cyclin B1	8
5533	PPP3CC	protein phosphatase 3 (formerly 2B), catalytic subunit, gamma isoform	5
7043	TGFB3	transforming growth factor, beta 3	5
3725	JUN	jun oncogene	2
8399	PLA2G10	phospholipase A2, group X	1
5879	RAC1	ras-related C3 botulinum toxin substrate 1 (rho family, small GTP binding protein Rac1)	1
5608	MAP2K6	mitogen-activated protein kinase kinase 6	1
5602	MAPK10	mitogen-activated protein kinase 10	1
5595	MAPK3	mitogen-activated protein kinase 3	1
5106	PCK2	phosphoenolpyruvate carboxykinase 2 (mitochondrial)	1
50487	PLA2G3	phospholipase A2, group III	1
399694	SHC4	SHC (Src homology 2 domain containing) family, member 4	1
2353	FOS	FBJ murine osteosarcoma viral oncogene homolog	1
2308	FOXO1	forkhead box O1	1
9047	SH2D2A	SH2 domain protein 2A	1
5747	PTK2	PTK2 protein tyrosine kinase 2	1

Table 5.3: Significantly enriched pathways which are covered by the genes of the deregulated subgraph of size 25. The p-values were computed by using the hypergeometric distribution test (ORA) with the genes of the subgraph as test set and the genes of the regulatory graph as reference set. The p-values are FDR adjusted.

Pathway Source	Pathway Name	p-value	number of genes in subgraph
KEGG	Pathways in cancer	0.000442969	12
KEGG	MAPK signaling pathway	0.000442969	11
KEGG	Focal adhesion	0.000442969	10
KEGG	VEGF signaling pathway	3.22812e-07	10
KEGG	Neurotrophin signaling pathway	4.64128e-05	9
KEGG	Renal cell carcinoma	5.15226e-07	9
KEGG	T cell receptor signaling pathway	0.000288768	8
KEGG	Toll-like receptor signaling pathway	0.000442969	7
KEGG	ErbB signaling pathway	0.000442969	7
KEGG	GnRH signaling pathway	0.000482278	7
KEGG	Insulin signaling pathway	0.00272689	7
KEGG	Chemokine signaling pathway	0.00333032	7
MSigDB	BOQUEST_CD31PLUS_VS_CD31MINUS_UP	0.017504	7
KEGG	Glioma	0.000452846	6
KEGG	Pancreatic cancer	0.000587537	6
KEGG	Fc epsilon RI signaling pathway	0.000879813	6
KEGG	Colorectal cancer	0.00118209	6
KEGG	B cell receptor signaling pathway	0.00333032	5
MSigDB	HYPOXIA_REVIEW	0.00484667	5
KEGG	Chronic myeloid leukemia	0.00486093	5
KEGG	Bladder cancer	0.00235075	4
KEGG	mTOR signaling pathway	0.00507319	4
KEGG	Epithelial cell signaling in Helicobacter pylori infection	0.00507319	4
KEGG	Non-small cell lung cancer	0.00670911	4
KEGG	Endometrial cancer	0.00803437	4
MSigDB	SHEPARD_CRASH_AND_BURN_MUT_VS_WT_UP	0.00860306	4
MSigDB	CHEN_HOXA5_TARGETS_UP	0.00922605	4
MSigDB	HYPOXIA_NORMAL_UP	0.0130697	4
MSigDB	METPATHWAY	0.014913	4
MSigDB	KERATINOCYTEPATHWAY	0.0205406	4
KEGG	Melanoma	0.0206123	4
KEGG	p53 signaling pathway	0.0210982	4
KEGG	Fc gamma R-mediated phagocytosis	0.0210982	4
MSigDB	SIG_PIP3_SIGNALING_IN_CARDIAC_MYOCTES	0.0267726	4
KEGG	Prostate cancer	0.0273038	4
KEGG	Small cell lung cancer	0.0288885	4
KEGG	Vascular smooth muscle contraction	0.0304471	4
MSigDB	ST_INTEGRIN_SIGNALING_PATHWAY	0.046262	4
MSigDB	RAS_ONCOGENIC_SIGNATURE	0.0471463	4
MSigDB	INSULIN_SIGNALING	0.0488835	4

5.2.3 Comparison to FiDePa

To compare the results of our ILP approach to FiDePa, we subjected the sorted list of genes containing the absolute values of the computed fold differences of the GSE13671 data set to a FiDePa analysis. We computed all significant paths of lengths up to eight. For each path length, the computation resulted in several hundred significant paths. Furthermore, the number of significant paths increased with the path length (147 at length 3, 478 at length 8). Interestingly, we found only four significant paths of different lengths that contained EGLN3 (see Table 5.4), which was the stable root node in the ILP analysis. Moreover, the overlap of these four significant paths to the most deregulated subgraph of size 25 in the ILP approach is two, namely EGLN3 and EPAS1. In addition, the most significant paths of length eight in the FiDePa analysis show only a minimal overlap with the most deregulated subgraph of the ILP approach.

Length	Path	p-value
4	EGLN3 → EPAS1 → PDGFB → PDGFRA → PDGFRB	0.027
6	EGLN3 → EPAS1 → TGFA → ERBB2 → EGFR → JAK1 → STAT1	0.027
8	EGLN3 → EPAS1 → TGFA → ERBB2 → EGFR → JAK1 → STAT3 → PRKAB2 → AGRP	0.032
8	EGLN3 → EPAS1 → TGFA → ERBB2 → EGFR → JAK1 → STAT3 → PRKAB2 → SLC2A1	0.032

Table 5.4: Four significant paths computed with FiDePa containing EGLN3

The most obvious disadvantage of the FiDePa approach is the huge number of significant paths that are found. This complicates the interpretation and evaluation of the results. Furthermore, the biological dependencies are much better mirrored in the computation of a deregulated subnetwork than in a simple path. Such a path cannot capture all effects that the deregulation has on the different components of a regulatory network. Therefore, we used in our previous application of the FiDePa algorithm the union graph consisting of the computed deregulated paths. The ILP approach computes direcly the most deregulated connected subgraph, however, this approach is dependent on the manual selection of a size for this network that influences the results. At the moment, we are working on an improvement of this algorithm to remove the dependency of this threshold. Moreover, our ILP approach enforces that the computed deregulated subnetwork is rooted in a possible key player. This key player may be responsible for the observed differences between the investigated conditions. In cancer, oncogenes, tumor suppressor genes, or other genetically altered genes that contribute to significant and crucial changes of regulatory and signaling processes can be considered as such key players. The long-term objective of this proposed model is to help identifying putative targets for an individualized tumor therapy.

5.2.4 Discussion and Conclusion

The identification of patterns of pathway deregulation is a crucial task in differential network analysis. Moreover, the determination of the initiators of the observed differences between the investigated conditions is a major challenge. With our connectivity model we do not only identify the most deregulated subgraph, but also a root node which may be one of the key players for the deregulation. We applied our method to expression profiles of nonmalignant primary mammary epithelial cells (PMECs) isolated from BRCA1 mutation carriers and women without BRCA1 mutations. BRCA1 germline mutations are associated with a predisposition for developing breast cancer. The cumulative breast cancer risk by 70 years of age in BRCA1 mutation carriers was estimated to be 65% [191]. Although familial breast cancers have been intensely studied, the exact processes influenced by the BRCA1 mutation which eventually result in the development of breast cancer are still elusive. Burga and co-workers found that the nonmalignant PMECs from BRCA1 mutation carriers contained a subpopulation of progenitor cells which showed an altered proliferation and differentiation in cell culture [188]. In concordance to these morphologic observations, the comparison of the expression profiles of the PMECs with and without BRCA1 mutations revealed an upregulation of the EGFR pathway, which they discussed as possible cause for the altered growth and differentiation properties. Our study confirms these results, since we also find in our deregulated subgraphs components of the EGF and p53 signaling pathway significantly enriched. Even more interesting, we were able to associate the genes in our deregulated subgraphs with oxidative stress. The designated root node of our deregulated network is the gene PHD3 (EGLN3), which is known to play an important role in hypoxia. Yan et al. [192] found that the occurrence of a HIF-1α positive phenotype and a PHD3 negative phenotype is correlated with BRCA1 tumors. However, in this study we find that PHD3 is overexpressed in the nonmalignant PMECs with BRCA1 mutations. Ginouves et al. discussed overactivation of PHDs during chronic hypoxia and its effects on HIFα [193]. They found that PHDs are the key enzymes triggering a feedback mechanism, which leads to a desensitization of HIF1/2α and protects cells against necrotic cell death. Additionally, the GADD (growth arrest and DNA damage-inducible) genes (GADD45B, DDIT3) found in our deregulated subgraph are involved in cell cycle arrest, repair mechanisms and apoptosis. An increased expression of these genes has also been described in studies examining cells in stressful conditions [194, 195]. The genes GADD45B and DDIT3 (GADD153) are also overexpressed in the BRCA1 mutation carrier expression data. This is another indication that the cells seem to be in a stressful state which may have origins in the processes involved in the hypoxia regulation. A recent study of Dai et al. [196] discussed the role of oxidative stress in dependence of obesity as a possible cause for increased breast cancer risk. When regarding cell cultures of PMECs as in our case, this factor should admittedly be of no relevance. We hypothesize that the described different growth properties of the PMECs with BRCA1 mutations are responsible for a disturbance in O_2 homeostasis, so that this may induce oxidative stress. Additionally, the activation of the aforementioned stress proteins can result in avoidance of necrosis or apoptosis and in this way lead to an increased overall survival of cells with genetic alterations. If the cells in risk of cancerous transformation show a different growth behavior which results in oxidative stress, targeting the genes involved in these processes to induce cell death may be a possible starting point for preventing the outbreak of the disease. The idea of using, e.g., PHDs, HIF-1α or its downstream targets as a potential therapeutic strategy has been already suggested by Ginouves et al. and Yan et al., respectively.

Taken together, the nonmalignant mammary epithelial cells with BRCA1 mutations exhibit many properties that are known from breast cancer. Our study indicates that the cells are in a stressful state potentially originated from the processes involved in the regulation of long-term oxidative stress. Moreover, it seems that it is a very thin line between a cancerous outcome and non-cancerous phenotype for BRCA1 mutated mammary epithelial cells considering the accumulated deregulation affecting multiple signaling pathways visible in our computed subgraphs. Performing a GSEA also reported hypoxia as a significant finding. However, since this category was just one of some hundred significant categories we may have as well missed this result or at least not have attached that much importance to it.

With our approach the most deregulated part of a network can be visualized and experts can directly grasp the processes involved in the deregulation. Although the interpretation is not always straightforward, our approach is at least a very powerful complement to the standard gene set and single gene analysis methods for microarray data. Furthermore, we showed that the application of our algorithm to already published data can yield new insights. As expression data and network data are still growing, methods as our ILP-based algorithm will be valuable to detect deregulated subgraphs in different conditions and help contribute to a better understanding of diseases.

5.3 Conclusion

In this chapter, we presented two sophisticated new algorithms for identifying deregulated components of regulatory networks. With FiDePa, we demonstrated that it is possible to derive patient specific deregulated subnetworks from the union of computed significant paths. Thus, we were able to impressively highlight the differences in network components for patients with same disease but different survival times. Our second algorithm uses an ILP and computes directly the most deregulated subgraph that is rooted in a special node from which all other nodes in the subgraph are reachable. With this model, we enforce that this root node has the properties of a putative key player that directly influences the observed differences in the considered states and may serve as a potential target for an individualized tumor therapy. Both approaches are helpful for gaining new insights from expression profiles.

CHAPTER 6

CONCLUSION

In this thesis, we have presented a comprehensive gene set analysis framework and its applications to different fields in cancer research. With GeneTrail we have developed a novel modular C++ framework suitable for the integration of various data sources, statistical methods, and algorithms for the computer-aided evaluation of high-throughput data. The basic functionality of GeneTrail is the detection of statistically enriched or depleted biological categories concerning the genes of examined data sets. Furthermore, we added a variety of features to our framework, e.g., the handling of expression data with GeneTrailExpress, the dynamical pathway visualization with BiNA, and the capability to perform differential network analysis.

GeneTrail is suitable for distinct groups of users, researchers without programming knowledge and developers. For the first group, we provide a straightforward and easy-to-use graphical interface as presented by our web-application that is worldwide accessible. The user is guided through the different necessary input steps and can access the results of the computation in different well-arranged file formats. For advanced users having at least a basic knowledge of the programming language C++, GeneTrail can be used as a rapid prototyping library to realize new biological categories, statistical methods, etc. or for processing and filtering the computed results.

While GeneTrail's capabilities are not limited to cancer related problems, we applied our framework to answer topical questions in different fields of cancer research. As a first application, we performed a comprehensive analysis of various putative characteristics of antigens that render them possible candidates for eliciting immune responses in cancer and autoimmune diseases. Our results provided further evidence for differences and similarities between tumor antigens and autoantigens. Furthermore, we disclosed a certain prevalence of sequence similarities to proteins of many organisms throughout all kingdoms of life in the tested antigen sets, which may be a possible cause why the autoantibody repertoire seems restricted to a limited number of self-proteins.

Second, we analyzed the putative target pathways and networks of miRNAs in different cancer types to further elucidate the methods of action of miRNAs in cancer. We performed a study of different cancer expression profiles which showed that targets of specific miRNAs were significantly enriched or depleted in these sets. Furthermore, we computed and illustrated the putative target networks of miRNAs and found indications that the regulation takes place on basis of balance and interplay of concentrations of miRNAs rather than by regulating some few important targets or hubs in the network. In summary, our findings confirmed and enforced the important role of miRNAs as key players of gene regulation in cancer.

Third, we performed differential network analyses with two different newly developed algorithms. The first algorithm, FiDePa, is based on the dynamic programming algorithm for the unweighted GSEA. Instead of computing the statistical significance of pre-defined biological pathways, FiDePa detects deregulated paths in a regulatory network that are statistically significant. Applying FiDePa to expression profiles of 100 high-grade glioma samples in comparison to 158 profiles of normal tissue samples, we demonstrated that it is possible to derive patient specific deregulated subnetworks from the union of computed significant paths. Our second algorithm for differential network analysis uses an ILP and computes directly the most deregulated subgraph that is rooted in a special node from which all other nodes in the subgraph are reachable. With this model we enforce that this root node has the properties of a putative key player that directly influences the observed differences in the considered states. To demonstrate the potential of this method, we computed the deregulated network of size 25 given expression profiles of BRCA1 mutation carriers and non-mutation carriers. Our evaluation indicates that oxidative stress plays an important role in epithelial cells of BRCA1 mutation carriers that may contribute to the later development of breast cancer. Both approaches may be suitable for facilitating the selection of optimal therapeutic agents and the identification of novel potential targets for an individualized therapy of cancer.

In summary, this thesis has led to the development of one of the most comprehensive non-commercial gene set analysis frameworks available for the research community. Besides our own contributions using GeneTrail for topical problems in cancer research, our web-application has also been successfully employed by groups worldwide working in various research areas and has been frequently cited. With the ongoing development and increasing user counts, we hope that our work will have a continuing positive impact on the research of other groups.

APPENDIX A: PUBLICATIONS

Peer reviewed journal articles

- **2010**
  - Keller, A., Leidinger, P., Bauer, A., ElSharawi, A., Haas, J. et al. miRNA signatures from human blood – promising biomarkers for human diseases. Manuscript in preparation.
  - Schuler, M., Keller, A., Backes, C., Phillipar, K., Lenhof, HP, Bauer, P. A comprehensive strategy for identifying functional categories from transcriptomic data sets in Arabidopsis thaliana. Manuscript in preparation.
  - Backes, C., Meese, E., Lenhof, HP, and Keller, A. A dictionary on microRNAs and their putative target pathways. Nucleic Acids Research.
  - Backes, C., Rurainski, A., Gerasch, A., Klau, G., Küntzer, J., Eggle, D., Hein, M., Keller, A., Burtscher, H., Kaufmann, M., Meese, E., and Lenhof, HP. An integer linear programming approach for finding deregulated subgraphs in regulatory networks using expression profiles. Submitted.
- **2009**
  - Keller, A., Backes, C., Gerasch, A., Kaufmann, M., Kohlbacher, O., Meese, E., and Lenhof, H.-P. (2009). A novel algorithm for detecting differentially regulated paths based on Gene Set Enrichment Analysis. Bioinformatics
- **2008**
  - Keller, A., Backes, C., Al-Awadhi, M., Gerasch, A., Küntzer, J., Kohlbacher, O., Kaufmann, M., and Lenhof, H.-P. (2008). GeneTrailExpress: a web-based pipeline for the statistical evaluation of microarray experiments. BMC Bioinformatics (9), 552.
  - Keller, A., Ludwig, N., Backes, C., Romeike, B.F., Comtesse, N., Henn, W., Steudel, W.I., Mawrin, C., Lenhof, H.-P. and Meese, E. (2008). Genome wide expression profiling identifies specific deregulated pathways in meningioma. Int J Cancer
- **2007**
  - Küntzer, J., Backes, C., Blum, T., Gerasch, A., Kaufmann, M., Kohlbacher, O., and Lenhof, H.-P. (2007). BNDB - The Biochemical Network Database. BMC

Bioinformatics, 8, 367.

- Keller, A., Backes, C., and Lenhof, H.-P. (2007). Computation of significance scores of unweighted Gene Set Enrichment Analyses. BMC Bioinformatics, 8, 290.
- Elnakady, Y. A., Rohde, M., Sasse, F., Backes, C., Keller, A., Lenhof, H.-P., Weissmann, K. J., and Müller, R. (2007). Evidence for the Mode of Action of the Highly Cytotoxic Streptomyces Polyketide Kendomycin. Chembiochem. 8(11), 1261-72.
- Backes, C., Keller, A., Küntzer, J., Kneissl, B., Comtesse, N., Elnakady, Y. A., Müller, R., Meese, E., and Lenhof, H.-P. (2007). GeneTrail–advanced gene set enrichment analysis. Nucleic Acids Res., 35, W186-192.

- **2006**
  - Küntzer, J., Blum, T., Gerasch, A., Backes, C., Hildebrandt, A., Kaufmann, M., Kohlbacher, O., and Lenhof, H.-P. (2006). BN++ - A Biological Information System. Journal of Integrative Bioinformatics, 3(2), 34.
- **2005**
  - Backes, C., Küntzer, J., Lenhof, H.-P., Comtesse, N., and Meese, E. (2005). GraBCas: a bioinformatics tool for score-based prediction of caspase- and granzyme b-cleavage sites in protein sequences. Nucleic Acids Res., 33, 208-213.
  - Comtesse, N., Zippel, A., Walle, S., Monz, D., Backes, C., Fischer, U., Mayer, J., Ludwig, N., Hildebrandt, A., Keller, A., Steudel, W.-I., Lenhof, H.-P., and Meese, E. (2005). Complex humoral immune response against a benign tumor: frequent antibody response against specific antigens as diagnostic targets. PNAS, 102(27), 9601-06.
- **2004**
  - Dönnes, P., Höglund, A., Sturm, M., Comtesse, N., Backes, C., Meese, E., Kohlbacher, O., and Lenhof, H.-P. (2004). Integrative analysis of cancer-related data using cap. FASEB J., 18(12), 1465-7.

Conference posters

- **2008**
  - Klatte, A., Schuler, M., Backes, C., Keller, A., Philippar, K., Fink-Straube, C.,Wirtz, M., Hell, R., Bauer, P. Nicotianamine is required for iron homeostasis and seed iron loading an example for applying the web-based gene chip data analysis software tool ’GeneTrail’. 5th Tri-National Arabidopsis meeting 2008.
  - Backes, C., Keller, A., Kuentzer, J., Gerasch, A., Kaufmann, M. and Lenhof, HP. GeneTrail - statistical evaluation and visualization of biological pathways. In BioSysBio 2008 conference.

APPENDIX B: CLASS DIAGRAM OF GENETRAIL

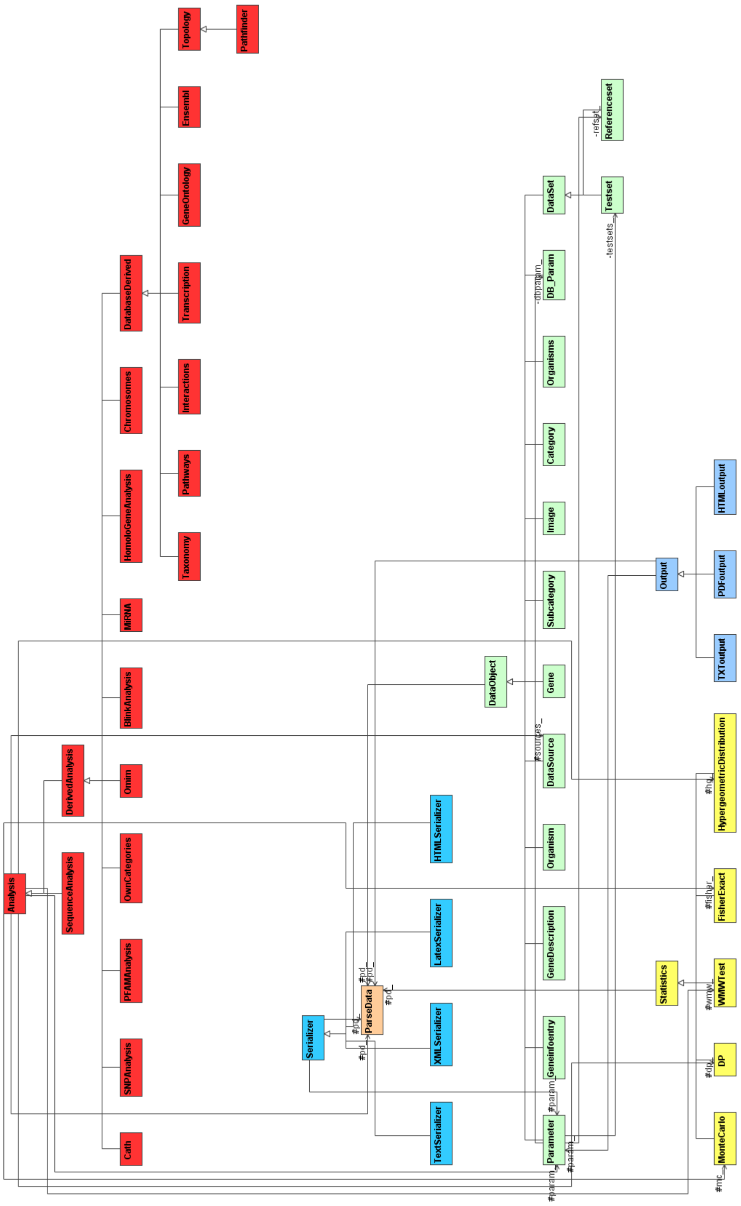


Figure B.1: Simplified class diagram of GeneTrail

APPENDIX C: DATA SOURCES

C.1 NCBI Entrez Gene

Entrez Gene[1] is the successor of the LocusLink database. This database provides information to genes and proteins, their official symbol, the gene description, PubMed links, and functional information if known. The Gene ID is a unique accession number which corresponds to exactly one gene in one organism. In GeneTrail we use this accession number as the origin for all analyses. If the user input contains other gene identifiers, we first transform them into Gene IDs before starting the analysis.

C.2 NCBI RefSeq

The NCBI reference sequences[2] (RefSeqs) are a non-redundant set of standard sequences including genomic DNA, transcript (RNA), and protein products. Additionally, the RefSeqs are annotated with information like chromosomal location or associated gene. RefSeq mRNAs have NM- or XM-prefixes, proteins NP- or XP-prefixes, correspondingly. We use these accession numbers to retrieve an amino acid sequence for the corresponding Gene ID. One Gene ID can be mapped to one or more RefSeq amino acid sequences (isoforms), but one RefSeq sequence can be mapped to exactly one Gene ID.

C.3 NCBI UniGene

The NCBI UniGene database[3] contains clustered sequences of transcripts. Each UniGene entry represents a set of transcript sequences that appears to come from the same transcription locus (gene or expressed pseudogene). Therefore, a UniGene ID is not necessarily unique for one gene. Different transcripts of a gene may be mapped to different UniGene cluster and vice versa one UniGene cluster can consist of transcripts of different genes. This type of accession number can be found frequently on microarray chips and can therefore be used for GeneTrail analyses.

[1] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene

[2] http://www.ncbi.nlm.nih.gov/RefSeq/

[3] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene

C.4 UniProt

The UniProt[4] Knowledgebase (UniProtKB) is a very comprehensive data source concerning functional information on proteins. Each UniProtKB entry contains besides the core data (amino acid sequence, protein name, taxonomic data, etc.) information about biological ontologies, classifications, and cross-references if available. The UniProtKB consists of a manually-annotated section (referred to as “UniProtKB/Swiss-Prot”) and a section with computationally analyzed records that await manual annotation (referred to as “UniProtKB/TrEMBL”). For our analyses, we always used the manually-annotated version of UniProtKB.

C.5 Ensembl

The Ensembl[5] project provides comprehensive species specific databases containing sequence information and additional annotations. For *H. sapiens*, we downloaded and installed a local version of the Ensembl MySQL core database. This database is the main source for gene and exon lengths, as well as for the mapping of NCBI Gene to Ensembl protein/gene IDs, and PDB[6] accession numbers. In addition, Ensembl provides web-services that can be accessed via a Perl API. We use this feature to retrieve a mapping of Ensembl IDs to NCBI Gene IDs for *M. musculus*. This way, we do not need to install an additional local database for this organism.

C.6 GEO

The Gene Expression Omnibus[7] (GEO) is a public repository for microarray, next-generation sequencing, and other forms of high-throughput data. GEO supports the MIAME (Minimum Information About a Microarray Experiment) standard that outlines the minimum information that should be included when publishing microarray data. GeneTrail and GeneTrailExpress make use of data and expression experiments in GEO, either for identifier mapping or for pre-processing deposited microarray experiments.

[4] `http://www.uniprot.org/`

[5] `http://www.ensembl.org/index.html`

[6] `http://www.rcsb.org/pdb/home/home.do`

[7] `http://www.ncbi.nlm.nih.gov/geo/`

C.7 Transpath

The Transpath[8] database focuses on regulatory pathways. The information available in Transpath has been manually curated from publications. Transpath provides detailed information about the intracellular signal transduction pathways, from signal induction on cell surface to the final target. Additionally, Transpath contains data about all molecules playing a role in signal transduction, as well as their reactions.

C.8 Transfac

Transfac[9] provides information about eukaryotic transcription factors, their genomic binding sites and DNA-binding profiles. Transfac is also integrated in the BNDB. GeneTrail uses this information to compute significant transcription factors regulating genes in an input set.

[8] http://www.gene-regulation.com/pub/databases.html

[9] http://www.gene-regulation.com/pub/databases.html

APPENDIX D: SUPPLEMENTAL MATERIAL

Table D.1: Scoring matrices for granzyme B and caspases 1-9. Amino acid preference distribution for each position Pi was extracted from Thornberry et al. [93] giving the most common amino acid a value of 1000.

	position Pi	AA of consensus recognition motif	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
Granzyme B	4	I	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1000	52	500	12	1	1
	3	E	1	297	54	1	153	477	1	198	1000	81	9	1	1	1	1	1	1	1	1	9
	2	P	1	752	544	1000	576	16	624	144	288	576	544	1	1	1	16	96	304	144	16	16
Caspase 1	4	W	1	48	48	48	48	16	16	80	96	16	128	1	1	1	96	288	80	496	576	1000
	3	E	1	357	442	1	442	425	119	374	1000	646	187	51	34	1	221	323	442	272	187	85
	2	H	1	144	396	144	180	18	72	36	72	108	1000	54	72	1	198	54	108	126	144	126
Caspase 2	4	D	1	1	50	10	1	1	10	1000	200	1	1	1	1	1	180	400	80	40	40	1
	3	E	1	425	884	1	680	119	119	102	1000	680	153	408	221	1	119	187	646	255	187	153
	2	H	1	624	528	352	336	48	96	1	1	16	1000	320	304	1	144	16	80	16	16	80
Caspase 3	4	D	1	40	50	1	10	1	20	1000	40	1	10	1	1	1	10	1	20	10	1	1
	3	E	1	306	357	1	357	85	153	255	1000	408	187	17	17	1	153	153	306	272	255	119
	2	V	1	14	378	406	182	1	14	1	0	14	196	42	0	1	714	224	1000	182	154	84
Caspase 4	4	W	1	80	208	144	96	48	80	288	256	96	48	1	1	1	304	848	224	384	352	1000
	3	E	1	187	119	34	119	17	85	221	1000	306	85	1	17	1	51	85	204	187	85	17
	2	H	1	102	119	119	425	17	102	221	357	153	1000	51	1	1	595	17	119	85	102	51
Caspase 5	4	W	1	14	56	1	56	126	42	98	98	84	56	1	1	1	280	1000	154	504	406	1000
	3	E	1	24	12	1	12	1	1	124	1000	12	12	1	1	1	12	12	12	12	12	1
	2	H	1	340	425	323	323	1	85	119	323	85	1000	17	34	1	272	17	272	153	204	102
Caspase 6	4	V	1	144	880	64	96	16	64	224	256	48	48	1	1	1	656	304	1000	80	48	48
	3	E	1	48	48	16	80	16	16	176	1000	144	48	1	1	1	16	16	48	48	16	48
	2	H	1	54	576	18	72	18	108	18	18	36	1000	54	36	1	648	486	918	288	216	558
Caspase 7	4	D	1	117	78	1	39	26	26	1000	104	26	39	1	1	1	13	1	13	13	13	1
	3	E	1	221	357	1	323	51	153	255	1000	425	187	85	102	1	306	221	697	204	204	102
	2	V	1	16	448	448	128	16	48	1	1	16	208	80	16	1	704	176	1000	160	128	48
Caspase 8	4	L	1	208	304	480	448	96	304	704	448	96	144	1	1	1	576	1000	720	224	256	144
	3	E	1	45	75	0	45	15	15	180	1000	150	45	1	1	1	45	15	105	45	45	15
	2	T	1	180	1000	216	324	18	126	72	198	108	306	72	72	1	720	108	792	180	198	306
Caspase 9	4	L	1	198	216	594	576	144	108	414	468	180	126	36	18	1	576	1000	684	252	216	144
	3	E	1	85	136	51	85	17	17	272	1000	187	119	1	17	1	102	119	204	102	85	51
	2	H	1	85	136	102	85	17	17	51	34	17	1000	34	1	1	187	17	153	51	34	51

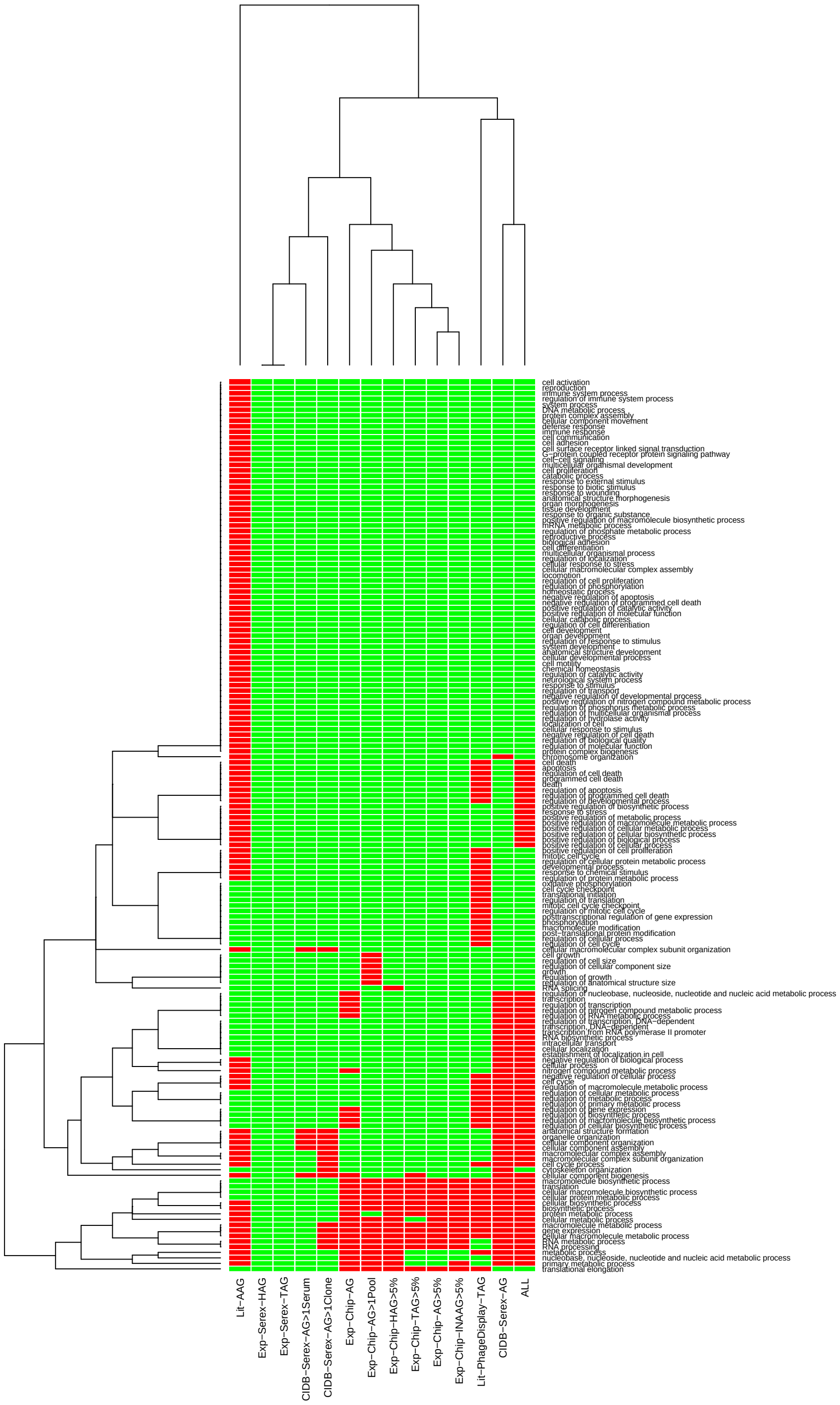


Figure D.1: The heatmap illustrates the significantly enriched GO terms of the biological process hierarchy in our antigen sets. Red = significantly enriched compared to the reference. Green = not significant or depleted.

Appendix E: Table of Abbreviations

APC	Antigen Presenting Cell
API	Application Programming Interface
BiNA	Biological Network Analyzer
BNDB	Biochemical Network Database
cDNA	complementary DNA
CSS	Cascading Style Sheets
CTL	Cytotoxic T Lymphocyte
DAG	Directed Acyclic Graph
DNA	DeoxyriboNucleic Acid
ECM	ExtraCellular Matrix
GEO	Gene Expressin Omnibus
GO	Gene Ontology
GSEA	Gene Set Enrichmen Analysis
GUI	Graphical User Interface
HLA	Human Leukocyte Antigen
HTML	HyperText Markup Language
ID	IDentifier
IEA	Inferred from Electronic Annotation
KEGG	Kyoto Encyclopedia of Genes and Genomes
MHC	Major Histocompatibility Complex
MIAME	Minimum Information About a Microarray Experiment
miRNA	microRNA
mRNA	messenger RNA
NCBI	National Center for Biotechnology Information
ORA	Over-Representation Analysis
ORF	Open Reading Frame
PHP	Hypertext PreProcessor
PSSM	Position Specific Scoring Matrix
RCGDB	Roche Cancer Genome Database

RNA	RiboNucleic Acid
rRNA	ribosomal RNA
SEREX	SErological identification of antigens by Recombinant EXpression cloning
SGD	Saccharomyces Genome Database
SVM	Support Vector Machine
TAIR	The Arabidopsis Information Resource
tRNA	transfer RNA
TS	tumor suppressor
UML	Unified Modeling Language
UniProtKB	UniProt Knowledgebase
WMW	Wilcoxon-Mann-Whitney
WWW	World Wide Web

Bibliography

[1] WHO - Cancer fact sheet February 2009. URL http://www.who.int/mediacentre/factsheets/fs297/en/index.html.

[2] Goodyear, M. D. E. Further lessons from the TGN1412 tragedy: New guidelines call for a change in the culture of research. *BMJ : British Medical Journal* **333**, 270–271 (2006).

[3] King in the kingdom of uncertainty. *Nat Biotech* **23**, 1025 (2005).

[4] Mootha, V. K. *et al.* PGC-1[alpha]-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* **34**, 267–273 (2003).

[5] Lamb, J. *et al.* A mechanism of cyclin d1 action encoded in the patterns of gene expression in human cancer. *Cell* **114**, 323–334 (2003).

[6] Dinu, I. *et al.* Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics* **8**, 242 (2007).

[7] Efron, B. & Tibshirani, R. On testing the significance of sets of genes. *Annals of Applied Statistics* **1**, 107 (2007).

[8] Ackermann, M. & Strimmer, K. A general modular framework for gene set enrichment analysis. *BMC Bioinformatics* **10**, 47 (2009).

[9] Al-Shahrour, F. *et al.* From genes to functional classes in the study of biological systems. *BMC Bioinformatics* **8**, 114 (2007).

[10] Backes, C. *et al.* GeneTrail–advanced gene set enrichment analysis. *Nucleic Acids Res.* **35**, W186–192 (2007).

[11] Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 15545–50 (2005).

[12] Nam, D. & Kim, S.-Y. Gene-set approach for expression pattern analysis. *Brief Bioinform* **9**, 189–197 (2008).

[13] Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucl. Acids Res.* **37**, 1–13 (2009).

[14] Ideker, T., Ozier, O., Schwikowski, B. & Siegel, A. F. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics (Oxford, England)* **18 Suppl 1**, S233–240 (2002).

[15] Rahnenführer, J., Domingues, F. S., Maydt, J. & Lengauer, T. Calculating the statistical significance of changes in pathway activity from gene expression data. *Statistical Applications in Genetics and Molecular Biology* **3**, Article16 (2004).

[16] Rajagopalan, D. & Agarwal, P. Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics (Oxford, England)* **21**, 788–793 (2005).

[17] Cabusora, L., Sutton, E., Fulmer, A. & Forst, C. V. Differential network expression during drug and stress response. *Bioinformatics (Oxford, England)* **21**, 2898–2905 (2005).

[18] Liu, C. *et al.* Topology-based cancer classification and related pathway mining using microarray data. *Nucleic Acids Research* **34** (2006).

[19] Liu, M. *et al.* Network-based analysis of affected biological processes in type 2 diabetes models. *PLoS Genetics* **3**, e96 (2007).

[20] Ulitsky, I., Karp, R. & Shamir, R. Detecting Disease-Specific dysregulated pathways via analysis of clinical expression profiles. In *Research in Computational Molecular Biology*, 347–359 (Springer Berlin / Heidelberg, 2008).

[21] Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T. & Muller, T. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics* **24**, i223–231 (2008).

[22] Nacu, S., Critchley-Thorne, R., Lee, P. & Holmes, S. Gene expression network analysis and applications to immunology. *Bioinformatics (Oxford, England)* **23**, 850–858 (2007).

[23] Chuang, H., Lee, E., Liu, Y., Lee, D. & Ideker, T. Network-based classification of breast cancer metastasis. *Molecular systems biology* **3**, 140 (2007).

[24] Keller, A. *et al.* GeneTrailExpress: a web-based pipeline for the statistical evaluation of microarray experiments. *BMC Bioinformatics* **9**, 552 (2008).

[25] Kuentzer, J. *et al.* BNDB - the biochemical network database. *BMC Bioinformatics* **8**, 367 (2007).

[26] Keller, A. *et al.* A novel algorithm for detecting differentially regulated paths based on gene set enrichment analysis. *Bioinformatics* btp510 (2009).

[27] Guarnieri, D. J. & DiLeone, R. J. MicroRNAs: a new class of gene regulators. *Ann. Med.* **40**, 197–208 (2008).

[28] Drakaki, A. & Iliopoulos, D. MicroRNA Gene Networks in Oncogenesis. *Curr. Genomics* **10**, 35–41 (2009).

[29] Vasudevan, S., Tong, Y. & Steitz, J. A. Switching from repression to activation: microRNAs can up-regulate translation. *Science* **318**, 1931–1934 (2007).

[30] Angelis, G. D., Rittenhouse, H. G., Mikolajczyk, S. D., Shamel, L. B. & Semjonow, A. Twenty years of PSA: from prostate antigen to tumor marker. *Reviews in Urology* **9**, 113–123 (2007). PMID: 17934568.

[31] Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).

[32] Wilcoxon, F. Individual comparison by ranking methods. *Biometric Bull* **1**, 80–83 (1945).

[33] Mann, H. & Wilcoxon, F. On a test of whether one of two random variables is stochastically larger than the other. *Ann Mat Stat* **18**, 50–60 (1947).

[34] Al-Shahrour, F., Diaz-Uriarte, R. & Dopazo, J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* **20**, 578–580 (2004).

[35] Maere, S., Heymans, K. & Kuiper, M. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics* **21**, 3448–3449 (2005).

[36] Beissbarth, T. & Speed, T. P. GOstat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics (Oxford, England)* **20**, 1464–1465 (2004).

[37] Lee, H. K., Braynen, W., Keshav, K. & Pavlidis, P. ErmineJ: tool for functional analysis of gene expression data sets. *BMC Bioinformatics* **6**, 269 (2005).

[38] Liu, C. *et al.* CRSD: a comprehensive web server for composite regulatory signature discovery. *Nucleic Acids Research* **34**, W571–577 (2006).

[39] Subramanian, A., Kuehn, H., Gould, J., Tamayo, P. & Mesirov, J. P. GSEA-P: a desktop application for gene set enrichment analysis. *Bioinformatics (Oxford, England)* **23**, 3251–3253 (2007).

[40] Breslin, T., Eden, P. & Krogh, M. Comparing functional annotation analyses with catmap. *BMC Bioinformatics* **5**, 193 (2004).

[41] Zhang, B., Kirov, S. & Snoddy, J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Research* **33**, W741–W748 (2005).

[42] Al-Shahrour, F., Minguez, P., Vaquerizas, J. M., Conde, L. & Dopazo, J. BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. *Nucleic Acids Research* **33**, W460–W464 (2005).

[43] Backes, C., Küntzer, J., Lenhof, H.-P., Comtesse, N. & Meese, E. GraBCas: a bioinformatics tool for score-based prediction of Caspase- and Granzyme B-cleavage sites in protein sequences. *Nucleic Acids Research* **33**, W208–W213 (2005).

[44] Küntzer, J. *et al.* BN++ - a biological information system. *Journal of Integrative Bioinformatics* **3** (2006).

[45] Backes, C. *GeneTrail - A statistical framework and web-application for analyzing gene set characteristics*. Master's thesis, Saarland University (2006).

[46] Kuentzer, J. *BN++ - A biological information system*. Ph.D. thesis, Saarland University (2008).

[47] Kanehisa, M. The KEGG database. *Novartis Found. Symp.* **247**, 91–101 (2002).

[48] Krull, M. *et al.* TRANSPATH: an information resource for storing and visualizing signaling pathways and their pathological aberrations. *Nucleic Acids Res.* **34**, D546–551 (2006).

[49] Matys, V. *et al.* TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–110 (2006).

[50] Salwinski, L. *et al.* The database of interacting proteins: 2004 update. *Nucl. Acids Res.* **32**, D449–D451 (2004).

[51] Zanzoni, A. *et al.* Mint: a molecular interaction database. *FEBS Lett.* **513**, 135–140 (2002).

[52] Hermjakob, H. *et al.* Intact - an open source molecular interaction database. *Nucl. Acids Res.* **32**, D452–D455 (2004).

[53] Peri, S. *et al.* Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* **13**, 2363–2371 (2003).

[54] Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A. & Enright, A. J. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* **34**, D140–144 (2006).

[55] Griffiths-Jones, S. miRBase: the microRNA sequence database. *Methods Mol. Biol.* **342**, 129–138 (2006).

[56] Griffiths-Jones, S., Saini, H. K., van Dongen, S. & Enright, A. J. miRBase: tools for microRNA genomics. *Nucleic Acids Res.* **36**, D154–158 (2008).

[57] Keller, A., Backes, C. & Lenhof, H.-P. Computation of significance scores of unweighted gene set enrichment analyses. *BMC Bioinformatics* **8** (2007).

[58] Kim, S. & Volsky, D. J. PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics* **6**, 144 (2005).

[59] Eden, E., Lipson, D., Yogev, S. & Yakhini, Z. Discovering motifs in ranked lists of DNA sequences. *PLoS Computational Biology* **3**, e39 (2007).

[60] Tiede, M. & Voß, W. *Schließen mit Statistik - Verstehen* (Oldenbourg Wissenschaftsverlag, 2000).

[61] Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Statist Soc B* **57**, 289–300 (1995).

[62] Yekutieli, D. & Benjamini, Y. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference* **82**, 171–196 (1999).

[63] Benjamini, Y. & Yekutieli, D. The Control Of The False Discovery Rate In Multiple Testing Under Dependency. *The Annals of Statistics* **29**, 1165–1188 (2001).

[64] Lee, L. Q., Lumsdaine, A. & Siek, J. G. *Boost Graph Library, The: User Guide and Reference Manual* (Addison Wesley Professional, 2001), 1st edn.

[65] Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* **13**, 2498–2504 (2003).

[66] Backes, C., Meese, E., Lenhof, H. & Keller, A. A dictionary on microRNAs and their putative target pathways. *Nucl. Acids Res.* gkq167 (2010).

[67] Elnakady, Y. A. *et al.* Evidence for the mode of action of the highly cytotoxic streptomyces polyketide kendomycin. *Chembiochem: A European Journal of Chemical Biology* **8**, 1261–1272 (2007).

[68] Keller, A. *et al.* Genome wide expression profiling identifies specific deregulated pathways in meningioma. *International Journal of Cancer* **124**, 346–351 (2009).

[69] Leidinger, P. *et al.* Novel autoantigens immunogenic in COPD patients. *Respiratory Research* **10**, 20 (2009).

[70] Chatterjee, I. *et al.* Staphylococcus aureus ClpC ATPase is a late growth phase effector of metabolism and persistence. *PROTEOMICS* **9**, 1152–1176 (2009).

[71] van Esse, H. P., Fradin, E. F., de Groot, P. J., de Wit, P. J. G. M. & Thomma, B. P. H. J. Tomato transcriptional responses to a foliar and a vascular fungal pathogen are distinct. *Molecular Plant-Microbe Interactions* **22**, 245–258 (2009).

[72] Hartlapp, I. *et al.* Depsipeptide induces cell death in hodgkin lymphoma-derived cell lines. *Leukemia Research* **33**, 929–936 (2009).

[73] Barenco, M. *et al.* Dissection of a complex transcriptional response using genome-wide transcriptional modelling. *Mol Syst Biol* **5** (2009).

[74] Si, H. *et al.* Human and murine kidneys show gender- and Species-Specific gene expression differences in response to injury. *PLoS ONE* **4** (2009).

[75] Vicentini, R. & Menossi, M. Pipeline for macro- and microarray analyses. *Brazilian Journal of Medical and Biological Research* **40**, 615–619 (2007).

[76] Wang, X. *et al.* NMPP: a user-customized NimbleGen microarray data processing pipeline. *Bioinformatics* **22**, 2955–2957 (2006).

[77] Pelizzola, M., Pavelka, N., Foti, M. & Ricciardi-Castagnoli, P. AMDA: an r package for the automated microarray data analysis. *BMC Bioinformatics* **7**, 335 (2006).

[78] Hokamp, K. *et al.* ArrayPipe: a flexible processing pipeline for microarray data. *Nucleic Acids Research* **32**, W457–459 (2004).

[79] Herrero, J. *et al.* GEPAS: a web-based resource for microarray gene expression data analysis. *Nucleic Acids Research* **31**, 3461–3467 (2003).

[80] Morris, J. A., Gayther, S. A., Jacobs, I. J. & Jones, C. A suite of perl modules for handling microarray data. *Bioinformatics (Oxford, England)* **24**, 1102–1103 (2008).

[81] Hu, Z., Mellor, J., Wu, J. & DeLisi, C. VisANT: an online visualization and analysis tool for biological interaction data. *BMC Bioinformatics* **5**, 17 (2004).

[82] Breitkreutz, B., Stark, C. & Tyers, M. Osprey: a network visualization system. *Genome Biology* **4**, R22 (2003).

[83] Los, M., Stroh, C., Jänicke, R. U., Engels, I. H. & Schulze-Osthoff, K. Caspases: more than just killers? *Trends in Immunology* **22**, 31–34 (2001).

[84] Algeciras-Schimnich, A., Barnhart, B. C. & Peter, M. E. Apoptosis-independent functions of killer caspases. *Current Opinion in Cell Biology* **14**, 721–726 (2002).

[85] Heusel, J. W., Wesselschmidt, R. L., Shresta, S., Russell, J. H. & Ley, T. J. Cytotoxic lymphocytes require granzyme b for the rapid induction of DNA fragmentation and apoptosis in allogeneic target cells. *Cell* **76**, 977–987 (1994).

[86] Sharif-Askari, E. *et al.* Direct cleavage of the human DNA fragmentation factor-45 by granzyme b induces caspase-activated DNase release and DNA fragmentation. *The EMBO Journal* **20**, 3101–3113 (2001).

[87] Casciola-Rosen, L., Andrade, F., Ulanet, D., Wong, W. B. & Rosen, A. Cleavage by granzyme b is strongly predictive of autoantigen status: implications for initiation of autoimmunity. *The Journal of Experimental Medicine* **190**, 815–826 (1999).

[88] Lohmüller, T. *et al.* Toward computer-based cleavage site prediction of cysteine endopeptidases. *Biological Chemistry* **384**, 899–909 (2003).

[89] Wee, L. J. K., Tan, T. W. & Ranganathan, S. SVM-based prediction of caspase substrate cleavage sites. *BMC Bioinformatics* **7 Suppl 5**, S14 (2006).

[90] Wee, L. J. K., Tan, T. W. & Ranganathan, S. CASVM: web server for SVM-based prediction of caspase substrates cleavage sites. *Bioinformatics (Oxford, England)* **23**, 3241–3243 (2007).

[91] Venkatraman, P., Balakrishnan, S., Rao, S., Hooda, Y. & Pol, S. A sequence and structure based method to predict putative substrates, functions and regulatory networks of endo proteases. *PloS One* **4**, e5700 (2009).

[92] Wee, L. J. K., Tong, J. C., Tan, T. W. & Ranganathan, S. A multi-factor model for caspase degradome prediction. *BMC Genomics* **10 Suppl 3**, S6 (2009).

[93] Thornberry, N. *et al.* A combinatorial approach defines specificites of members of the caspase family and granzyme B. Functional relationships established for key mediators of apoptosis. *J Biol Chem* **272**, 17907–17911 (1997).

[94] Darmon, A. J., Nicholson, D. W. & Bleackley, R. C. Activation of the apoptotic protease CPP32 by cytotoxic t-cell-derived granzyme b. *Nature* **377**, 446–448 (1995).

[95] Andrade, F. *et al.* Granzyme b directly and efficiently cleaves several downstream caspase substrates: implications for CTL-induced apoptosis. *Immunity* **8**, 451–460 (1998).

[96] Sutton, V. R. *et al.* Initiation of apoptosis by granzyme b requires direct cleavage of bid, but not direct granzyme b-mediated caspase activation. *The Journal of Experimental Medicine* **192**, 1403–1414 (2000).

[97] Thomas, D. A., Du, C., Xu, M., Wang, X. & Ley, T. J. DFF45/ICAD can be directly processed by granzyme b during the induction of apoptosis. *Immunity* **12**, 621–632 (2000).

[98] Janeway, C. A., Travers, P., Walport, M. & Shlomchik, M. *Immunobiology* (New York and London: Garland Science, 2001), 5th edn.

[99] Fujinami, R. S. & Oldstone, M. B. Molecular mimicry as a mechanism for virus-induced autoimmunity. *Immunologic Research* **8**, 3–15 (1989).

[100] Reuschenbach, M., von Knebel Doeberitz, M. & Wentzensen, N. A systematic review of humoral immune responses against tumor antigens. *Cancer Immunology, Immunotherapy: CII* **58**, 1535–1544 (2009).

[101] Engvall, E. & Perlmann, P. Enzyme-linked immunosorbent assay (ELISA). quantitative assay of immunoglobulin g. *Immunochemistry* **8**, 871–874 (1971).

[102] Sahin, U. *et al.* Human neoplasms elicit multiple specific immune responses in the autologous host. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 11810–11813 (1995).

[103] Türeci, O., Sahin, U. & Pfreundschuh, M. Serological analysis of human tumor antigens: molecular definition and implications. *Molecular Medicine Today* **3**, 342–349 (1997).

[104] Chen, Y., Gure, A. O. & Scanlan, M. J. Serological analysis of expression cDNA libraries (SEREX): an immunoscreening technique for identifying immunogenic tumor antigens. *Methods in Molecular Medicine* **103**, 207–216 (2005).

[105] Büssow, K. *et al.* A method for global protein expression and antibody screening on high-density filters of an arrayed cDNA library. *Nucleic Acids Research* **26**, 5007–5008 (1998).

[106] Naour, F. L. *et al.* Proteomics-based identification of RS/DJ-1 as a novel circulating tumor antigen in breast cancer. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* **7**, 3328–3335 (2001).

[107] Horn, S. *et al.* Profiling humoral autoimmune repertoire of dilated cardiomyopathy (DCM) patients and development of a disease-associated protein chip. *Proteomics* **6**, 605–613 (2006).

[108] Stadler, M., Arnold, D., Frieden, S., Luginbühl, S. & Stadler, B. Single nucleotide polymorphisms as a prerequisite for autoantigens. *European Journal of Immunology* **35**, 371–378 (2005).

[109] Doennes, P. *et al.* Integrative analysis of cancer-related data using CAP. *FASEB* **18**, 1465–1467 (2004).

[110] Angelopoulou, K., Yu, H., Bharaj, B., Giai, M. & Diamandis, E. P. p53 gene mutation, tumor p53 protein overexpression, and serum p53 autoantibody generation in patients with breast cancer. *Clinical Biochemistry* **33**, 53–62 (2000).

[111] Soussi, T. p53 antibodies in the sera of patients with various types of cancer: a review. *Cancer Research* **60**, 1777–1788 (2000).

[112] Mitrunen, K. & Hirvonen, A. Molecular epidemiology of sporadic breast cancer. the role of polymorphic genes involved in oestrogen biosynthesis and metabolism. *Mutation Research* **544**, 9–41 (2003).

[113] Cheng, L. *et al.* Glutathione-S-transferase polymorphisms and risk of squamous-cell carcinoma of the head and neck. *International Journal of Cancer. Journal International Du Cancer* **84**, 220–224 (1999).

[114] Kuentzer, J., Eggle, D., Lenhof, H., Burtscher, H. & Klostermann, S. The Roche Cancer Genome Database (RCGDB). *Human Mutation* **4**, 407–413 (2010).

[115] Ross, D. T. *et al.* Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics* **24**, 227–235 (2000).

[116] Smigielski, E. M., Sirotkin, K., Ward, M. & Sherry, S. T. dbSNP: a database of single nucleotide polymorphisms. *Nucl. Acids Res.* **28**, 352–355 (2000).

[117] Plotz, P. H. The autoantibody repertoire: searching for order. *Nat Rev Immunol* **3**, 73–78 (2003).

[118] Dohlman, J., Lupas, A. & Carson, M. Long charge-rich alpha-helices in systemic autoantigens. *Biochem Biophys Res Commun* **195**, 686–696 (1993).

[119] Strieter, R. M. *et al.* The Functional Role of the ELR Motif in CXC Chemokine-mediated Angiogenesis. *Journal of Biological Chemistry* **270**, 27348–27357 (1995).

[120] Wakasugi, K. *et al.* Induction of Angiogenesis by a Fragment of Human Tyrosyl-tRNA Synthetase. *J Biol Chem* **277**, 20124–20126 (2002).

[121] Buckley, C. D. *et al.* RGD peptides induce apoptosis by direct caspase-3 activation. *Nature* **397**, 534–539 (1999).

[122] Lupas, A., Dyke, M. V. & Stock, J. Predicting Coiled Coils from Protein Sequences. *Science* **252**, 1162–1164 (1991).

[123] Kanehisa, M. *et al.* From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **34**, D354–357 (2006).

[124] Anderton, S. M. Avoiding autoimmune disease - t cells know their limits. *Trends in Immunology* **27**, 208–214 (2006).

[125] Tatusov, R. L. *et al.* The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research* **29**, 22–28 (2001).

[126] Tatusov, R. L. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4** (2003).

[127] Jensen, L. J. *et al.* eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Research* **36** (2008).

[128] Muller, J. *et al.* eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucl. Acids Res.* gkp951 (2009).

[129] Greene, L. H. *et al.* The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Research* **35**, D291–297 (2007).

[130] Sonnhammer, E. L., Eddy, S. R. & Durbin, R. Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins: Structure, Function, and Genetics* **28**, 405–420 (1997).

[131] Finn, R. D. *et al.* The pfam protein families database. *Nucleic Acids Research* **36**, D281–288 (2008).

[132] Berman, H. M. *et al.* The protein data bank. *Nucleic Acids Research* **28**, 235–242 (2000).

[133] Yeats, C. *et al.* Gene3D: comprehensive structural and functional annotation of genomes. *Nucl. Acids Res.* **36**, D414–418 (2008).

[134] Lees, J., Yeats, C., Redfern, O., Clegg, A. & Orengo, C. Gene3D: merging structure and function for a thousand genomes. *Nucl. Acids Res.* gkp987 (2009).

[135] Lee, D., Grant, A., Marsden, R. L. & Orengo, C. Identification and distribution of protein families in 120 completed genomes using Gene3D. *Proteins: Structure, Function, and Bioinformatics* **59**, 603–615 (2005).

[136] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–410 (1990).

[137] Keller, A. *Understanding Cancer with Bioinformatics*. Ph.D. thesis, Saarland University (2009).

[138] Rurainski, A. *Optimization in Bioinformatics*. Ph.D. thesis, Saarland University (2010).

[139] Amedei, A. *et al.* Molecular mimicry between Helicobacter pylori antigens and H+, K+ –adenosine triphosphatase in human gastric autoimmunity. *J Exp Med.* **198**, 1147–1156 (2003).

[140] Ryan, K. R., Patel, S. D., Stephens, L. A. & Anderton, S. M. Death, adaptation and regulation: the three pillars of immune tolerance restrict the risk of autoimmune disease caused by molecular mimicry. *Journal of Autoimmunity* **29**, 262–271 (2007).

[141] Fadok, V. A. *et al.* Macrophages that have ingested apoptotic cells in vitro inhibit proinflammatory cytokine production through autocrine/paracrine mechanisms involving TGF-beta, PGE2, and PAF. *Journal of Clinical Investigation* **101**, 890–898 (1998).

[142] Speckmann, E.-J. *Physiologie* (Urban & Fischer Verlag, 2008).

[143] Jiang, H. *et al.* An affinity/avidity model of peripheral t cell regulation. *The Journal of Clinical Investigation* **115**, 302–312 (2005).

[144] Wu, Y., Zheng, Z., Jiang, Y., Chess, L. & Jiang, H. The specificity of t cell regulation that enables self-nonself discrimination in the periphery. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 534–539 (2009).

[145] Jiang, H. & Chess, L. How the immune system achieves self-nonself discrimination during adaptive immunity. *Advances in Immunology* **102**, 95–133 (2009).

[146] Knipe, D. M., Howley, P. M. & Griffin, D. E. *Fields Virology 2 Vol Set* (Lippincott Williams & Wilkins, 2006).

[147] Naparstek, Y. & Plotz, P. The role of autoantibodies in autoimmune disease. *Annu Rev Immunol* **11**, 79–104 (1993).

[148] Ruggero, D. & Pandolfi, P. P. Does the ribosome translate cancer? *Nature Reviews. Cancer* **3**, 179–192 (2003).

[149] Montanaro, L., Trere, D. & Derenzini, M. Nucleolus, ribosomes, and cancer. *The American Journal of Pathology* **173**, 301–310 (2008).

[150] Lu, H., Goodell, V. & Disis, M. L. Humoral immunity directed against tumor-associated antigens as potential biomarkers for the early diagnosis of cancer. *Journal of Proteome Research* **7**, 1388–1394 (2008).

[151] Tan, H. T., Low, J., Lim, S. G. & Chung, M. C. M. Serum autoantibodies as biomarkers for early cancer detection. *The FEBS Journal* **276**, 6880–6904 (2009).

[152] Medina, P. P. & Slack, F. J. microRNAs and cancer: an overview. *Cell Cycle* **7**, 2485–2492 (2008).

[153] Zhang, B., Pan, X., Cobb, G. P. & Anderson, T. A. microRNAs as oncogenes and tumor suppressors. *Dev. Biol.* **302**, 1–12 (2007).

[154] John, B. *et al.* Human MicroRNA targets. *PLoS Biol.* **2**, e363 (2004).

[155] Krek, A. *et al.* Combinatorial microRNA target predictions. *Nat. Genet.* **37**, 495–500 (2005).

[156] Kiriakidou, M. *et al.* A combined computational-experimental approach predicts human microRNA targets. *Genes Dev.* **18**, 1165–1178 (2004).

[157] Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15–20 (2005).

[158] Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215–233 (2009).

[159] Ruby, J. G. *et al.* Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *Genome Res.* **17**, 1850–1864 (2007).

[160] Lall, S. *et al.* A genome-wide map of conserved microRNA targets in C. elegans. *Curr. Biol.* **16**, 460–471 (2006).

[161] Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U. & Segal, E. The role of site accessibility in microRNA target recognition. *Nat. Genet.* **39**, 1278–1284 (2007).

[162] Miranda, K. C. *et al.* A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* **126**, 1203–1217 (2006).

[163] Maragkakis, M. *et al.* DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Res.* **37**, W273–276 (2009).

[164] Papadopoulos, G. L., Alexiou, P., Maragkakis, M., Reczko, M. & Hatzigeorgiou, A. G. DIANA-mirPath: Integrating human and mouse microRNAs in pathways. *Bioinformatics* **25**, 1991–1993 (2009).

[165] Nam, S. *et al.* MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression. *Nucleic Acids Res.* **37**, W356–362 (2009).

[166] Wachi, S., Yoneda, K. & Wu, R. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics* **21**, 4205–4208 (2005).

[167] Barrett, T. & Edgar, R. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Meth. Enzymol.* **411**, 352–369 (2006).

[168] Yanaihara, N. *et al.* Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. *Cancer Cell* **9**, 189–198 (2006).

[169] Dalmay, T. & Edwards, D. R. MicroRNAs and the hallmarks of cancer. *Oncogene* **25**, 6170–6175 (2006).

[170] Fabbri, M. *et al.* MicroRNA-29 family reverts aberrant methylation in lung cancer by targeting DNA methyltransferases 3A and 3B. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 15805–15810 (2007).

[171] Keller, A. *et al.* miRNAs in lung cancer - Studying complex fingerprints in patient's blood cells by microarray experiments. *BMC Cancer* **9** (2009).

[172] Su, A. I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 6062–6067 (2004).

[173] Lu, M. *et al.* An analysis of human microRNA and disease associations. *PLoS ONE* **3**, e3420 (2008).

[174] Chang, S. S. *et al.* MicroRNA alterations in head and neck squamous cell carcinoma. *Int. J. Cancer* **123**, 2791–2797 (2008).

[175] Volinia, S. *et al.* A microRNA expression signature of human solid tumors defines cancer gene targets. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 2257–2261 (2006).

[176] Calin, G. A. *et al.* A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. *N. Engl. J. Med.* **353**, 1793–1801 (2005).

[177] Chen, Y. T., Kitabayashi, N., Zhou, X. K., Fahey, T. J. & Scognamiglio, T. MicroRNA analysis as a potential diagnostic tool for papillary thyroid carcinoma. *Mod. Pathol.* **21**, 1139–1146 (2008).

[178] Phillips, H. *et al.* Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer Cell* **9**, 157–173 (2006).

[179] Su, A. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 6062–6067 (2004).

[180] Chattopadhyay, N., Tfelt-Hansen, J. & Brown, E. M. PKC, p42/44 MAPK and p38 MAPK regulate hepatocyte growth factor secretion from human astrocytoma cells. *Brain Research. Molecular Brain Research* **102**, 73–82 (2002).

[181] Cuevas, P. *et al.* Dobesilate diminishes activation of the mitogen-activated protein kinase ERK1/2 in glioma cells. *Journal of Cellular and Molecular Medicine* **10**, 225–230 (2006).

[182] Kam, A. Y. F., Tse, T. T. M., Kwan, D. H. T. & Wong, Y. H. Formyl peptide receptor like 1 differentially requires mitogen-activated protein kinases for the induction of glial fibrillary acidic protein and interleukin-1alpha in human u87 astrocytoma cells. *Cellular Signalling* **19**, 2106–2117 (2007).

[183] Schlegel, J., Piontek, G., Budde, B., Neff, F. & Kraus, A. The akt/protein kinase b-dependent anti-apoptotic pathway and the mitogen-activated protein kinase cascade are alternatively activated in human glioblastoma multiforme. *Cancer Letters* **158**, 103–108 (2000).

[184] The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).

[185] Parsons, D. W. *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science* **321**, 1807–1812 (2008).

[186] Holden, M., Deng, S., Wojnowski, L. & Kulle, B. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics* **24**, 2784–2785 (2008).

[187] Lin, R. *et al.* Gene set enrichment analysis for non-monotone association and multiple experimental categories. *BMC Bioinformatics* **9**, 481 (2008).

[188] Burga, L. N. *et al.* Altered proliferation and differentiation properties of primary mammary epithelial cells from BRCA1 mutation carriers. *Cancer Res* **69**, 1273–1278 (2009).

[189] Nemhauser, G. L. & Wolsey, L. A. *Integer and Combinatorial Optimization* (John Wiley and Sons, New York, 1988).

[190] Schrijver, A. *Theory of linear and integer programming* (John Wiley and Sons, 1998).

[191] Antoniou, A. *et al.* Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: a combined analysis of 22 studies. *American Journal of Human Genetics* **72**, 1117–1130 (2003).

[192] Yan, M., Rayoo, M., Takano, E. A., Thorne, H. & Fox, S. B. BRCA1 tumours correlate with a HIF-1[alpha] phenotype and have a poor prognosis through modulation of hydroxylase enzyme profile expression. *Br J Cancer* **101**, 1168–1174 (2009).

[193] Ginouves, A., Ilc, K., Macias, N., Pouyssegur, J. & Berra, E. PHDs overactivation during chronic hypoxia ”desensitizes” HIFalpha and protects cells from necrosis. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 4745–4750 (2008).

[194] Scott, D. W., Mutamba, S., Hopkins, R. G. & Loo, G. Increased GADD gene expression in human colon epithelial cells exposed to deoxycholate. *Journal of Cellular Physiology* **202**, 295–303 (2005).

[195] Oh-Hashi, K., Maruyama, W. & Isobe, K. Peroxynitrite induces GADD34, 45, and 153 VIA p38 MAPK in human neuroblastoma SH-SY5Y cells. *Free Radical Biology & Medicine* **30**, 213–221 (2001).

[196] Dai, Q. *et al.* Oxidative stress, obesity, and breast cancer risk: Results from the shanghai women’s health study. *J Clin Oncol* **27**, 2482–2488 (2009).