LUDWIG-MAXIMILIANS-UNIVERSITÄT
TECHNISCHE UNIVERSITÄT MÜNCHEN

Institute of Computational Biology (ICB),
Helmholtz Zentrum, München

Bachelorarbeit
in Bioinformatik

Development of Biomarkers by Integration of Data and Prior Knowledge

Toray Akcan

Aufgabensteller:	Prof. Dr. Dr. Fabian Theis
Betreuer:	Dr. Bettina Knapp
Abgabedatum:	15. Februar 2014

Ich versichere, dass ich diese Bachelorarbeit selbständig verfasst und nur die angegebenen Quellen und Hilfsmittel verwendet habe.

15. Februar 2014 _____

Toray Akcan

Abstract

Biological systems, as complex as they may be, exhibit certain behavioural patterns as a response to certain conditions. Although each system's pattern on the same condition differs to a certain extent from the others, it still remains a pattern and one can find analogies, when comparing them with each other. Key players, holding these patterns together, are called biomarkers. Biomarkers are biological measures of a biological system's state and can be used as indicators or predictors of conditions, be it internal or external. Biomarkers allow us to compare biological systems or processes, in particular, to examine a wealthy processe against a pathogenic processes or pharmacologic response to therapeutic treatment. Based on biomarkers, hypotheses about present or future biological conditions can be made, providing us with crucial biological knowledge, which in turn may serve as the basis of other research, particularly in disease diagnosis and treatment.

The discovery of biomarkers is still a challanging task and great efforts were undertaken to develop techniques and methods to approach this problem. To counter this problem, one of these methods (stSVM [10]) integrates biological network information as well as experimental data into one classifier. By smoothing genewise t-statistics over the graph structure of a PPI-network and subsequent classification, it is capable to provide us with accurate and in particular biologically interpretable results with high signature stability. Our approach makes use of this existing method and extends it by two components, an automated network retrievel system and an automated validation system through PCR data.

Our datasets mainly derive from former work done by Quaranta et al. [22], where the detection of biomarkers in two common inflammatory skin diseases, psoriasis and eczema, were targeted. Psoriasis and Eczema are two common widespread inflammatory skin diseases, whose phenotypic outcomes are quite similar, thus hampering to clearly differentiate between these two. The high inter-individual variability, partially based upon gender, age and short-term environmental exposure, makes it even harder to get a comprehensive understanding of disease pathogenesis. In addition, psoriasis ad eczema respond quite differently, even antipodal, to particular therapy methods, hence it is of high interest to develop specific therapies or diagnostic tools, in order to combat these diseases. This former research revealed 174 significantly up- or down-regulated genes either in psoriasis, eczema or both and is our primary input, available as microarray data. The second input forms the PCR data, where a few biologically significant genes, manually selected from the set of significantly up- or down-regulated genes, were remeasured for validation.

The two main targets of this study are to build an interface to the STRING database in order to fetch protein interaction data to compile biological networks and secondly to asses genes differential expression via gene-wise t-statistics based on microarray data and subsequent correction through gene-wise t-statistics obtained through PCR data. Fortunately, we were able to improve the aforementioned method (stSVM) and hope to alleviate further efforts of biomarker detection, on the basis of our work.

Abstract

Biologische Systeme, so komplex wie sie auch sein mögen, entwickeln bestimmte Verhaltensmuster als Reaktion auf einen Reiz bzw. Einfluss. Obwohl sich diese Muster von System zu System unterscheiden, können bestimmte wiederkehrende Charakteristika festgestellt werden. Charakteristika, die hauptsächlich fr die Struktur eines Musters verantwortlich sind, werden auch Biomarker genannt. Biomarker erfassen den Zustand eines biologischen Systems und können dafür eingesetzt werden, interne oder externe Einflüsse auf biologische Systeme zu identifizieren. Die Verwendung von Biomarkern erlaubt uns einen Vergleich zwischen biologischen Systemen oder Prozessen durchzuführen, insbesondere kann ein intakter biologischer Pozess mit einem fehlerhaften biologischem Prozess oder eine pharmakologische Antwort mit einer therapeutischen Behandlung verglichen werden. Auf Grundlage von Biomarkern ist es möglich Hypothesen über gegenwärtige oder zukünftige biologische Zustände zu formulieren, welche uns zu neuem biologischem Wissen verhelfen und als Grundlage anderer Forschungsgebiete dienen könnten, insbesondere bei der Diagnose und Behandlung von Krankheiten.

Das Auffinden von Biomarkern ist noch immer eine herausfordernde Aufgabe und viel Aufwand wurde betrieben um Methoden oder Techniken zu entwickeln um dieses Problem anzugehen. Eines dieser Methoden (stSVM [10]), welches versucht das Problem zu l¨sen, integriert biologische Netzwerkinformationen, sowie experimentelle Daten, gemeinsam in einen Klassifizierer. Durch die Kombinierung von Netzwerkinformationen mit t-Statistiken der gemessenen Gene, ist es der Methode möglich uns mit stabilen, akkuraten und insbesondere biologisch interpretierbaren Ergenissen zu versorgen. Unser Vorhaben macht Gebrauch von dieser Methode und erweitert sie um eine automatisierte Netzwerk Abfrage und einer automatiserten Validierung durch PCR Daten.

Unsere Daten entstammen hauptsächlich einer bestehenden Studie von Quaranta et al. [22], bei der das Auffinden von Biomarkern in zwei verbreiteten entzündlichen Hautkrankheiten, Psoriasis und Eczema, abgezielt wurde. Psoriasis und Eczema sind zwei der weit verbreitetsten entzündlichen Hautkrankheiten, deren sehr ähnliches phänotypisches Erscheinungsbild uns die Unterscheidung beider Krankheiten erschweren. Deren hohe Variabilität, welche teilweise vom Geschlecht, Alter oder kurzzeitigen Umwelteinflüssen beeinflusst wird, erschweren es nur noch mehr, einen umfassenden Überblick über die Krankheitshintergründe zu erlangen. Hinzu kommt, dass beide Krankheiten unterschiedlich, manchmal sogar gegensetzlich, auf bestimmte Therapien reagieren, deshalb ist es von groem Interesse spezifische Therapien oder diagnostische Mittel zu entwickeln, um sie zu bekämpfen. Die bestehende Studie war in der Lage 174 signifikant hoch- oder runter-regulierte Gene zu identifizieren, welche entweder in Psoriasis oder Eczema oder in beiden, signifikant reguliert waren und bilden unseren primären Datensatz, der Microarray Datensatz. Unseren zweiten Datensatz bilden PCR Messwerte für eine manuell selektierte Teilmenge von Genen der vorher erwähnten signifikanten Gene und dient zur Validierung der Microarray Daten.

Hauptaugenmerk dieser Studie liegt in der Entwicklung einer Schnittstelle zur STRING Datenbank um Proteininteraktionsdaten zum Zweck der Kompilierung von biologischen Netzwerken zu sammeln und in der Korrektur von auf Microarray Daten beruhenden t-Statistiken der Gene durch t-Statistiken welche auf Grundlage der PCR Daten gewonnen wurden. Erfreulicherweise ist es uns gelungen die vorher genannte Methode (stSVM) zu

verbessern und hoffen auf Grundlage unserer Studie weitere Bemühungen zur Entdeckung von Biomarkern erleichtern zu können.

Contents

List of Figures

List of Tables

1 Introduction

The main subject of our study is the identification of biomarkers for two widespread inflammatory skin diseases, namly *Psoriasis* and *Eczema*. To investigate this problem, we will first give an overview about these diseases, present existing approaches and formulate our idea. The second chapter builds the basis for our appraoch, capturing some graph theory, statistics and former efforts. The third chapter describes the realization of the idea and holds the results, lastly followed by a summary and some future prospects.

1.1 Disease Information

Psoriasis and eczema are complex imflammatory skin diseases mainly affected by genetic background and a modified immune response. The following should give an appropriate outline.

1.1.0.1 Psoriasis

Psoriasis is an organ-specific, chronic, uncontagious autoimmune disease, which affects approximately 1 -3 % of the world population and is primarily initiated by the immune system [12]. Other immune-mediated diseases based on epidemiological studies are Chrohn's disease, type-1 diabetes or multiple sclerosis, encouraging us to extend our knowledge on immune molecular mechanisms. Another important factor is the association of comorbidities with psoriasis, like psychiatric and psychosocial disorders [13], psoriatic arthritis or inflammatory bowel disease and in particular cardiovascular comorbidities, like obesity, dyslipidemia, hypertension or coronary heart disease [12].

Psoriasis can be categorized into five types, namely plaque, pustular, inverse, guttate and erythrodermic psoriasis, in which plaque psoriasis is the most common one, preferably affecting the elbows, knees and scalp [17] [2].

Triggers, which cause psoriasis or contribute to the aggravation of treatment, are, for instance, bacterial or viral infections, too much or too little sunlight, dry air or dry skin, injury to the skin, e.g. cuts, burns, insect bites, and some medicines, like antimalaria drugs, beta-blockers or lithium, just to mention a few [17]. As already noted, psoriasis is an autoimmune disease, thus people having a debilitated immune system, as it is the case in patients, who suffer from other autoimmune disorders or other diseases having great negative impact on the immune system, like AIDS, are more likely to be afflicted with severe psoriasis [17].

Following figures in Fig. 1.1 illustrate the possible degree of disease infestation.

(a) Psoriasis plaque, small-area

(b) Psoriasis plaque, large-area

Figure 1.1: Types of disease outbreak (Psoriasis)

To mention the most common symptoms, patients suffering from psoriasis often have irritated, red flaky patches of skin, which in turn are raised, thick and itchy, in addition adopting a color comparable to salmon and covered with silvery-white scales [17], as in figure 1.1 b).

Fortunately, there exist treatment methods, at least to control the symptoms and prevent further infection. The most conservative option is the topical treatment, where skin lotions, creams etc. are used to preserve and moisten the skin, to prevent desiccation. But also medical creams, shampoos or ointments, containing cortisone, salicylic acid or Vitamin D/A, prove of value to embank the disease symptoms.

Depending on the disease severity, one could take medicine beyond the ones described, so called systemic treatments. This includes medicine that supresses the immune system's faulty respone, retinoids or newer drugs called biologics, like Adalimumab or Alefacept, just to mention the most common. Unfortunately, systemic medicines operate body-wide, likely to currupt other metabolic processes, which weren't targeted, and eventually causing unwanted sideeffects.

A more locally specific approach is the phototherapy, where involved skin parts are exposed to ultraviolet light, eventually after increasing the skin's sensitivity to light with appropriate drugs to gain better results in the exposure step [17]. Ultraviolet light induces the decreasing of the number of skin cells that grow too quickly and kills T-cells, which may result in the clearing of psoriatic lesions. Phototherapy may be consulted when conventional treatments have not been effective [**?**].

All in all, psoriasis is a severe disease, making life much difficult for the affected. The high inter-individual variability of disease outcome and the analogy of appearance to other skin diseases makes it difficult to issue a correct diagnosis in the first place, which may lead to false treatment methods, urging us to get a better understanding of the disease background.

1.1.0.2 Eczema

Eczema, more specifically, atopic dermatitis is a common chronic, inflammatory skin disease of childhood, nevertheless, adults comprise one-third of all cases. The exact cause of atopic eczema is unknown so far, but it is believed that a genetic predisposition and the interplay of alllergic and non-allergic factors appear to play an important role in disease expression determination [33].

As in psoriasis, there exist other forms of eczema, namely contact dermatitis, dyshidrotic eczema, nummular eczema and seborrheic dermatitis.

People with atopic exzema suffer from ongoing swelling and redness of involved skin parts. Similar to an allergy, a faulty skin reaction seems to be the reason. Interestingly patients often test positive to allergy skin test, but it is proven that atopic dermatitis is not caused by allergies.

Factors, like allergies to pollen, mold or animals, dry skin, emotional stress, sudden temperature changes, cold, parfumes or dyes can worsen the disease symptoms [33].

To name a few, people having atopic dermatitis exhibit blisters, ear discharge or bleeding, skin color changes or skin redness or imflammation around the blisters [16]. Following figures 1.2 should give the reader an appropriate image of the disease.

(a) Severe eczema on both arms

(b) A patch of eczema that has been scratched

Figure 1.2: Types of disease outbreak (Eczema)

The diagnosis is primarily based on the overall appearance of the symptoms and personal and family history [16].

The treatment methods are quite similar to those available for psoriasis, but to sum up, it can be said that there exist by far more factors to take into account in daily lifetime, in order to avoid symptoms getting worse. Mainly topical medicines are used for most causes of atopic dermatitis [16].

On the whole, eczema is another life impeding burden, whose phenotypic outcome aggravates a clear differentiation from other skin diseases, like psoriasis, and whose concrete biochemical background is not yet deciphered, encouraging us for further investigations.

1.2 Former Approaches for Biomarker Detection

Biomarkers are biological measures of biological states and can be used as indicators of some conditions. Given a set of biomarkers, one can examine a normal wealthy biological process to a pathogenic process or a pharmacologic response to a therapeutic treatment. Biomarker measures enable us to formulate accurate hypotheses about present or future biological conditions or causes, hence playing a significant role in disease diagnosis and treatment. Great efforts have been made to identify such biomarkers, with some of them listed below.

PAM: Prediction Analysis for Microarrays PAM [65] was developed at the Stanford University Labs and is a freely abvailable class prediction and survival analysis tool, which performs a sample classification from gene expression data by implementing the *nearest shrunken centroid method* given in Literatur [65]. As a result, the method provides a list of significant genes whose combined expression pattern induces a categorization of the diagnostic classes, in addition, it estimates the prediction error through cross-validation, features false discovery rates (FDRs) and for survival outcomes it implements a supervised principle component method. PAM works with data from cDNA and oligo microarrays, protein expression data and SNP chip data. Besides, PAM is available as an R-package [64].

SAM: Significance Analysis of Mircoarrays SAM [31] is a supervised learning software for genomic expression data mining, developed at the Stanford University Statistics and Biochemistry Labs and is freely available. SAM is based on [31]. It can be applied to data from oligo, cDNA, SNP and protein arrays, or RNAseq data using the *SAMSeq* method [36]. SAM provides parametric and non-parametric tests to correlate expression data to clinical parameters, like treatment, diagnosis categories or survival time, just to mention few. For incomplete datasets, it offers an automated imputation of data points via the *nearest neighbor algorithm* [4]. Threshold adjustment enables the user to modify the classification outcome, that is the number of significant genes. In addition, it estimates FDR's for multiple testing through data permutation, but also reports local fase discovery rates or miss rates. In contrast to PAM, SAM is a statistical technique for finding significant genes rather than performing a sample classification like in PAM. is used to SAM is also available as an R-package, see [30].

RFE-SVM: Support Vector Machine Recursive Feature Elimination SVM-RFE [73] promises to eliminate gene redundancy in the detection of biomarker signatures for disease gene finding from microarray data, resulting in more reliable and compact gene subsets. The features are eliminated based on a decisive factor which is related to their support to the discrimination function. Furthermore, the SVM is embedded in a highly parallel GPU based environment, saving plenty of computation time.

Since SVM-RFE is a *greedy* method, several improvements are being made to this approach. One of them is given in literature [58]. It tries to compensate for this limitation and combines the SVM-RFE with local search operators based on operational research and artificial intelligence. In short, the core statement of this approach is, that the reuse of previously eliminated features improves the quality of the final classifier.

Network and Data Integration via Network Smoothed T-Statistics Cun et. al. [10] propose a technique that integrates network information for biomarker signature discovery. By smoothing t-statistics of individual genes over the structure of a PPI-network (2.2.1), possibly combined with a miRNA-target gene network. This is another SVM based approach, providing highly accurate, cross-validated results, primarily because of the bigger information content, deriving from the networks. As this thesis is based on this most recent approach, we will discuss this method in detail in section 2.6. The so called *netClass* R-package, implementing this idea, is available at [11].

1.3 Aim of this Work

This work ties in with former work by Quaranta et. al (see 2.5) for biomarker signature discovery for the aforementioned diseases in section 1.1. The main goal is to reproduce, verify and improve the results of this former work by the integration of network and PCR-data. This is done by extending the existing method described in 1.2 by an automated network compilation system with interaction data from the STRING database (see 2.1.0.2.4) and by integrating PCR-data for the correction of gene expression measurements. With this work we hope to alleviate biomarker signature discovery which are based on network information and subsequent validation through PCR.

2 Materials & Methods

2.1 Gene Query Engines

Bioinformatics is a data-driven discipline and it is of high interest for all participants to have free access to biological data in various levels, from tiny molecule structures, over organelles and cells, up to complex organisms. Each entity comprises information and the combination of these information may result in knowledge, knowledge, essential to get a better understanding of biological processes, to construct more sophisticated tools to extract information and gain even more knowledge or in the best case, to develop mechanisms to manipulate these processes and develop pharmaceuticals to combat diseases or provide early diagnosis of disorders.

Some well known biological databases, primarily capturing protein interactions, are for instance the Molecular Interactions (MINT) [1], the Database of Interacting Proteins (DIP) [34], the BioGRID database [7] or the Human Protein Reference Database (HPRD) [66].

In the following, we will give a short overview of biological resources used in this or reffered work. These biological resources provide information through computer readable formats, like the *Systems Biology Markup Language (SBML)* [43], similar to the XML-language. It can represent metabolic networks, cell signaling pathways, regulatory networks and other kinds of systems. Other common formats are the *Proteomics Standards Initiative Interaction (PSI-MI)* [32], the *Chemical Markup Language (CML)* [52,57] for chemicals or *BioPAX* [50] for pathways. This kind of standardization increases the compatibility of information gathered from distinct sources and alleviates information parsing and saves time [54].

2.1.0.2.1 Pathway Commons Database Pathway Commons [15] is a free of charge network biology resource, providing a comprehensive collection of publicly available pathways from multiple organisms represented in a common language. The provided data is being extracted from several source databases, like REACTOME, PID, PhosphoSitePlus and several others. One might visit [18] for a complete source listing as well as statistical information, like how many pathways or interactions were included from database x.

Given pathways can include biochemical reactions, complex assembly, transport and catalysis events, physical interactions involving proteins, DNA, RNA, small molecules and complexes, gene regulation events and genetic interactions involving genes.

Access to the data is given by the web interface, by batch downloads or the Web API, which enables us to programmatically access information and process the computer readable responses for integration with other network analysis components.

The quality of the data depends on the quality of the data from the source databases mentioned before. But for a higher flexibility, Pathway Commons allows several filter options, including data source, such that high quality data can be subsetted.

2.1.0.2.2 Kyoto Encyclopedia of Genes and Genomes (KEGG) The KEGG database [46] has been established in 1995 in Japan and has become a very popular, free of charge network biology resource. It integrates genomic, chemical and systemic functional information. Mainly, gene catalogs from completely sequenced genomes are linked to higher-level systemic functions of the cell, the organism and the ecosystem. The most exciting thing about KEGG is the major effort to manually create a knowledge base for such systemic functions and the ongoing efforts to develop and improve the cross-species annotation procedure for linking genomes to the molecular networks.

Despite the high-quality data, KEGG provides a great set of software, which enables the user to access and process the data more efficiently. These include a graphical interface for the exploration of KEGG global maps, an automatic annotation tool or a similarity search tools for chemical structures or sequences. Following link comprises a complete listing of the available software [45].

One might be interested in KEGG statistics, which lists the amount of, for example, pathway maps, organisms, metabolites or human diseases in the databases. For a complete listing one might visit [44]

2.1.0.2.3 EMBL - European Bioinformatics Institute EMBL [28] is an intergovernmental organisation and one of the world's leading free of charge research institutes, with its main laboratory located in Heidelberg, Germany. It provides massive high-quality data, such as an archive of protein expressions data determined by mass spectrometry, a database that shows which genes are expressed under which conitions, a resource for the analysis of metagenomic data, a database for the classification of proteins into families, domains or conserved sites, three-dimensional structural data on biological macromolecules and their complexes, biological pathways and much more. Following link [25] gives acces to documents giving an extensive overview about EMBL in general and specific. EMBL is truly an invaluable institution providing the user with a broad selection of biological data.

As expected, EMBL provides programmatic access to various data resources and analysis tools. Available software facilitates data retrievel, analysis of data, sequence similarity search, building of pairwise or multiple sequence alignments, contruction of phylogenetic trees, sequence translation, statistics and format conversion, moreover structural analysis and literatur and ontology search. Following webpage [26] provides a detailed listing of the software available for each mentioned topic.

Because of the rich data supply and diversity of tools, EMBL developed additionally a training programm for the inexperienced, to make the most of their services [27].

2.1.0.2.4 Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) STRING [21] is another free of charge network biology resource, with known and predicted interaction data, physical or functional for various species, mainly derived from four sources, specifically from genomic context, high-throughput experiments, (conserved) coexpression analysis and from previous knowledge.

A web interface is available to access the data and get a nice overview of the proteins and their interactions, represented as a colored network, either directed or undirected, moreover, with the option to filter out evidences, like textmining or gene fusion, in order to restrict the result to more reliable interaction evidences, like experiments only. Each interaction is scored, in particular, a combined score is evaluated, which combines the different evidence scores, for a more detailed description please see [21]. But there are

far more options to manipulate the results or process them further with different tools, e.g. clustering tools or to pipeline it with other resources like KEGG or perform a GO enrichment analysis. In addition to the web interface, STRING provides an application programming interface (API), which enables the user to easily request data by only constructing an appropriate URL.

Mainly, further work (3.2) will source information from the STRING database.

2.2 Graph Theory

The graph theory is an upcoming subject in a wide variety of disciplines. It arises in computer and math sciences (e.g. world wide web, complexity theory, logic, algebra), sociology (e.g. human relationship networks), business sciences (e.g. cost and time minimization problems) or biochemistry and biology (e.g. cristallography, protein interaction networks, neural networks) just to mention a few. This topic and its application is essential to many problem statements, as it is for this work. In specific, we will concentrate on biological networks, as part of the systems biology approach. As a start, we will give a short overview of the different types of networks, which occur in biological systems.

2.2.1 Biological Network Types

In this section we will concentrate on the various types of biological networks, leaving the mathematical implementation aside, that is the actual representation as a graph.

Protein-Protein Interaction (PPI) Networks Protein-Protein interactions play a major role in almost all biological processes. A PPI is characterized by the interplay of two or more proteins through diverse types of bonds, like covalent, ionic, polar or hydrogen bonds. Each type of bond has a major effect on the properties of the interaction, e.g. the interaction strength or polarity, which in turn may result in other conditions, leading to the interplay with other components and in general, to the dynamics of the whole system.

PPI's can be detected by pull down assays, tandem affinity purification (TAP), yeast two-hybrid (Y2H), mass spectrometry, microarrays or phage display, just two mention the most common [54].

Gene/Genetic Regulatory Networks (GRN) GRN's are another type of networks, occuring in biological systems and incorporating information about the control of gene expression. They play a major role in understanding the dynamics of a biological system. This highly dynamical system is influenced by a variaty of variables, such as transcription factors, post-translational modifications, presence or absence of specific molecules, so in general, molecule concentrations or the association with other biomolecules. Interestingly, these networks exhibit specific motifs and patterns concerning their topology, allowing us to biologically interpret and generalize certain properties.

Data, based on Protein-DNA interactions, is available in databases like JASPAR [59] or TRANSFAC [14], while post-translational modifications can be found in databases like Phospho.ELM [24], NetPhorest [51] or PHOSIDA [23] [54].

Signal Transduction Networks Signal transduction is the capability of a cell, to recieve extracellular signals, remit it in the intracellular space through the membrane, eventually

amplifying it, and finally to activate signal dependent bioentities within the cell, as a respone to the signal, and causing specific global effects. All in all, signal transduction allows the cell to respond to certain environmental parameters. The resulting signal transduction networks share common behavior with GRN's, in the sense that they exhibit specific patterns and motifs either. Data about signal transduction pathways can be accessed by the MiST [38] or TRANSPATH [47] databases [54].

Metabolic Networks A series of chemical reactions occuring within a cell is called a *metabolic pathway*. Since numerous distinct pathways exist in a cell, the entirety of metabolic pathways is defined as the metabolic network. Enzymes play a major role in such networks, since they catalyze these chemical reactions, but vitamins or cofactors play an important role either, as enzymes require them to function properly. Modern sequencing techniques allow the comprehensive reconstruction of such biochemical reactions, not only in humans. Among the databases capturing information about biochemical networks are the Kyoto Encyclopedia of Genes and Genomes (KEGG, see 2.1.0.2.2), EcoCyc [35], BioCyc [55] or metaTiger [37] [54]. In section 3.2 we will make use of PPI-networks.

2.2.2 Definitions and Properties

In order to correctly define and characterize the aforementioned network types, we will give a basic introduction into graph theory.

Definition A graph G is a pair (V, E) where V is a set of vertices representing the nodes of the graph and E is a set of edges representing the connections between the nodes in the graph. More specifically, $E = \{(i, j)|i, j \in V\}$, such that for any two nodes i and j, with either a directed or non-directed connection and a possible connection weight, the pair (i, j) will be in E, representing the edge between the nodes i and j. We call i and j *neighbors* or *adjacent*.

In the case of directed graphs, E is a set of *ordered* pairs (i, j), such that $(i, j) \neq (j, i)$.

As a edge weight, one can define anything appropriate to the problem statement, but it can be said that biological networks mainly hold real numbers as connection weights, giving the confidence of the connection.

In some cases, two nodes may have more then one connection between them, called a multi-edge connection. This kind of graphs occur, when a connection between two nodes incorporates more than one information, meaning that multiple connection weights with different properties exist, e.g. experimental confidence or literatur confidence score. In the upcoming section 3.2 we will see an application of such multi-edge graphs in connection with the STRING (2.1.0.2.4) database [54].

In biological networks, nodes may have self loops, meaning a node has an edge directing it to itself [68].

Finally, the definition of a *simple graph* is required. A simple graph is an unweighted, undirected graph with no self loops or multiple edges [69].

2.2.3 Network Graph Representations

There exist two main data structures to store network graphs representations, namely the *adjacency matrix* or the *adjacency list*.

In case of fully connected networks, in general, dense networks, the adjacency matrix is highly suggested. For a given graph $G = (V, E)$, this is a $|V|x|V|$ matrix A, with $A(i, j) = 1$ if $(v_i, v_j) \in E$, else 0, where $v \in V$. If G is a weighted graph capturing connection weights $w(v_i, v_j)$, then $A(i, j) = w(v_i, v_j)$. Please note, that in case of undirected graphs, the matrix is symmetric because $(v_i, v_j) = (v_j, v_i)$, both being in E, hence $A(i, j) = A(j, i)$, so that both, the upper and lower matrix triangle parts, comprise exactly the same network information.

Conversely, the adjacency list representation is appropriate for graphs with low density of connections, so called sparse graphs. But is it also depends on the use of the graph, since matrix operations are much easier than on lists, whereas lists handle more easy childnodes. So the choice strongly depends on the targeted application. The representation consists of an array A with linked lists corresponding to each $v \in V$ and containing all its adjcent neighbors, so all $u \in V$ which satisfy $(v, u) \in E$ [54].

In further work, we will make use of adjacency matrices, for reasons being later discussed.

Graph and Local Properties It is very usful to get an overall image of the network, for that, diverse graph attributes can be computed. In the following we cover the most common graph properties as defined in [54].

Node Degree One might compute the *degree* $deg(i)$ *of a node* i, which gives the number of connections of node i, in specific, $deg(i) = |\{(i, j)|(i, j) \in E\}|$, $\forall j \in V$ and a fix, but abitrary $i \in V$. In case of directed graphs, one has to consider the incoming and outgoing edges seperately, resulting in two different degrees, the *in-degree* $deg_{in}(i)$ and the *out-degree* $deg_{out}(i)$, respectively.

Nodes having a significantly high degree compared to other nodes often play a major role in the corresponding network and are called *hubs*.

Graph Density The graph density reflects the sparse- or denseness of a graph in response to the number of connections per node set. The density *dens* for a graph G is defined as $dens_G = \frac{2|E|}{|V|(|V|-1)}$. A graph is considered to be sparse, if following condition is met, $|E| = O(|V|^k)$ with $2 > k > 1$, else dense.

In general, biological networks are sparsely connected, thus preserving robustness, since, for instance, the probability to disturb a critical node, like a hub (see 2.2.3) for example, is far lower than in dense graphs.

Clustering Coefficient A cluster of a graph is a subset of vertices being highly connected to each other. Then *local clustering coefficient* C_i of vertex i in an undirected graph G is given by $C_i = \frac{2e}{k(k-1)}$, where $k = deg_G(k)$ and e is the number of edges between the k neighbors of i in G, thus C_i is in the range of $[0; 1]$, with $C_i = 1$ indicating that the neighbors of i are fully connected, otherwise for $C_i = 0$ fully unconnected.

To compute the global tendency of a graph to be divided into clusters, the *average clustering coefficient* $C_{average}$ is defined as $C_{average} = \frac{1}{N} \sum_{i=1}^{N} \frac{E_i}{k_i(k_i-1)}$, where $N = |V|$. High $C_{average}$ values indicate the tendency of a network to form clusters.

Completeness and Cycles A graph is a *complete* graph, if every pair of nodes is adjacent, that is if $\forall i, j \in V : (i, j) \in E$. For a complete graph with $|V| > 2$ there exist always a

cycle, which is a *walk* through the graph beginning at a node i, passing several other nodes and returning back to node i. In other words, it is a specific sequence $(v_1, v_2, \ldots, v_e)$ such that $\{(v_1, v_2), (v_2, v_3), \ldots, (v_{e-1}, v_e)\} \subseteq E$, where $v_e = v_i$, with no other node repeated, and the length of the sequnce must be greater than 3, otherwise it is just a connection between two nodes. If the graph doesn't contain any cycle, then it is called *acyclic*.

Node Distance and Graph Diameter Lastly we define the terms *distance*, *average path length* and *dimater* in the connection with graphs. The distance $\delta(i, j)$ from network node i to j is the length of the *shortest path* from node i to j in the corresponding graph. If there is no connection between node i and j, such that $(i, j) \notin E$, then we set $\delta(i, j) = \infty$, assuming that the distance is so far, that they are not connected. The shortest path problem plays a major role in graph and complexity theory and there exist several algorithms to solve the problem, being the reason why we are going to examine it in an own section (2.3).

Finally, the diameter D of a graph is the longest distance within a network, given by $D = max(\{\delta(i, j) | \forall i, j \in V \wedge (i, j) \in E\})$ and the average path length δ_{avg} is defined as, $\delta_{avg} = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \delta(i, j)$, where $N = |V|$.

2.3 The Shortest Path Problem

2.3.1 Definition

Given a graph $G = (V, E)$, the *shortest path problem* is the problem of finding a trail between two nodes $u, v \in V$ with minimal edge costs. In case of edge weights of 1, the solution to the shortest path problem is a path between u and v with the minimal amount of edges between u and v. If the graph is directed, the shortest path is a trail with consecutive vertices being connected by an appropriate directed edge. In the following we will give a precise definition.

A *path* or *trail* in an undirected graph is a *sequence* S of consecutive adjacent nodes, so $S = (v_1, v_2, \ldots, v_e)$, where $v_1, \ldots, v_e \in V$ and v_i is adjacent to v_{i+1} for $1 \leq i < e$. Let f be a real-valued weight function with $f : E \longmapsto \mathbb{R}$ and $e_{i,j}$ the edge incident to both v_i and v_j, then the shortest path from node v to v' is the sequence $S = (v_1, v_2, \ldots, v_n)$ that minimizes the sum $\sum_{i=1}^{n-1} f(e_{i,i+1})$ over all possible n, where $v_1 = v$ and $v_n = v'$. As already mentioned, in case of $f : E \longmapsto 1$, the shortest path problem is equivalent to finding a path with fewest edges. For directed graphs, the definition of a *path* slighty differs, in such a way that the sequence S comprises nodes which are adjacent through appropriate directed edges.

The shortest path problem can be subdivided into the *single-source shortest path problem*, in which all shortest paths from a source vertex v to all other vertices in the graph are computed, the *single-destination shortest path problem*, where shortest paths from all vertices to a specific vertex v are searched, which indeed is the reverse of the prior subproblem, and finally the *all-pairs shortest path problem*, in which all shortest paths for all possible pairs of vertices v and v' are computed [8].

Actually, the shortest path problems are closely related to our daily life or present in many computer applications, for instance, the finding of the shortest route when traveling, finding of shortest trading routes, finding the shortest path in peer-to-peer applications or finding the shortest path in scocial networks (degree of relatedness).

There are several algorithms available to solve these kind of problems and is our topic in the next section.

2.3.2 Algorithms

Common algorithms are the (a) *Dijkstra*, (b) *Bellman-Ford*, (c) *A*- search* or the (c) *Floyd-Warshall* algorithm. These algorithms differ in their purpose, (a) solves the single-source shortest path problem, (b) as (a), but with possible negative edge weights, (c) as (a), but tries to save computation time by heuristics and finally (d), which solves the all-pairs shortest path problem [8].

In fact, the most widely used algorithms are the Dijkstra's greedy algorithm and Floyd's dynamic algorithm with running time complexity of $O(N^2)$ and $O(N^3)$, respectively, where $N = |V|$.

The following gives a description about the functioning of the latter.

Given a graph $G = (V, E)$ and the single-sorce shortest path problem to be solved for node v_{start}, the Dijkstra algorithm first assigns every node $v \in V$ a tentative distance value, 0 for v_{start} and infinity for all other nodes, so

$$dist(v_{start}, v_{start}) = 0 \tag{2.1a}$$

$$dist(v_{start}, v') = \infty, \tag{2.1b}$$

where $v' \in V \backslash \{v_{start}\}$. Secondly, a new set U of *unvisited nodes* is created, consisting of all nodes at the beginning, so

$$U = V \tag{2.2}$$

These two steps can be categorized into the *preprocessing stage*, and are only performed once when the algorithm starts.

Next, for the *current node* v_{curr}, at the beginning

$$v_{curr} = v_{start}, \tag{2.3}$$

the tentative distances to all its *unvisited* neighbors are calculated. These are the distances

$$dist(v_{curr}, v') = x, \forall v' \in U \tag{2.4}$$

If $(v_{curr}, v') \notin E$ then

$$dist(v_{curr}, v') = \infty \tag{2.5}$$

Afterwards, v_{curr} is being removed from U, such that

$$U = U \setminus \{v_{curr}\}, \tag{2.6}$$

this is being considered as marking the current node v_{curr} as visited. An important fact is, that visited nodes are never being checked again. At the end of this step, the algorithm checks if the destination node has been marked as visited or if the smallest tentative distance to the nodes in U is infinity, if so, the algorithm stops, either finding the shortest path or no path at all. If the latter is not the case, the node with the smallest tentative distance among the unvisited nodes is taken, marked as the *current node*, so

$$v_{start} = min\{dist(v_{curr}, v')\}, \tag{2.7}$$

where $v' \in U_{new}$, and all steps, beginning at step 2.3, are repeated with the new *current node* v_{start}.

Because of step 2.7, where the minimum distance is taken, the algorithm is characterized as being *greedy*.

The main difference between the Dijkstra algorithm and the Bellman-Ford algorithm is that the former is incapable in handling negative edge weights. An additional feature of the Bellman-Ford algorithm is the recognition of negative graph cycles, which can produce infinitely long paths with increasing negative weights, so that a solution can never be computed, more specifically it doesn't exist in that case.

The application of the Dijkstra algorithm follows in section 3.2, when we are going to build biological networks.

2.4 T-Statistics

The *Student's t-distribution* [70], or just t-distribution, is a continous probability distribution for estimating the mean of normally distributed populations where sample sizes are small and population standard deviations are unknown.

For N independent measurements x_i and $\overline{x}$ as the sample mean, μ as the population mean and s as the population standard deviation estimator, let

$$t = \frac{\overline{x} - \mu}{s/\sqrt{N}}, \tag{2.8}$$

where s^2 is defined by

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2, \tag{2.9}$$

then the t-distribution gives the distribution of the random variable t, with following density function

$$f(t) = \frac{(\frac{r}{r+r^2})^{(1+r)/2}}{\sqrt{r}B(\frac{1}{2}r, \frac{1}{2})}, \tag{2.10}$$

where $r = n - 1$ is the number of degrees of freedom and $B(a, b)$ the beta function.

For increasing N, the t-distribution approaches the normal distribution, moreover it becomes the normal distribution for parameters $\sigma = s$ and $t = z$, where z is the Student's z-distribution, with

$$z = \frac{\overline{x} - \mu}{s} \tag{2.11}$$

The *t-test* [3,49] is based on the t-distribution and examines, if the means of two groups systematically differ from another. So, it provides a decision support if the found mean difference for two given groups has arisen by chance or if it is really based on significant differences of the two groups. There exist three types of t-tests, the one-sample t-test, the t-test for two unpaired samples and the t-test for two independent samples. A one sample t-test tests the mean of a sample against a set value, whereas a paired t-test compares the mean values of two samples, where the two samples derived from the same entity. For instance the comparison before and after a medical treatment for the same patients. The t-test for two independet samples, so the converse of the paired t-test, compares for instance the means in two different therapy groups. In order to apply the t-test, two conditions must be met, firstly, the sample values need to be drawn from a normally distributed population and secondly, in case of independent samples, their

variences need to be identical. In case of two samples having possibly unequal variances, a *Welch t-test* [3] is appropriate.

In specific, a t-test calculates the difference between the observations and the mean and standard error of these differences. Then, dividing the mean by the standard error of the means gives rise to a test statistic $\mathbf{t}$, that is t-distributed with degrees of freedom equal to one less than the number of pairs. Once the t-statistic has been computed, it can be used together with the t-distribution to test the null hypothesis that the two population means are equal or the null hypotheses that one of the population means is greater than or equal to the other. In particular, it will yield a p-value, giving us the probability of obtaining such a test statisitc $\mathbf{t}$ under the assumption that the null hypothesis is true.

In section 3.3 we will see an application of the paired t-test.

2.5 Former Work

This work mainly follows up prior analysis on gaining profound insight into the pathogenesis of psoriasis and eczema, done by Quaranta et. al [22]. In the following we will give a short overview of this study.

2.5.1 Materials & Methods

The first study cohort (n=20) consisted of patients with co-existing plaque-type psoriasis and eczema (atopic eczema, n=3; nummular or dishydrotic eczema, n=7). All patients were caucasians, whereof 35 % were male, 40 % were smokers and the mean age and mean body-mass index were 47 +/- 12 and 24.4 +/- 5.2, respectively. Immune-efficient medications prior to material sampling with a wash-out phase of 6 weeks for systemic and 2 weeks for topical treatments were exclusion criteria.

Afterwards, $6mm$ skin punch biopsies from eczema lesions, psoriasis plaque and clinically non-involved skin were obtained from all patients after local anesthesia. The acquired skin samples were then divided for histologic evaluation and isolation of total RNA [22].

The obtained microarray data was then processed with R software and the limma package from Bioconductor . Processing includes, background correction (*normexp* method), normalization (*cylicloess* method), exclusion of control probes and low expressed probes, linear model fitting to compute moderated paired t-statistics and computation of log-odds of differential expression by empirical Bayes moderation of standrad errors towards a common value. Subsequently, fold changes for each condition (psoriasis, induced eczema, naturally occuring eczema and all eczema) against the corresponding non-involved skin samples were computed. Selection of significant differentially expressed hits was based on an absolute $\log_2$ fold change larger than 2 and a Benjamin-Hochberg corrected p-value smaller than 0.05. In addition, a principal component analysis (PCA) was conducted on the normalized and averaged gene expression data. Finally the enrichment analysis was performed with the *topGO* package from Bioconductor using Fisher's exact test and the *weight01* method [39]. The input for the enrichment analysis consisted of three distinct groups, each group comprising the significant differentially regulated hit genes being present in either psoriasis or eczema and lastly in both samples [22]. Lastly, a classifier on the PCR data was build up, for that, one half of the eczema and one half of the psoriasis patients were used as training set. A 10-fold cross validation was used to train the classifier, and finally the classifier was tested by predicting the disease class on the remaining data samples [22].

2.5.2 Results

Using a cohort of patients affected by both diseases simultaneously, it could be confirmed that unsupervised clustering of whole genome expression (PCA) resulted in patient-related, rather than disease-related grouping. Intra-individual comparison of the molecular signatures revealed genes and signaling pathways being present in both or just one of the diseases, providig a comprehensive picture of the disease pathogenesis [22].

A disease classifier consisting of 15 genes was build up, which was able to accurately diagnose either psoriasis or eczema in an independent patient cohort. Only one patient was misdiagnosed, which reflects the power of this classifier. As already mentioned, prior to the classifier step, a principal component analysis was performed, resulting in no clustering according to the disease, but rather a clustering of individual patients, when the analysis was repeated on a non-supervised hypothesis. These results suggest, that differences in psorisasis, eczema and non-involved skin are being disguised by inter-individual differences [22].

The second step involved the identification of significantly up- or downregulated genes, as compared to non-involved skin in each patient. To sum up, 85 (66 up-regulated, 19 down-regulated) genes uniqe for psoriatic plaques, 55 (36 up-regulated, 19 down-regulated) genes unique for eczema and 34 (23 up-regulated, 11 down-regulated) genes regulated in both diseases, could be identified being significantly different regulated, as compared to non-involved skin [22].

Categorization of these in total 174 significantly different regulated genes resulted in a grouping consisting of three groups, namely the immune system, epidermal component and metabolism. Mainly the immue system and epidermal components are affected, with a substantial difference, being that eczema is characterized by severe defects in epidermal cornification and barrier function, whereas disturbance of epidermal development and differentiation is observed in psoriasis. Moreover, it is suggested, that the immune system directly regulates both epidermal barrier disruption and regeneration [22].

Concerning the immune system, cytokines belonging to the IL-10 family, like IL-9, IL-20, IL-36A and IL-36G were significantly upregulated in psoriasis, in contrast to eczema, in which the upregulation of IL-6 and IL-13 cytokines was significant. But also non-significant upregulation of diverse Th17 and Th2 associated cytokines were observed in psoriasis and eczema, respectively [22].

With regard to psoriasis, this cytokine network induces epidermal metabolism and proliferation and inhibits its differentiation. This indicates an interplay between the epidermal compartement and the immune system, resulting in a wound-healing reaction.

On the other hand, the cytokine network, that corresponds to eczema, is characterized by the pro-inflammatory marker IL-6. Since patients respond well to therapeutic neutralization of IL-6, it is likely to play a pathogenic role.

In contrast to psoriatic skin, various chemokines were up-regulated in eczematous lesions, among them CCL8, CCL17, CCL18, CCL19, CXCL9 and CXCL11, whereas CXCL1 and CXCL8 were up-regulated in psoriatic plaques. So in general, significant regulation of chemokines was primarily observed in eczematous lesions [22].

Numerous antimicrobial peptides (AMPs) were found to be up-regulated in both diseases as compared to non-involved skin. These include the defensin members DEFB4 and DEFB103B as well as some S100 proteins, namely S100A7A, S100A7, S100A8, S100A9 and S100A12. Moreover, IL-20 induced Kallikrein-related peptidases KLK6, KLK9 and KLK13 were exclusively up-regulated in psoriatic skin which induce AMPs and it can be said that the induction of detected AMPs was much higher in psoriatic than in eczematous

skin [22].

Additional differences were observed with regard to early differentiation markers, namely the small proline-rich protein (SPRR) family and the late cornified envelope family (LCE). The former includes the up-regulation of SPRR1B, SPRR2A, SPRR2B, and SPRR2D, and the latter includes LCE3C, LCE3D and LCE3E. Up-regulation of these were exclusively observed in psoriatic plaques. But also down-regulation of LCE members in both diseases was observed, but to a higher degree in eczema. These are LCE1B, LCE1E, LCE2B and LCE5A [22].

With regards to keratin regulation, high up-regulation of KRT6A, KRT6B, KRT6C, KRT16 and KRT75 was observed in psoriatic skin, whereas down-regulation of KRT2 and KRT77 was observed in both diseases, but to a higher degree in psoriatic skin [22].

Another important fact is, that genes involved in metabolism, especially glucose, lipid and amino acid metabolism, are intensively up-regulated in psoriasis, which might explain the clinical association with metabolic syndrome and cardiovascular diseases [22]. Among them, PLA2G4D, iNOS, ABCG4, AKR1B15, AKR1b10 and Wnt signaling inhibitor sclerostin [22].

Given these 174 genes, being either up- or down-regulated in psoriasis, eczema or both, we want to identify the key players and reduce the gene space. These 174 genes will be our primary input for our classfier, which is introduced in the upcoming section and extended in the next chapter.

2.6 Network Smoothed T-Statistics (stSVM)

As already mentioned in section section 1.2, a recent method for biomarker discoveries is a filter based feature selection approach, integrating network information by smoothing gene-wise t-statistics over the graph structure using a random walk kernel [10]. It has been shown that this appraoch yields high signature stability and allows for clear association to biological knowledge, in comparison to other competing methods, like PAM (1.2), SAM (1.2) and several others.

The following gives a short overview of the functioning and will serve as scaffold for our extension, as pointed out in 1.3.

To begin with, network information derived from the Pathway Commons database (2.1.0.2.1) and as an alternative, from the KEGG 2.1.0.2.2 database. Genes, for which no network information existed, were added as unconnected network nodes.

In order to rank these networks, several different appraoches were developed, among them the *kernelized spatial depth* (KSD) measure. The KSD is the spatial depth on the feature space induced by a positive kernel, so the result gives the depth of every vertex of the graph. Various kernels on graphs exist, among them the *Laplacian Kernel*, which gives rise to a dissimilarity matrix, such that a vertex close to the center in the graph turns into a vertex far from the center, so small KSD values indicate vertex significance. The *Diffusion Laplacian Kernel*, which performs opposite operations of the Laplacian, with high KSD values indicating central vertices in the graph, and lastly the *p-Step Random Walk Kernel*, which we will discuss in detail later. All in all, all aforesaid kernels and some additional others, given in literature [29], are based on *Laplacian*, which contains information of the topological structure of a graph, e.g. the degree of relatedness between two nodes or the centrality of a node, and they all perform equally well, but differ in computation time, difficulty of interpretation and in attractivity for practical purposes.

It turns out, that the p-step random walk kernel is the most interpretable and time saving method, being the reason of choice [29].

First, it follows a short explanation on random walks. A random walk on a graph with p steps on a possibly infinite graph G with root vertex r is a stochastic process with random variables $X_1, X_2, \ldots, X_p$ such that $X_1 = r$ and $X_i + 1$ is a vertex chosen randomly from the neighbors of X_i, where $n = |V|$ and $i \in [1, n]$. Given that, one can compute the probabilty $P_{r,e,p}$ that a random walk on the graph of length p starting at node r ends in vertex e [60].

Now we can specify how network smoothed t-statistics are computed. Given a simple, undirected graph $G = (V, E)$ with adjacency matrix A and let $deg(v)$ denote the degree of vertex v, then the graph Laplacian L is defined as

$$L(u,v) = \begin{cases} deg(v), & u = v \\ -1, & (u,v) \in E \\ 0, & otherwise \end{cases}$$

where $u, v \in V$. A more compact term would be

$$L = D - A, \tag{2.12}$$

where $D = diag(deg(v_1), \ldots, deg(v_n))$, with $diag(M)$ returning the main diagonal ($\searrow$) of matrix M, and A, as already mentioned above, is the adjacency matrix of the graph G. Since only undirected graphs are allowed as input, for directed graphs one has to seperate the two cases, meaning that the Laplacian can only be computed on the resulting graphs of incoming or outgoing edges. A normalized version of the Laplacian matrix is similarly defined by

$$L^{norm}(u,v) = \begin{cases} 1, & u = v \wedge deg(u) \neq 0 \\ -\frac{1}{\sqrt{deg(u)deg(v)}}, & u \neq v \wedge (u,v) \in E \\ 0, & otherwise \end{cases}$$

After having defined the Laplacian, we can now give the formula for the p-step random walk kernel

$$K = (\alpha I - L^{norm})^p = ((\alpha - 1)I + D^{-\frac{1}{2}} A D^{-\frac{1}{2}})^p, \tag{2.13}$$

which gives rise to a matrix comprising the degree of topological relatedness between network nodes. The advantage, in comparison to shortest path distance measures is, that alternative routes between two nodes are taken into account, so that nodes being connected via different paths of the same length, mark a higher similarity.

Now having extracted the network information, the next step is to assess genes differential expression by gene-wise t-tests (2.4), resulting in a t-statistic t_i for network node i. The parameters (paired/unpaired, welch-t-test) of the t-test depend on the available datasets and at this point we are just targeting to give a general overview of the procedure. After summarizing all t_i into a vector $\mathbf{t}$, consider the following score vector

$$\tilde{\mathbf{t}} = \mathbf{t}^T K, \tag{2.14}$$

where $\mathbf{t}^T$ describes the transposition of $\mathbf{t}$. So,

$$\tilde{\mathbf{t}_i} = \sum_{j=1}^{n} t_j K_{ij}, \tag{2.15}$$

where $n = |V|$, hence $\tilde{\mathbf{t}}_i$ is a network smoothed version of t_i. Please note that $\tilde{\mathbf{t}}$ does not follow a t-distribution any more.

Subsequently, a permutation test will be applied to the 10% highest ranked genes according to the network smoothed t-score to obtain p-values for each gene. Restriction to the 10% highest ranked genes was for reasons of computation time. After performing multiple testing correction using the FDR approach, a Support Vector Machine (SVM) will be trained with the significant genes with $FDR < 5\%$.

2.7 Goal

In the preceding sections we gave an overall view on methods, definitions and former work, which build the skeletal structure of our approach and provide us with appropriate tools to formulate our idea. To clearify what exactly is being approached in the upcoming chapter, a short description follows.

In section 2.5 174 genes could be identified to be significantly up- or down-regulated in psoriasis, eczema or both. These genes and their expression measurements are our primary input, with the secondary input being PCR measurements of manually selected genes, which will serve as validation data. Since the aforementioned method, the stSVM, tries to integrate network information, we have to compile a network for our significant genes. This step includes retrieval of protein interaction data, which build the basis for our underlying biological network for our significant genes, which in turn will be used to smooth our gene-wise t-statistics. The compilation of these interaction networks will be automated and be based on interaction data from the STRING database. The validation through the integration of PCR data is mainly a slight modification to the equation in 2.15 and will be explained in 3.3.

3 Application and Results

In this chapter, we will first give an overview on the available data for our study, then show how prior knowledge is being fetched and processed and how information content from two distinct data types are being combined and integrated into the disease classifier. Finally we will evalute the results and compare them to former work, to see, if any improvements could be achieved.

3.1 Preprocessing

This section gives attention to the microarray and PCR data available for our study, visualizes and describes how they were preprocessed for further work.

3.1.1 Microarray-Data

For the in 2.5 mentioned 174 significantly up- or down-regulated genes, microarray measurements of patients (n=2*20) affected by plaque-type psoriasis and atopic eczema exist, but only for 13 patients, measurements of non-involved skin parts were available.

The first step applied was the normalization process. Unfortunately, because of a small sample size of non-involved skin measurements, a clean normalization could not be achieved, so it was decided to estimate the missing values through the mean of existing values. To summarize this, let D be a nxm matrix capturing for n patients m gene expression measurements for either psoriasis or eczema affected patients and let N be a kxm matrix comprising the gene expression measurements of non involved skin parts, where the first k patients of the n patients were the same, in particular $k < n$, then the normalisation of D was done via

$$D^{norm}(i,j) = \begin{cases} D(i,j) - N(i,j), & n_i = k_i \\ D(i,j) - \frac{1}{k}\sum_{l=1}^{k} N(l,j), & otherwise \end{cases}$$

where n_i and k_i are the i-th patient of D or N, respectively.

After having normalised the data, we must check if the samples are normally distributed, in order to fulfill the criteria for a t-test, which will be performed in section 3.3. For this, for each gene in the normalised dataset D^{norm} of each diesease a Shapiro-Wilk test of normality was performed. Since we analyse a large number of genes the α error (Type I error: number of false positives) holds only for a single test, hence it is too large and correction for this is required. There exist several methods with distinct properties, like the *Bonferroni* [67] method, which is the most conservative approach, lowering the α value by dividing it by the amount of comparisons or Bonferroni derivatives like Holm [5], Hochberg [5], or Hommel [5]. A Mathematical breakdown proofs that the power of the aforementioned methods increases in the given order [5]. Another powerfull correction method is the so called *false discovery rate* estimation [62, 72] approach, which gives the quantity of the expected proportions of false positive findings amongst the rejected

hypotheses and is biologically motivated, being the reason of our choice. The widely common used shapiro test was used to test for normality and the results suggest that our genes are normally distributed. Exclusion criteria was a p-value of 0.05. Finally both datasets were standardized, for that, each obersavation was replaced by its corresponding *z-score* [71]. The z-score of an observation x_i of a random variable X is defined as

$$z_i = \frac{x_i - \overline{x}}{\sigma}, \tag{3.1}$$

where $\overline{x}$ is the mean and σ the standard deviation of the random variable X. Thus, the z-score gives us the number of standard deviations an observation is either below (negative values) or above (positive values) the mean. This conversion does not change the information content and was performed for practical purposes.

3.1.2 PCR-Data

For a subset of genes (n=28), manually selected from the set of 174 significantly up- or down-regulated genes, normalised PCR measurements for n=10 patients were obtained, for both, psoriasis and eczema. The data was already normalised, so test for normality could be done subsequently. Unfortunately, the results showed that approximately 30% and 52% of the genes in psoriasis and eczema,respectively, are non-normally distributed. In such cases, a log-transformation [48] is appropriate and may lead to better results. Data transformation is common in statistics, but one has to choose the right transformation, such as the square-root transformation for count data or the log transformation for size data [48]. Many variables in biology have a log-normal-distribution, which results from the product of many independent factors which are subject to the biological system [48], hence both datasets were log-transformed (here: natural logarithm), meaning that each measurement was replaced by its natural logarithm. Subsequent test for normality, as already described in 3.1.1, yielded normality for both datasets. Lastly, both datasets were standardized through the z-score either (see 3.1.1).

3.2 Prior Knowledge

In our study, prior knowledge refers to genes relational information gained on the basis of biological interaction networks. This network information is required for the existing method described in 2.6, thus has to be fetched from appropriate source databases, like the ones mentioned in section 2.1. We decided to build our networks based on interaction data available in the STRING database (2.1.0.2.4). Of course, any other choice would be equally appropriate, but the possibility to select between many different evidences and its application programming interface (API) are quite attractive. We first concentrate on the retrievel of the data and than go on to how it is used to build networks.

3.2.1 Interaction-Data Retrieval

Although STRING provides us with a web interface where we can enter gene names and query an interaction network, and even extend it by additional nodes, in oder to get a more comprehensive one, it does not support automated *network* retrievel for a list of genes perse, but provides us with pairwise interaction data for genes available in the gene pool. This means, we have to compile the networks from scratch, based on these pairwise

interaction data, but first we have to construct an interface to STRING to actually get these interaction data.

STRING's application programming interface was used for this purpose. This step mainly requires to build an appropriate uniform resource identifier (URI) to call the API. The structure of the URI is as follows, where values in square brackets can vary:

http://[database]/api/[format]/[request]?[parameter]=[value]

In the following we will just concentrate on the important variables, so an extensive description of the API is not targeted, but can be found in [19].

Possible values for *database* are $string - db.org$, *string.embl.de* and *stitch.embl.de*, whereof the first is the main entry point of STRING, the second the alternative entry point of STRING and the third the sister database of STRING. Our STRING interface allows to select any of these entry points, because of potential server down-times.

As $format$, one can select between different formats like the PSI-MI 2.5 XML format, as already mentioned in section 2.1, Tab-Seperated-Value format (TSV) or even a jpg-image. Our choice is being restricted to PSI-MI 2.5 XML and a flattened version of the PSI-MI 2.5 XML format, due to the forth parameter *request*. So we decided to use the flattened PSI-MI 2.5 XML format, which is in fact like the original PSI-MI 2.5 XML format, but easier to parse. The corresponding value is $psi - mi - tab$.

The forth parameter *request* can take several values, such as *resolve* to get information about a single gene, like its identifer, taxonomic classification and description. The value *interactors* lists all interaction partners for the query item. Lastly, *interactions*, which we used for *request* and which requests all interactions with all possible evidences for a specific gene input.

Additional $[paramteter] = [value]$ pairs, seperated by &, can be added to further specify the request. But since we are using *interactions* for the *request* parameter, additional $[paramteter] = [value]$ pairs are restricted to specify the gene, the required score and to request additional transitive interactions, if there are no more direct connections to the query gene. Corresponding variables are $identifier$,$required_score$ and $additional_network_nodes$. So, for instance, if we set $required_score = 800$ and $additional_network_nodes = 10$ and assume we already requested an interactions list for a gene with gene name *genename*, so $identifier = genename$, which resulted in an initial list of five interactions with minimum score of 800, then the next query will try to find 10 additional best scoring interactions which met the score condition and which are not neccessarily directly connected to the query gene but rather transitively, so one can find a path starting at the query gene, trailing through several interaction partners and lastly ending in that added node. To sum up, the extension of interaction partners always results in a connected network and ends if there are no more interaction partners or if the condition of the required score can not be met. Please note, that values for the parameter $required_score$ are in the range of $[0; 1000]$.

A typical URI for a gene with gene name *genename* would be

http://string-db.org/api/psi-mi-tab/interactions?identifier=genename&
required_ score=800& additional_ network_ nodes=10

By automatically incrementing the *additional_network_nodes* parameter we can exceedingly extend the genes corresponging list of (transitive) interactions partners, with all interactions fulfilling the minimum score condition. Basically, the varying parameters

are $identfier$ and $additional\_network\_nodes$, so the corresponding implementation in R was easily achieved.

Once submitted such an URI, the answer from the database will be a matrix where each row represents a scored connection between two genes, found in the corresponding connected network for the input gene. As alredy pointed out in the introduction of this section, an automated network retrieval service for a *list* of genes is not supported by STRING. So we are not able to query a connected interaction network, where every gene in the input list is present in the network, in particular as a *connected* node. Even if we ease the condtions and abstain from a *connected* network, it will not work. These are the problems we are facing and the basis of the next section, where we exploit the aforementioned options we have at least, to *build* our desired networks.

3.2.2 Network Compilation

To counter those problems mentioned earlier, we present following procedure. The idea is to query each gene in the gene list seperately, get its initial top scoring network, so its best scoring interaction partners, then exceedingly extend its initial network until either no (transitive) interaction partner can be found or no (transitive) interaction partner with the specified connection score can be found. Afterwards, each other gene in the gene list is being searched in this network and pretended to have a connection to the initial query gene if it is present, otherwise not. The according network will be saved for further work, in the positive case.

It is worth to mention, that the actual implementation in R differs, to reduce the broadband and STRING server load. Hence, the initial network is extended stepwise and and in each step the gene-search is repeated, so we give the programm the chance to stop if it already has found the solution, being that he found all genes in the initial network. The according URI parameter $additional\_network\_nodes$ is parameterized in the corresponding R function, thus allowing us to query either fine or coarse grained information. The main difference is that with the coarse grained approach, fetching data is significantly faster, but in general we get additional information which we do not need for further processing, thus longer parsing times, so in general, an increase in computation time. The fine grained approach is exactly the converse, long fetching times, but reduced computation time.

However, to wrap up the first paragraph, let G be the gene pool, our gene query list, with $|G| = n$ and N the retrieved networks, with size $|N| = m$, in the best case $m = n$, meaning that we found for every $g \in G$ a $g' \in G \setminus \{g\}$ with which it interacts, either directly or indirectly (transitive). Furthermore, let E be a list with $|E| = m$ and each list l_i, with $l \in E$, containing the interaction partners for the gene to which network k_i corresponds, where $k \in N$. Then, after some minor formatting work on each $k \in N$, the programm goes on to the next step, the actual network compilation.

At this point, we first concentrate on the possible interaction scores. STRING holds for every interaction nine different scores, if available. The following table which is extracted from [20], gives an overview:

Abbr.	Type	Description
cscore	combined score	see [21]
nscore	neighborhood score	computed from the inter-gene nucleotide count
fscore	fusion score	derived from fused proteins in other species
pscore	cooccurence score of the phyletic profile	derived from similar absence/presence patterns of genes
hscore	homology score	the degree of homology of the interactors trivial and normally not reported in STRING
ascore	coexpression score	derived from similar pattern of mRNA expression measured by DNA arrays and similar technologies
escore	experimental score	derived from experimental data, such as, affinity chromatography
dscore	database score	derived from curated data of various databases
tscore	textmining score	derived from the co-occurrence of gene/protein names in abstracts

Table 3.1: Available interaction evidences (STRING) [20]

The *combined score* combines the probabilites from the other evidences and corrects for the probability of randomly observing an interaction [21]. In the preceding section we introduced the URI parameter *required_score*, which is in fact the lower bound for the combined score. Unfortunately, the API does not support programmatic selection of any of the aforelisted scores, but only to check if the combined score $cs > required_score$. But, the interaction matrices we get, contain the other scores, allowing us to post-process each network $k \in N$, in the sense that we select one of the given evidence types and retain each interaction in network k_i if it provides us with the specified score, or abolish it otherwise. During this process, essential interactions could be removed, so we have to correct the lists of interaction partners in E. Imagine l_i, with $l \in E$, comprises the interaction partners for the query gene $g \in G$ with corresponding network k_i, then *essential* interactions would be connections in k_i, which are compulsive to get from g to any of its interaction partner $p \in l_i$. So, after the removal, the two pretended connected genes, have to be considered to be unconnected, which results in a reformatting of E. We will come to this point again later.

We now reached the point, where we actually try to combine those n networks. For this, we set up a new network without connections but containing each unique interactor in E as network node. If we summarize the uniques into a vector U with $|U| = p$, then we get a pxp adjacency matrix A, representing the interaction network for the initial gene query list, where $A(i, j) = 0$ at the begnning, saying that network node i has a connection with network node j with a confidence score of 0. The next step is to fill the matrix with correct values. This is done by solving the *single-source shortest path problem* for each network $k \in N$, with the source node s being the corresponding query gene for network k_i. The shortest-path problem was solved with the R package *igraph*, providing the corresponding function *get.shortest.paths*. The Dijkstra algorithm was selected instead of the Bellman-Ford algorithm, due to improved computation time and the absence of negative evidence scores. Now, for each shortest path, from the source vertex v to destination vertex v', the minimum confidence score s that is present along that actual path is taken and $A(v, v')$

is set to s if $A(v, v') > s$, otherwise $A(v, v')$ is retained. We took the minimum to be on the safer side. These steps are being repeated for each network $n \in N$. Please note that the Dijkstra algorithm tries to minimize path weights, so it always selects shortest paths which have minimum edge weights amongst all other shortest paths, so it fits perfectly for our purposes. The test $A(v, v') > s$ is necessary, since more than one shortest path could exist to destination vertex d.

One thing is left to consider. In the second to last paragraph we mentioned that a reformatting of E is neccessary due to potential removals of essential interactions, but at this point we could not know if we have removed an interaction contained in a shortest-path, which we did not calculated so far and potentially had selected it. We just knew that if we remove an interaction and if one of these interaction partners is present in l_i for network n_i, then removal from l_i is required. So, apparently we have to correct for this. This is done by calculating for each node in A its degree and check if it is greater then 0, if not, then it has no interaction partner, and is removed from A.

After this step, we get an adjacency matrix A which represents the underlying biological network with real evidence score for our input genes. For further processing real edge weights were not required, so A was discretized, entries bigger than 0 were replaced by 1.

An aditional implemented feature is the possibility to select if intermediate actors on the (shortest) path should be included in the final network. This is done by extending A with each intermediate interaction, so intermediates $I_1,\ldots,I_n$ are added as network nodes and $A(I_j, I_{j+1})$ is set to the corresponding interaction score, where n is the number of intermediates in the actual path and $j \in [1; n]$.

Following figure gives an impression of how a graph would look like. This graph comprises only experimentally approved interactions with a minimum confidence score of 0.8. Main gene input was our set of 174 significantly different regulated genes. Intermediate interactors were included.

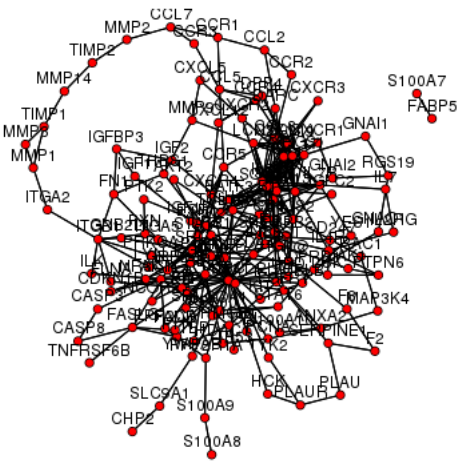


Figure 3.1: Sample network image with experimentally approved interactions with confidence >= 0.8

We finally build our biological network retrieval system and examine in the upcoming

section the integration of the PCR data.

3.3 PCR-Data Integration

In this section we will concentrate on the integration of the PCR data into the disease classfier, namely the stSVM (see 2.6). First it follows the main idea behind this step, how and why this integration is targeted and finally the actual implementation and mathematical presentation.

3.3.1 Method

The great advantage of the PCR technique in comparison to microarrays is its high sensitivity [56], providing highly accurate measurements and enabling us to set up robust statistics. PCR data is necessarily used for validation of microarray data due to its predominance, as it is in our study.

The keynote is to correct the genes t-statistic calculated in equation 2.15 for which PCR measurements exist. This is done by taking the mean of the t-statistics of the microarray and PCR data.

Let S_1 be a sample of a genes differential expression, measured with the microarray technology and and let S_2 be the same measurement, but obtained through PCR techniques. Furthermore, let P_1 and P_2 be the corresponding unknown populations of S_1 and S_2, with true but unknown mean μ_1 and μ_2, respectively. Additionally, let $\overline{x}_1$ and $\overline{x}_2$ be the means of S_1 and S_2, so the estimates of μ_1 and μ_2. Following figure should give an approriate illustration. Please note that no further information, in particular relational information, about $\overline{x}_1$, $\overline{x}_2$, μ_1 and μ_2 exist so far, so we do not know at this point if $\overline{x}_1 > \overline{x}_2$ or vice versa, or if $\overline{x}_1 > \mu_1$ etc. , hence the placements in the following figure are arbitrary.

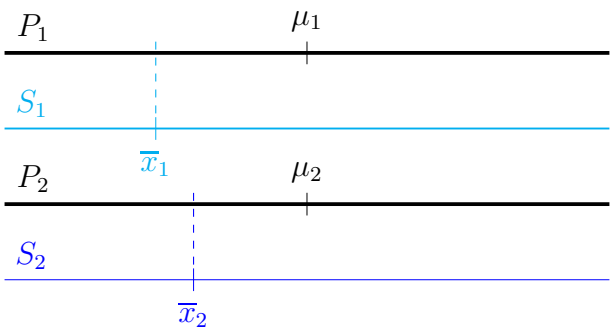


Figure 3.2: Initial sample relation to true population

The samples S_1 and S_2 were measured with different techniques, but they underly the same population. So we summarize P_1 and P_2 into P^* and get following figure. Again, relational information on the variables is not given so far, hence arbitrary positioning.

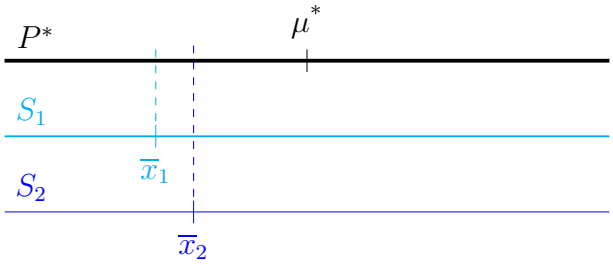


Figure 3.3: Relatedness of measurements between microarray and PCR data

Now, we assume that S_2 is the more reliable dataset, providing us with a better estimate of μ^*, through more accurate expression measurements which form $\overline{x}_1$. We justify this assumption with the PCR techniques predominance. Additionally, assume that we have assessed each genes differential expression via a *paired welch t-test* for both sample sets. Because standard deviations are non-equal and because both clinical cases, psoriasis and eczema, were measured for the same patient, a *paired welch* t-test is required, instead of the standard t-test. Each genes t-statistic tells us on the basis of the sample means if and by how far the true means of the corresponding populations differ from each other. This is done for each dataset seperately, so that we get n t-statistics based on the microarray data and m t-statistics based on the PCR data, where n and m are the amounts of genes for which measurements exist for each technique, in particular $n > m$, so not every gene measured in the microarray process is also remeasured through PCR. If we introduce the distance in terms of t-statistics t_1 and t_2 for the samples S_1 and S_2, respectively, we get following situation:

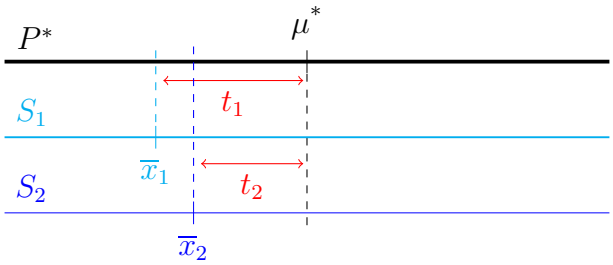


Figure 3.4: Samples t-statistics in comparison

Now, we combine both t-statistics t_1 and t_2 into t^* with $t^* = \frac{t_1+t_2}{2}$ which results in a shift of $\overline{x}_1$ in the direction of μ^*. Finally, let $\overline{x}^*$ be the corrected version of $\overline{x}_1$ induced by t^*, and S^* the resulting combined virtual sample space, then we come to an end with:

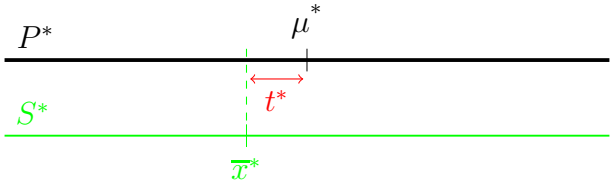


Figure 3.5: Corrected t-statistic

So at this point we strongly believe that the PCR's t-statistic must be better, thus telling us that our estimate $\overline{x}_2$ is near the truth, that is μ^*.

After having discussed the actual idea behind the integration, we will now give the mathemathical description in the upcoming section.

3.3.2 Implementation

In section 2.6 we computed the network smoothed t-scores for each gene. Now we are going to compute the network smoothed corrected/validated t-score for each gene, by introducing the PCR's t-statistic.

Imagine we assessed genes differential expression, as described in the preceding section, so in particular for both datasets, which results in two vectors t^{pcr} and t^{ma} comprising the t-statistics for each gene available for each method, respectively. Remember, that $|t^{pcr}| < |t^{ma}|$, because not every gene existent in the microarray dataset is validated through PCR. However, we now combine both vectors into a vector $\tilde{t}$ according to

$$\tilde{t}_k = \begin{cases} \frac{t_i^{pcr} + t_j^{ma}}{2}, & g_{pcr} = g_{ma} \\ t_j^{ma}, & otherwise \end{cases}$$

where g_{pcr} and g_{ma} are the genes for which a t-statistic t_i^{pcr} and t_j^{ma} exist, respectively. So the positive case, $g_1 = g_2$, comes into effect if there exist t-statistics t^{pcr} and t^{ma} for the same gene. If there is no PCR data with which we can validate the microarray data, then we take the uncorrected t-statistic obtained on the basis of the microarray data, which corresponds to the second case in the above definition. With that, the decisive equation 2.15 alters, in the sense that

$$\tilde{t}_k = \begin{cases} \sum_{j=1}^{n} \frac{t_j^{ma} + t_l^{pcr}}{2} K_{ij}, & g_{pcr} = g_{ma} \\ \sum_{j=1}^{n} t_j K_{ij}, & otherwise \end{cases}$$

where l is the index of the t-statistic of gene g_{pcr} in t^{pcr}.

Finally we have come to an end with descriptions, definitions and explanations and are going to actually test our approach and present our results, which is the topic of the ongoing sections.

3.4 Evaluation & Results

We finally come to the part, where we actually test our approach. We first compiled eight networks based on either experimentally approved interaction data or on all evidences, which are given in table 3.1. Following table gives some properties of these networks as defined in 2.2.

Evidence	Confidence	#Nodes	#Edges	Diameter	Density	$C_{average}$
cscore	0.8	93	1370	4	0.32	0.78
escore	0.8	28	207	3	0.54	0.77
cscore	0.85	79	928	3	0.30	0.76
escore	0.85	15	34	2	0.32	0.54
cscore	0.9	58	498	3	0.30	0.75
escore	0.9	12	15	2	0.22	0.45
cscore	0.95	37	135	3	0.20	0.54
escore	0.95	6	3	1	0.2	0.0

Table 3.2: Compiled networks with different properties

As expected, with increasing confidence bound, available interaction data decreases, resulting in networks of small size, especially for experimentally approved networks, being the reason why we proceed with the non-experimentally approved networks. To remember, we had 174 genes, so none of these networks fully cover our gene space. The overall conclusion we can draw is, that our underlying biological network for our genes is sparsely connected and has a tendancy to form clusters because of the high clustering coefficients $C_{average}$. In addition, a low diameter means that with just a few steps, at most with $diameter = x$ steps, we can reach every node in the graph regardless of where we start and a low density tells us that the network is sparsely connected, so that we reach crabwise any node. This is somehow contradictory and if we take the diameter and density together into account, we can assume that several key nodes, so called hubs, exist. For further processing we decided to train our classifier on the network given in the first row of the above table.

We targeted a classification, where 25% of the datasets was used for testing and 75% for prediction. The classification was started with different SVM parameters a and p, whereof a is the constant value of the random walk kernel and p the random walk steps of the random walk kernel, where $a \in \{3, 5, 10, 15\}$ and $p \in \{1, 2, 3, 4\}$. The cut-off p-value for the permutation test was $p = 0.01$, with $aa = 1000$ permutations test steps. Let *run* describe a classification with a concrete combination of the aforementioned parameters, then each run was 10 times repeated.

We then calculated for each run and each gene, which was selected as a signature in that run, the relative frequency of selection over all repeats. We pre-selected the 10% most frequent genes for each run, to exclude non-significant signature results. Table .1 gives an overview of the frequencies for the modified and unmodified version of the stSVM for the pre-selected genes over all runs. To sum up, our method selected eight more signatures, hence is less robust, in terms of signature selection stability, as compared to the stSVM, and some of them were equally often selected, which compensates the lack of robustness. Subsequently, the top two genes were selected and represent our final signatures for a classification.

Our classifier revealed *NOS2* and *KLK13* as gene signatures. NOS2 acts as a biologic mediater in several processes, like neurotransmission and antimicrobial and antitumoral activities [41], whereas KLK13 belongs to the group of kallikreins, which are a subgroup of serine proteases with diverse physiological functions [40]. It is strongly believed that many kallikreins are associated with carcinogenesis [40]. Given these two genes, we started a functional annotation with the annotation tool from the DAVID website [39]. We focused only on biological pathways from Gene Ontology and it revealed that KLK13 can

be categorized into *proteolysis*, *protein processing* and *protein maturation*, whereas NOS2 matched many diverse biological processes, among them *blood vessel development*, *cell proliferation*, *positive regulation of immune response*, *defense response*, *innate immune response*, *positive regulation of cell killing*, *blood vessel morphogenesis* and *tissue remodeling*. Please note that this list is manually selected and there exist far more categorizations, but this list is adjusted to our problem statement. The table .2 gives the full list.

It also revealed the presence in two cancer related pathways (KEGG pathways) .1 and .2, where NOS2 *directly* contributes to *sustained angiogenesis*. Angiogenesis is the process through which new blood vessels form and can be observed during growth and development as well as in wound healing. Especially tumors need blood vessels to grow and spread, so inhibition of angiogenesis can be used to stop or slow growth or spread of tumors [61]. NOS2 also plays a major role in the activation of monocytes, as a BBID pathway shows .3. This might be the reason, why NOS2 is matched to the GO term *innate immune system*, since monocytes main function is the defense of foreign structures and the activation of the innate immune system [6]. Moreover, they play an important role in the inflammatory response [6].

In the following we will discuss these results.

3.5 Discussion

In 2.5.2 we already mentioned that numerous antimicrobial peptides were upregulated in both diseases, in particular in psoriasis. Moreover, NOS2 was *highly* upregulated in psoriasis and since NOS2 acts as mediater for antimicrobial and antitumoral activities as pointed out in the last section, it may explain the latter and in particular the functional annotations to *positive regulation of immune response*, *defense response* and *immune response*, which in turn might explain its angiogenic behaviour.

One of KLK13's main functions is the hydrolysis of proteins into smaller polypeptides and/or amino acids by cleavage of their peptide bonds, but it is also implicated in protein maturation. Protein maturation is the process leading to the attainment of the full functional capacity of a protein [42]. A study revealed that kallikreins are essential for steady desquamation and skin barrier function [53].

If we now take both signatures together into account, the positive regulation of the immune response might be supported by KLK13, or in general by kallikreins, in the sense that the body tries to remodel involved skin parts by first denaturating existing proteins (proteolysis-KLK13; positive regulation of cell killing-NOS2), which leads to desquamation. Subsequent renewal of skin (tissue remodeling-NOS2) might be positively regulated by NOS2, in terms of blodd vessel development and cell proliferation (angiogenesis). This processes might in turn be supported by KLK13 through protein maturation.

3.6 Comparison

Lastly we want to compare the original stSVM to our modified version presented in our study, which we name *vstSVM* from now on. For this, we calculated the area under receiver operator characteristic curve (AUC) with the R-package *ROCR* [63], to compare the prediction power. The AUC values were taken from all classifications with all combinations of parameters as described in 3.4. The following boxplot illustrates the results:

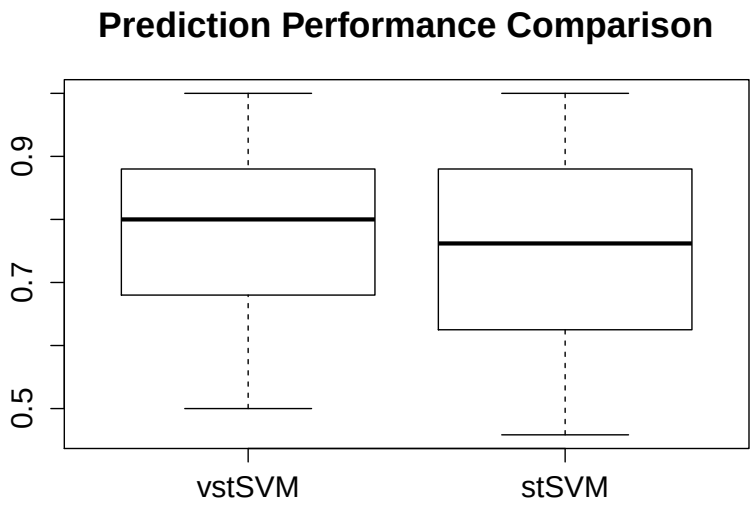


Figure 3.6: Prediction performance comparison of the modified vs unmodified version over 10 repeats of the cross-validation step

As we can see, the median of the vstSVM is higher than compared to the original stSVM which proves the success of our method. In order to assess the degree of improvement, an unpaired t-test was applied, to test wether the mean AUC values significantly differ from each other. This revealed a p-value of 0.039, although the p-value suggests a significant improvement with an *alpha level* at 0.05, we must consider that the result is quite near the rejection border. Moreover, our approach yields less variation in prediction performance, as we can deduce from the higher compression of the interquartile range and also the lower whisker incorporate higher AUC values as compared to the stSVM. Additionally, we would like to mention that a common problem is that inclusion or exclusion of a few patients can lead to quite different signature outcomes, so that reproducibility of signatures is difficult, hence leading to different prediction perfomances, but we believe that with increasing classification repetitions or with increasing number of patient, this problem can be countered.

Since our approach yields better results than the original stSVM, we disclaimed further comparison with other related methods, like the ones given in 1.2, for which extensive comparison have been made [9] already. To summarize these, incorporating prior knowledge, in terms of network information, does not significantly improve classification accuracy, but rather interpretability and and stability of the gene signatures, as compared to classical methods(1.2) [9]. This could be approved when starting classification with no network information. We can say that improved gene selection stability does not necessarily concur with improved prediction performance and it is supposed that the reason could be high correlations in gene's differential expression and if those highly correlated genes are itself correlated with the patient group, then selection of any of these genes would lead to similar prediction performance [9]. This statement is supported by our datasets, since our genes derived from former work [22] which already pre-filtered non significant regulated genes and provided us with a set of genes which all were believed to correlate with the two diseases.

Moreover, it is known that performance of machine lerning methods strongly depends

on the available data [9] and since we have unfortunately small sample sizes, in particular PCR data with only 10 measurements per clinical case, it is likely to play a major role in the outcome of prediction performance, so we are convinced that classification on larger datasets reveals better results. Especially, *small* PCR samples sizes may *worsen* the calculated t-statistics t_1 derived from the micorarray data (3.3), by shifting t_1 away from the real population mean μ^*. So we strongly recommend to apply our method only on relatively big PCR sample sizes to counter this problem. The optimal patient size for a PCR dataset would be at least the patient size of the microarray dataset, such that the resulting t-statistics t_1 and t_2 are calculated on equal sample sizes.

One major advantage of network-based methods is the biological interpretability of gene signatures, which often reveal enrichment of disease related genes, known drug targets or KEGG pathways [9], as it is in our case.

All in all, a comparison of 14 network based and non-network based methods with respect to prediction accuracy, biomarker signature stability and biological interpretability, showed that, in general, no algorithm performed best with regard to all three categories [9]. Network-based classifiers greatly inprove gene selection stability and interpretability, but yield poor prediction performance, whereas other methods yield moderate prediciton accuracy, but low stability, or high prediction perfomance but difficult interpretability [9]. So we can not clearly discriminate any of the compared algorithms given in [9] and should concentrate on alternative ways to enhance available methods, for instance, by integrating additional experimental data like miRNA, SNP or CNV data [9].

4 Summary & Outlook

In this study, we targeted the discovery of biomarkers in two common widespread inflammatory skin diseases, namely *psoriasis* and *eczema*. We tied in with former work [22], which revealed a set of significantly up- or down-regulated genes, either in psoriasis, eczema or both. This set consists of 174 genes and was available as raw microarray data. Measurements of non-involved skin were also available and were used for normalization. A subset of genes of these significantly different regulated genes were remeasured through PCR for validation of the microarray data. The PCR data was already fully normalized. A most recent method called *stSVM* served as scaffold for our approach. This method is able to integrate expression data and network information to train a classifier. It basically ranks networks through a kernelized spatial depth measure, induced by a kernel, namely the p-step random walk Laplacian kernel. Our main goal was to improve this method by integrating PCR data for gene's expression validation. Moreover, an automated network retrieval system was targeted.

We build an interface to the STRING database, in order to get protein interaction data, which in turn was used to compile our networks. This system allows the user to manually select the sources of the interaction data, like experiments or coexpression data only, which yields highly comprehensive and reliable networks. The network compilation step was mainly based on finding pairwise shortest-paths for our genes. The validation through PCR was mainly done by correcting gene's differential expression on basis of t-statistics. For both datasets, the microarray and PCR data, gene's differential expression was assessed via paired t-tests. The gene's t-statistic, calculated on the basis of the microarray dataset, were corrected through the corresponding t-statistic derived from the PCR data, if such PCR data existed for the particular gene. The correction was done by taking the mean of both t-statistics. This correction was done under the assumption that PCR measurements are more reliable as compared to microarray measurements, because of its predominance in terms of sensitivity. Hence, the resulting t-statistics, based on the PCR data, should reflect genes differential expression more accurate.

Our approach yielded better results than the original unmodified method. The results were cross validated, in particular both datasets. For that we excluded a 25% of measurements to test the classifier and the remaining 75% to train the classfier. Our classification results are consistent with those described in section 2.5.2 and are assumed to be the one of the key players to distinguish psoriasis and eczema. A comparison to the unmodified version of the classifier (stSVM) via an unpaired t-test showed a significant improvement with a p-value of 0.039 at a significance level $alpha = 0.05$, but we have to consider that it is actually quite near the rejection border.

To summarize the advantages and disadvantages, network-based classifiers do not neccessarily improve prediction performance but rather biological interpretability and signature selection stability, which enables us to link signatures to diseases easier. This is supported by the fact that network ranking with network kernels, ranks central nodes (hubs) much higher, which in turn are often well studied and directly known to be disease related. A major problem is, that if highly correlated genes are itself correlated with

the patient group, then selection of any of these genes would lead to similar prediction performances. This was indeed the case in our study, since our classification was done on a set of genes which was believed to correlate with the aforementioned diseases. Another problem is, that inclusion or exclusion of patients often leads to quite different signature outcomes, which alters the prediction performance. So, signature reproducability is another problem we encounter in SVMs. In particular, our method is prone to this, because of small PCR patient sample sizes, which may lead to a *worse* t-statistic, if inappropriate patient combinations were included to calculate the statistics. Unfortunately, this would worsen rather than improve the calculated t-statistics derived from the microarray dataset. So, it is highly recommended to apply our method cautious, in the sense that sufficiently big PCR sample sizes must be at hand. We additionally recommend to classify with high repetitions, in order to counter the problem of the selection of patients.

In order to improve our approach, one might consider extending the network retrieval system by the component to compile *directed* networks. This would be a major improvement, since the relational information content would increase. This extension would imply an alteration to the equations given in section 2.6, in the sense that the kernelized spatial depth measure induced by the Laplacian would be applied to the corresponding graphs with incoming or outgoing edges. Another improvement would be the possibility to select *combinations* of given evidences, which would provide us with more comprehensive and reliable networks. This is due to insufficent experimental evidences, so one could select a subset of the next most reliable evidences, to avoid the lack of experimentally approved interactions. Although STRING provides us with sufficient protein interaction data with multiple evidences, its limitation to these kind of data may be insufficient for further ambitions. Moreover, its application programming interface was indeed satisfactory, but may not compete with other source database's tools to fetch data. So, it is recommended to build further interfaces to source databases, which provide more data and in a more flexible manner, hence we introduced EMBL in section 2.1, which would be a preferable alternative. Since we just achieved a marginal enhancement by the integration of PCR data, the inclusion of further experimental data like RNASeq, SNP or CNV data would be the next step.

With this work, we hope to alleviate further efforts of biomarker discovery and data combination and integration.

Acknowledgements

Special thanks goes to my advisor Bettina Knapp, who was a great and affectionate supervisor. I also would like to thank the whole ICB group for the remarkably pleasant working athmosphere and seminars. Additional special thanks goes to my friends and family who supported me during my work.

Appendices

Gene	modified	unmodified
NOS2	14.678899	21.621622
KLK13	10.091743	-
TREX2	10.091743	-
GJB6	9.174312	18.918919
IL8	9.174312	24.324324
PLAT	9.174312	-
PLA2G4D	5.504587	18.918919
IL6	4.587156	18.918919
MMP1	4.587156	-
TNC	4.587156	8.108108
LTF	3.669725	-
RHCG	2.752294	-
SOCS3	2.752294	-
IL36G	1.834862	-
LCN2	1.834862	

Table .1: Signature frequencies for the modified and unmodifed version of the stSVM over all runs for a classification with 16 parameter combinations

GO Terms [KLK13]
blood vessel development
response to hypoxia
regulation of leukocyte mediated cytotoxicity
positive regulation of leukocyte mediated cytotoxicity
endothelial cell proliferation
vasculature development
blood vessel remodeling
innate immune response in mucosa
organ or tissue specific immune response
mucosal immune response
positive regulation of immune system process
regulation of immune effector process
positive regulation of immune effector process
regulation of peptide secretion, circulatory system process
arginine metabolic process
arginine catabolic process
oxygen and reactive oxygen species metabolic process
superoxide metabolic process, nitric oxide biosynthetic process
defense response, immune response
cell surface receptor linked signal transduction
G-protein coupled receptor protein signaling pathway
G-protein signaling
coupled to cyclic nucleotide second messenger
G-protein signaling
coupled to cGMP nucleotide second messenger
intracellular signaling cascade

protein kinase cascade
blood circulation
regulation of heart contraction
regulation of blood pressure
cell proliferation
cellular amino acid catabolic process
glutamine family amino acid metabolic process
glutamine family amino acid catabolic process
amine catabolic process
response to bacterium
organic acid catabolic process
second-messenger-mediated signaling
cGMP-mediated signaling
cyclic-nucleotide-mediated signaling
intracellular receptor-mediated signaling pathway
regulation of cell killing
positive regulation of cell killing
regulation of cell proliferation
defense response to bacterium
regulation of cellular respiration
regulation of generation of precursor metabolites and energy
regulation of multi-organism process
positive regulation of multi-organism process
regulation of system process
nitrogen compound biosynthetic process
innate immune response
nitric oxide metabolic process
carboxylic acid catabolic process
regulation of hormone secretion
retinoic acid receptor signaling pathway
blood vessel morphogenesis
tissue remodeling
regulation of insulin secretion
defense response to Gram-negative bacterium
regulation of secretion
regulation of killing of cells of another organism
positive regulation of killing of cells of another organism
oxidation reduction
regulation of cellular localization
response to oxygen levels

Table .2: GO Terms for KLK13

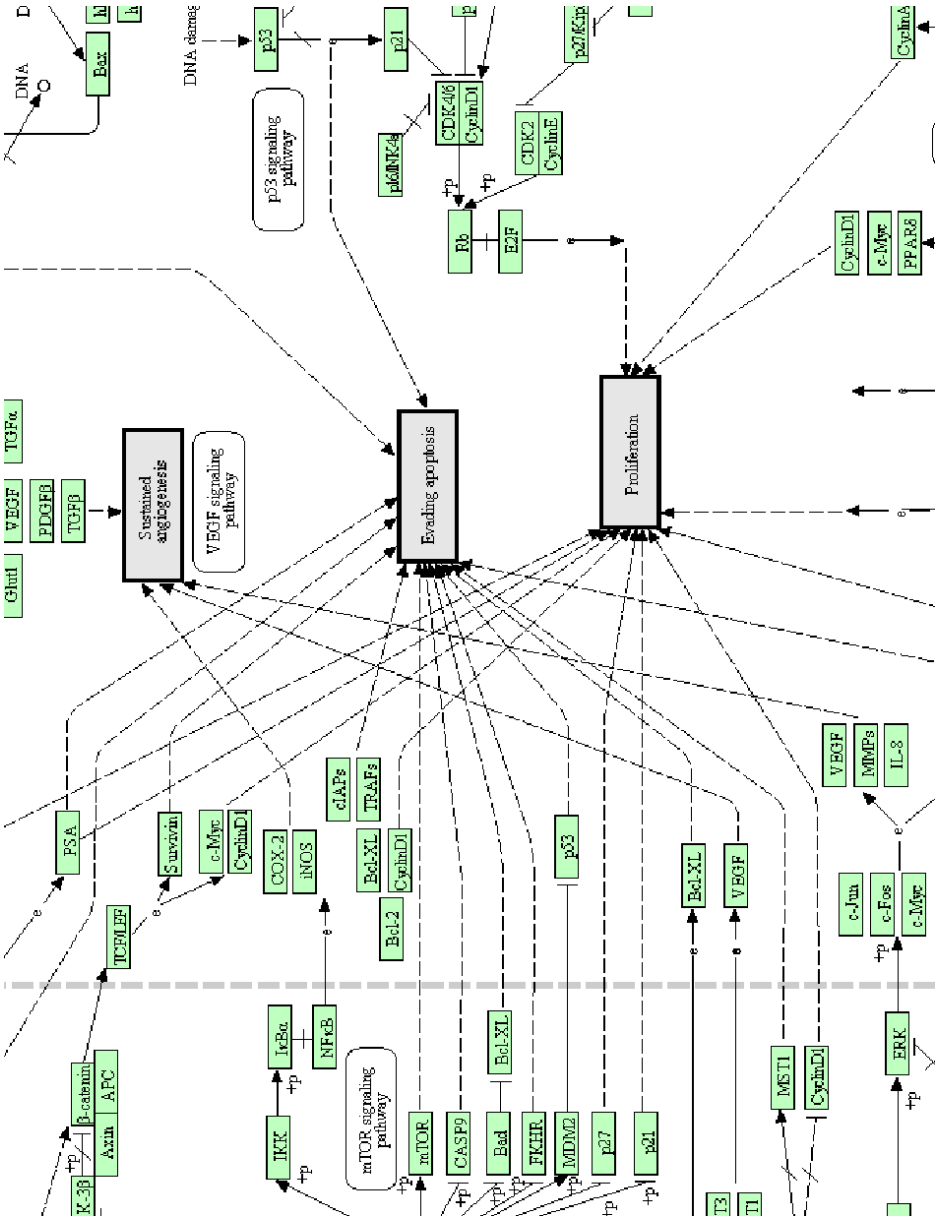


Figure .1: Image section that contains NOS2 (iNOS) in cancer related pathways

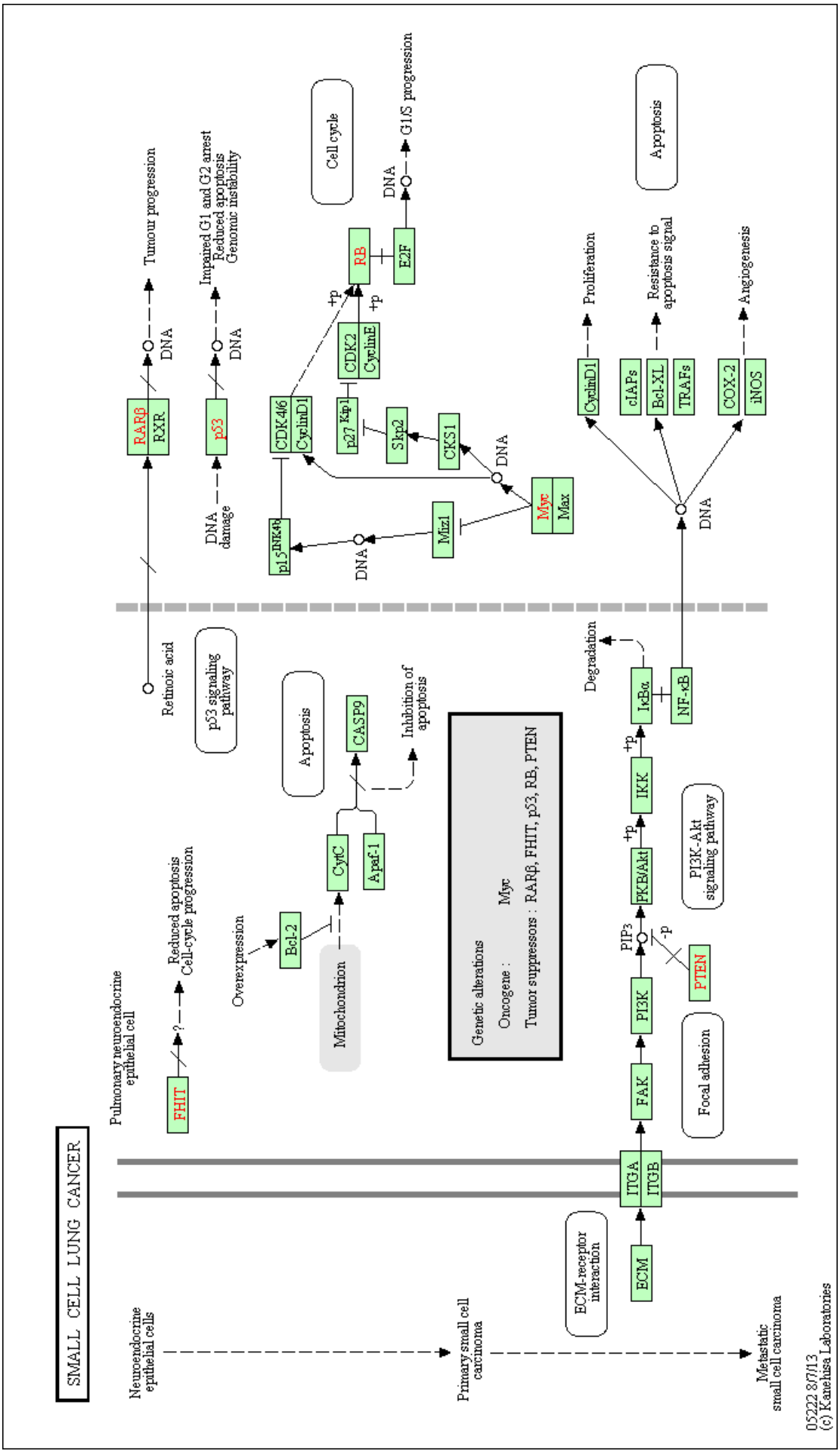
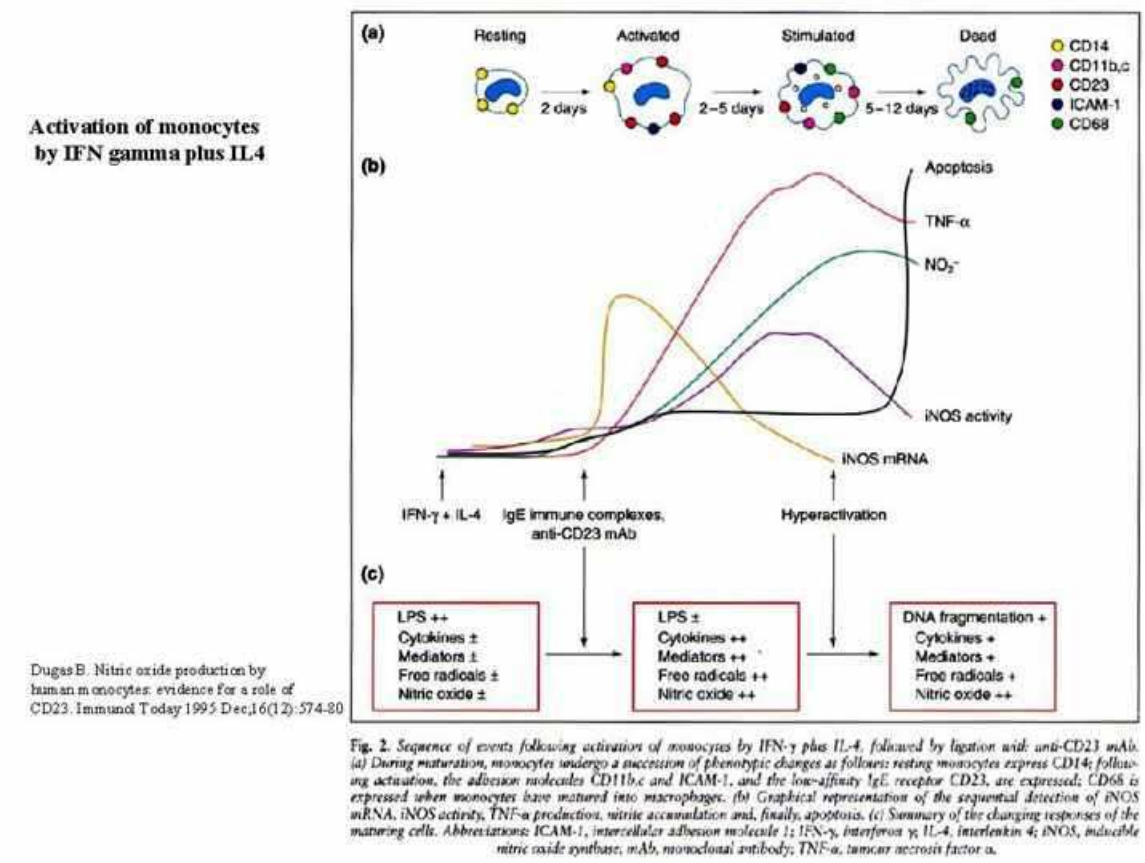


Figure .2: NOS2 acting in small cell lung cancer



Dugas B. Nitric oxide production by human monocytes: evidence for a role of CD23. Immunol Today 1995 Dec;16(12):574-80

Fig. 2. *Sequence of events following activation of monocytes by IFN-γ plus IL-4, followed by ligation with anti-CD23 mAb. (a) During maturation, monocytes undergo a succession of phenotypic changes as follows: resting monocytes express CD14; following activation, the adhesion molecules CD11b,c and ICAM-1, and the low-affinity IgE receptor CD23, are expressed; CD68 is expressed when monocytes have matured into macrophages. (b) Graphical representation of the sequential detection of iNOS mRNA, iNOS activity, TNF-α production, nitrite accumulation and, finally, apoptosis. (c) Summary of the changing responses of the maturing cells. Abbreviations: ICAM-1, intercellular adhesion molecule 1; IFN-γ, interferon γ; IL-4, interleukin 4; iNOS, inducible nitric oxide synthase; mAb, monoclonal antibody; TNF-α, tumour necrosis factor α.*

Figure .3: NOS2 (iNOS) dependent monocyte activation

Bibliography

[1] Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, and Cesareni G. Mint: a molecular interaction database. *FEBS Lett*, 513(1):135–140, 2002.

[2] Mason AR, Mason J, Cork M, Dooley G, and Hancock H. Topical treatments for chronic plaque psoriasis. *Cochrane Database Syst Rev*, 28:3:CD005028, 2013. doi:10.1002/14651858.CD005028.pub3.

[3] Rasch B., Friese M., Hofmann W.J., and Naumann E. *Quantitative Methoden: 1. Einführung in die Statistik für Psychologen und Sozialwissenschaftler*. Springer Verlag, 2010.

[4] Gustavo E. A. P. A. Batista and Maria Carolina Monard. A study of k-nearest neighbour as animputation method. *P. O. Box*, 668:13560–970, 2002.

[5] Richard E. Blakesley, Sati Mazumdar, Mary Amanda Dew, Patricia R. Houck, Gong Tang, Charles F. Reynolds, and Meryl A. Butters. Comparisons of methods for multiple hypothesis testing in neuropsychological research. *Neuropsychology*, 23(2):255–264, 2009.

[6] Auffray C, Sieweke MH, and Geissmann F. Blood monocytes: development, heterogeneity, and relationship with dendritic cells. *Annu Rev Immunol*, 27:669–692, 2009.

[7] Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, and Tyers M. Biogrid: a general repository for interaction datasets. *Nucleic Acids Res*, 34(Database Issue):D535–539, 2006.

[8] Boris V. Cherkassky, Andrew V. Goldberg, and Tomasz Radzik. Shortest path algorithms: Theory and experimental evaluation. *Mathematical Programming*, 73:129–174, 1996. doi:10.1007/BF02592101.

[9] Yupeng Cun and Holger Fröhlich. Prognostic gene signatures for patient stratification in breast cancer - accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions. *BMC Bioinformatics*, 13:69:n.pag., 2012.

[10] Yupeng Cun and Holger Fröhlich. Network and data integration for biomarker signature discovery via network smoothed t-statistics. *PLosS ONE*, 8(9):n.pag., 2013.

[11] Yupeng Cun and Holger Fröhlich. An r package for network-based biomarker discovery, Dez. 2013. R-package. URL: http://cran.fhcrc.org/web/packages/netClass/index.html.

[12] Naiara Abreu de Azevedo Fraga, Maria de Fátima Paim de Oliveira, Ivonise Follador, Bruno de Oliveira Rocha, and Vitória Regina Rêgo. Psoriasis and uveitis: a literature review. *An Bras Dermatol*, 87(6):877–883, 2012.

[13] John de Korte, Mirjam A G Sprangers, Femke M C Mombers, and Jan D Bos. Quality of life in patients with psoriasis: A systematic literature review. *J Investig Dermatol Symp Proc*, 9(2):140–7, 2004.

[14] Wingender E, Dietze P, Karas H, and Knuppel R. Transfac: a database on transcription factors and their dna binding sites. *Nucleic Acids Res*, 24(1):238–241, 1996.

[15] Cerami EG, Gross BE, Demir E, Rodchenkov I, Babur O, and et al. Pathway commons, a web resource for biological pathway data. *Nucleic Acid Res.*, 39(Database Issue):D685–D690, 2008. doi:10.1093/nar/gkq1039.

[16] A.D.A.M Medical Encyclopedia, Nov. 2012.

[17] A.D.A.M Medical Encyclopedia. Psoriasis, Nov. 2012. URL: http://ncbi.nlm.nih.gov/pubmedhealth/PMH0001470/.

[18] Cerami et al. Pathway commons - source databases. URL: http://pathwaycommons.org/pc2/datasources.html.

[19] Jensen et al. String database: Developer-documentation. ch. 6. URL: http://string-db.org/help/index.jsp?topic=/org.string-db.docs/api.html.

[20] Jensen et al. String database: Developer-documentation - frequently asked question (faq). ch. 4. URL: http://string-db.org/help/index.jsp?topic=/org.string-db.docs/api.html.

[21] Jensen et al. Search tool for the retrieval of interacting genes/proteins (string). *Nucleic Acids Res.*, 37(Database issue):D412–416, 2009.

[22] Quaranta et al. Intra-individual genome expression analysis reveals a specific molecular signature of psoriasis and eczema. -, 2013.

[23] Gnad F, Ren S, Cox J, Olsen JV, Macek B, Oroshi M, and Mann M. Phosida (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol*, 8(11):n.pag., 2007.

[24] Diella FCS, Gemnd C, Linding R, Via A, Kuster B, Sicheritz-Pontn T, Blom N, and Gibson TJ. Phospho.elm: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*, 5:n.pag., 2004.

[25] Stoesser G., Sterk P., Tuli M., Stoehr P., and Cameron G. European bioinformatics institute (embl) - information. URL: http://ebi.ac.uk/about/brochures.

[26] Stoesser G., Sterk P., Tuli M., Stoehr P., and Cameron G. European bioinformatics institute (embl) - software. URL: http://ebi.ac.uk/Tools/webservices/.

[27] Stoesser G., Sterk P., Tuli M., Stoehr P., and Cameron G. European bioinformatics institute (embl) - training. URL: http://ebi.ac.uk/training/online/.

[28] Stoesser G., Sterk P., Tuli M., Stoehr P., and Cameron G. The embl nucleotide sequence database. *Nucleic Acids Research*, 25(1):714, 1997. doi:10.1093/nar/25.1.7.

[29] Cuilan Gao1, Xin Dang1, Yixin Chen, and Dawn Wilkins. Graph ranking for exploratory gene data analysis. *BMC Bioinformatics*, 10(Suppl 11):S19:n.pag., 2009. doi:doi:10.1186/1471-2105-10-S11-S19.

[30] I Guyon, J Weston, S Barnhill, and V Vapnik. Sam: Signifance analysis of microarrays. R-package. URL: http://statweb.stanford.edu/~tibs/SAM/Rdist/index.html.

[31] I Guyon, J Weston, S Barnhill, and V Vapnik. Gene selection for cancer classification using support vector machines. *Mach. Learn*, 46:389–422, 2002.

[32] Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, and von Mering C. The hupo psis molecular interaction format - a community standard for the representation of protein interaction data. *Nature Biotechnology*, 22(2):177–83, 2004.

[33] C Hoare, A Po Li Wan, and H Williams. Systematic review of treatments for atopic eczema. *Health Technol Assess*, 4(37):1–191, 2000.

[34] Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, and Eisenberg D. Dip: the database of interacting proteins. *Nucleic Acids Res*, 28(1):289–291, 2000.

[35] Keseler IM, Bonavides-Martinez C, Collado-Vides J, Gama-Castro S, Gunsalus RP, Johnson DA, Krummenacker M, Nolan LM, Paley S, and Paulsen IT. Ecocyc: a comprehensive view of escherichia coli biology. *Nucleic Acids Res*, 37(Database Issue):D464–470, 2009.

[36] Li J and Tibshirani R. Finding consistent patterns: a nonparametric approach for identifying differential expression in rna-seq data. *Stat Methods Med Res.*, 22(5):519–36, 2013. doi:doi: 10.1177/0962280211428386.

[37] Whitaker JW, Letunic I, McConkey GA, and Westhead DR. metatiger: a metabolic evolution resource. *Nucleic Acids Res*, 33(19):6083–6089, 2009.

[38] Ulrich LE and Z IB. Mist: a microbial signal transduction database. *Nucleic Acids Res*, 35(Database issue):D386–390, 2007.

[39] Richard A. Lempicki. David bioinformatics resources 6.7. *Nature Protocols & Nucleic Acid Res*, 4(1) & 37(1):44 & 1, 2009.

[40] Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, He S, Liu C, Shi W, and Bryant SH. Klk13 kallikrein-related peptidase 13 [homo sapiens (human)]. URL: http://www.ncbi.nlm.nih.gov/gene/26085.

[41] Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, He S, Liu C, Shi W, and Bryant SH. Nos2 nitric oxide synthase 2 [homo sapiens (human)]. URL: http://www.ncbi.nlm.nih.gov/gene/4843.

[42] Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, He S, Liu C, Shi W, and Bryant SH. Protein maturation. URL: http://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0051604.

[43] Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, Kitano H, Arkin AP, Bornstein BJ, Bray D, and Cornish-Bowden A. The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003.

[44] Kanehisha M, Araki M, Goto S, Hattori M, and Hirakawa M et al. Kyoto encyclopedia of genes and genomes (kegg) - database statstics. URL: http://kegg.jp/kegg/docs/statistics.html.

[45] Kanehisha M, Araki M, Goto S, Hattori M, and Hirakawa M et al. Kyoto encyclopedia of genes and genomes (kegg) - software. URL: http://kegg.jp/kegg/kegg4.htm.

[46] Kanehisha M, Araki M, Goto S, Hattori M, and Hirakawa M et al. Kegg for linking genome to life and environment. *Nucleic Acid Res.*, 36:D480–D484, 2008.

[47] Krull M, Voss N, Choi C, Pistor S, Potapov A, and Wingender E. Transpath: an integrated database on signal transduction and a tool for array analysis. *Nucleic Acids Res*, 31(1):97–100, 2003.

[48] J.H. McDonald. *Handbook of Biological Statistics (2nd ed.)*. Sparky House Publishing, 2009.

[49] John H. McDonald. *Handbook of Biological Statistics (2nd ed.)*. Sparky House Publishing, 2009.

[50] Huaiyu Mi, Anushya Muruganujan, Emek Demir, Yukiko Matsuoka, Akira Funahashi, Hiroaki Kitano, and Paul D. Thomas. Biopax support in celldesigner. *Bioinformatics*, 27 (24):3437–3438, 2011.

[51] et al Miller ML. Linear motif atlas for phosphorylation-dependent signaling. *Sci Signal*, 1(35):n.pag., 2008.

[52] Wright M Murray-Rust P, Rzepa HS. Development of chemical markup language (cml) as a system for handling complex chemical content. *New Journal of Chemistry*, 25:618–34, 2001.

[53] Komatsu N, Saijoh K, Toyama T, Ohka R, Otsuki N, Hussack G, Takehara K, and Diamandis EP. Multiple tissue kallikrein mrna and protein expression in normal skin and skin diseases. *Br J Dermatol*, 153(2):274–281, 2005.

[54] Georgios A Pavlopoulos, Maria Secrier, Charalampos N Moschopoulos, Theodoros G Soldatos, Sophia Kossida, Jan Aerts, Reinhard Schneider, and Pantelis G Bagos. Using graph theory to analyze biological networks. *BioData Mining*, 4:10:n.pag., 2011. doi:10.1186/1756-0381-4-10.

[55] Karp PD, Ouzounis CA, Moore-Kochlacs C, Goldovsky L, Kaipa P, Ahren D, Tsoka S, Darzentas N, Kunin V, and Lopez-Bigas N. Expansion of the biocyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res*, 33(19):6083–6089, 2005.

[56] George Quellhorst, Ying Han, and Ray Blanchard. Validating microarray data using rt real-time pcr. http://sabiosciences.com. URL: http://sabiosciences.com/manuals/MicroarrayValidation.pdf.

[57] Murray RP and S RH. Chemical markup, xml, and the worldwide web. 1. basic principles. *Chem Inf Comput Sci*, 39:928–42, 1999.

[58] Mouhamadou Lamine Samb, Fode Camara, Samba Ndiaye, Yahya Slimani, and Mohamed Amir Esseghir. A novel rfe-svm-based feature selection approach forclassification. *International Journal of Advanced Science and Technology*, 43:n.pag., 2012.

[59] Sandelin, Alkema W, Engstrm P, Wasserman WW, and Lenhard B. Jaspar: an open-access database for eukaryotic. *Nucleic Acids Res*, 32(Database Issue):D91–94, 2004.

[60] Kyle Siegrist. Random walks on graphs. http://www.math.uah.edu/. URL: http://www.math.uah.edu/stat/markov/WalkGraph.html.

[61] Cancer.gov: National Cancer Institute Web site. Angiogenesis inhibitors, Oct 2011. URL: http://www.cancer.gov/cancertopics/factsheet/Therapy/angiogenesis-inhibitors.

[62] Korbinian Strimmer. A unified approach to false discovery rate estimation. *BMC Bioinformatics*, 9:303, 2008. doi:10.1186/1471-2105-9-303.

[63] Sing T, Sander O, Beerenwinkel N, and Lengauer T. Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):394041, 2005.

[64] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Pam: Prediction analysis for microarrays. R-package. URL: http://statweb.stanford.edu/~tibs/PAM/Rdist/index.html.

[65] Robert Tibshirani, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *PNAS*, 99(10):6567–72, 2002.

[66] Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, and Venugopal A. Human protein reference database. *Nucleic Acids Res*, 37:Database: D767–772, 2009.

[67] Eric Weisstein. Bonferroni correction. From MathWorld–A Wolfram Web Resource. URL: http://mathworld.wolfram.com/BonferroniCorrection.html.

[68] Eric Weisstein. Graph loop. From MathWorld–A Wolfram Web Resource. URL: http://mathworld.wolfram.com/GraphLoop.html.

[69] Eric Weisstein. Simple graph. From MathWorld–A Wolfram Web Resource. URL: http://mathworld.wolfram.com/SimpleGraph.html.

[70] Eric Weisstein. Student's t-distribution. From MathWorld–A Wolfram Web Resource. URL: http://mathworld.wolfram.com/Studentst-Distribution.html.

[71] Eric Weisstein. z-score. From MathWorld–A Wolfram Web Resource. URL: http://mathworld.wolfram.com/z-Score.html.

[72] Daniel Yekutieli Yoav Benjamini. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):116588, 2001.

[73] Ying Yu. Svm-rfe algorithm for gene feature selection. Department of Electrical and Computer Engineering, University of Delaware, Newark. URL: https://www.google.de/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&ved=0CC4QF