universität wien

DIPLOMARBEIT

Titel der Diplomarbeit

Mapping of differential gene expression data on protein networks

angestrebter akademischer Grad

Magistra der Naturwissenschaften (Mag. rer.nat.)

Verfasserin:	Irmgard Mühlberger
Matrikel-Nummer:	0105478
Studienrichtung (lt. Studienblatt):	Molekulare Biologie A490
Betreuer:	Univ.Doz.Dr. Bernd Mayer

Wien, am 6.Juni 2008

DANKSAGUNGEN

Ganz besonders möchte ich meinem Betreuer Bernd Mayer danken, der mich während meiner Diplomarbeit im aller besten Sinne unterstützt und betreut hat.

Prof. Rainer Oberbauer und Dr. Peter Hauser möchte ich für die Bereitstellung der experimentellen Daten danken, sowie Andreas Bernthaler, der mir seine methodischen Ansätze zur Verfügung gestellt hat.

Meine Familie und Freunde haben mich bis hierhin begleitet und meine Erfolge, vorallem aber auch Krisen mit mir getragen. Ohne ihnen wäre diese Diplomarbeit nicht möglich gewesen.

Ein sehr herzliches Dankeschön geht an alle meine Kollegen für das angenehme Arbeitsklima, die Unterstützung und unglaubliche soziale Kompetenz die es mir möglich gemacht hat, auch trotz meiner Behinderung, diese Arbeit so problemlos zu verwirklichen.

TABLE OF CONTENTS

INTRODUCTION

1.1 Omics Technologies

During the last 50 years, the field of molecular biology underwent particular technological advances. This progress started with the discovery of the DNA double helix and a most recent milestone was the completion of sequencing of the human genome.

Nowadays, high-throughput (HT) techniques are fully established in the daily routine in the labs, enabling the study of thousands of genes or proteins simultaneously. The most popular techniques are microarrays but also other procedures like 2D gel electrophoresis have widely expanded. These new technologies allow a more global view on cellular processes on different levels of observation and can be embraced by the term omics technologies, including genomics, transcriptomics, proteomics and metabolomics, and many more.

The term "omics" stands for the comprehensive analysis of the respective level of biological systems, as proteomics refers to the study of the proteome of a cell in a given state. Genomics attempts to describe a cell or an organism in terms of the sequence of its genome, whereas transcriptomics examines gene expression on the RNA level. The most recent subdiscipline is metabolomics, which aims to detect and quantify the low molecular weight molecules known as metabolites.

As technical and experimental limitations restricted scientists to a purely hypothesis-driven research, the progress of omics technologies enables a complementation with explorative, data-driven approaches. The rational behind an unbiased analysis is that features of biological processes may be found in an initially unexpected context.

All these advances have induced a novel paradigm in molecular biology research:

- HT technologies allow a global and integrative insight into biological systems and led to the development of **systems biology** as a multidisciplinary approach.

- The need of administration and interpretation of the increasing amount of heterogenous data, as derived from HT experiments, triggered the emergence of the subdiscipline **computational systems biology**.

The next section provides an overview of the concepts of systems biology, as well as of the subdiscipline computational systems biology.

1.2 Concepts of Systems Biology

The systems biology approach arises from the rational that properties of biological systems cannot be reduced to those of their parts. The advance of omics technologies has established a basis for integrative studies of biological processes on more than one level of observation. Thus, generation, management, and interpretation of the available data poses a multidisciplinary challenge.

The field of systems biology attempts to provide a systems-level understanding by systematically organizing the different omics data, using it to build a descriptive and mechanistic model of the underlying biological phenomena [1]. Systems modeling of a physiological process beginning at the level of genes and gene networks is a highly iterative process involving cycles of data collection, quantitative modeling, hypothesis formulation and testing, and model refinement [2]. With the possibility to generate quantitative data, a shift towards dynamic, quantitative models was induced. Whereas qualitative biological models are in their nature discrete, the quantitative approach tries to provide an understanding of the dynamical characteristics of the whole system.

With these evolving concepts in systems biology, the integration of experimental and computational research has become necessary. The management of the large amount of data and the complex coherences between different levels of observations require computational power. Protein as well as DNA sequence databases are growing at steady pace. Examples are SWISSPROT [3], the Protein Data Bank (PDB) [4] or the NCBI (National Center for Biotechnology Information) RefSeq [5]. Furthermore, databases providing functional analysis tools have developed, including the protein analysis through evolutionary relationships (PANTHER) classification system [6] and the pathway databases KEGG (Kyoto Encyclopedia of Genes and Genomes) [7].

A further challenge is to resolve the problem of integrating heterogeneous data from different sources with varying degrees of reliability. Functional mapping projects have recently emerged, generating large-scale maps from functional omics data. The

UBC Bioinformatics Center developed Atlas, a biological data warehouse that locally stores and integrates biological sequences, molecular interactions, homology information, functional annotations of genes, and biological ontologies [8]. Aerts et al. designed a web tool named Endeavour, that prioritizes candidate genes underlying biological processes or diseases, based on their similarity to known genes by combining multiple data sources [9].

Of particular interest in context of data integration are protein-protein interactions. Represented as molecular interaction networks, they can serve as a basis for analysis of the dynamics of biological systems. An outline of the major concepts of protein interaction networks and an overview of recent approaches is given in the following section.

1.3 Protein Interaction Networks

Networks of molecular interactions are widely studied to reveal the complex roles played by genes, gene products and the cellular environment in biological processes. These networks can either represent direct physical binding of proteins, or functional relationships between the involved objects. The former approach is widely covered by protein interaction databases like OPHID [10], BIND [11], HPRD [12, 13], MIPS [14] or the MINT database [15].

However, understanding biological processes equally requires knowledge of indirect relationships. These can include shared pathway or process memberships, the same cellular localization or similar tissue specific expression levels. Von Mering and colleagues developed the database STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) [16], holding interactions derived from high-throughput experimental data, from the mining of databases and literature, and from predictions based on genomic context analysis. A further approach is a system designed by Myers et al. for predicting process-specific networks, including information derived from microarray data sets, protein-protein interactions, as well as sequence data [17].

Within this thesis, networks representing physical protein interactions, as well as a dependency graph analysis were integrated in the applied workflows. Subjects of study were two sets of gene expression data characterizing an animal model of membranous nephropathy and pathophysiological processes encountered after transplantation of cadaveric donor kidneys.

1.4 Kidney Diseases

In the recent years, a multitude of studies have been performed tackling a wide range of kidney diseases. High-throughput technologies facilitated the identification of relevant molecular mechanisms and protein biomarkers associated with common kidney diseases. In this thesis, two different aspects of kidney diseases were analyzed. The following subsections provide an overview of the pathophysiology of these diseases as well as of the current status of protein biomarkers, identified in recent studies in the context of acute renal failure (ARF) and chronic kidney disease (CKD).

1.4.1 Pathophysiology

Experimental Membranous Nephropathy

Membranous nephropathy (MN) is an antibody mediated and complement dependent disease, which can lead to ARF as well as CKD. The target of injury in MN is the glomerular visceral epithelial cell or podocyte, a highly specialized and terminally differentiated cell that is located on the outside of the glomerular basement membrane [18]. Binding of antibodies to membranous antigenes, identified as megalin associated complexes [19], leads to complement activation and formation of subepithelial immune deposits. The hallmarks of MN are thickening of the basement membrane due to an increase in the accumulation of extracellular matrix protein synthesis by injured podocytes and loss of the glomerular filtration barrier followed by proteinuria.

Passive Heymann Nephritis (PHN) is a rat model that shows similarity to human MN. Rats are immunized with an antibody directed against the crude renal fraction Fx1A that contains megalin as a principal component [20]. This animal model was used as a basis for experimentally determining the gene expression pattern of glomeruli in diseased rats for learning more on the pathophysiology of MN.

Renal Transplants

The incidence of end-stage renal disease increased over the last years. Kidney transplantation is the treatment of choice for most of these patients, but the number of

kidneys available for transplantation is limited [21]. More and more patients are reliant on cadaveric donor kidneys, but clinical observations report severe organ inflammation, post-transplant ARF and a reduced long-term survival of transplants when compared to patients recieving transplants from living donors.

The risc factors for a delayed graft function include donor age and cause of death, the duration of cold ischemia, or intraoperative diuresis [22]. However, intrinsic donor factors are main contributors to post-transplant ARF. Brain-death often leads to the development of a central diabetes insipidus (excretion of large amounts of severely diluted urine) and following dehydration, in turn causing arterial hypertension [23]. All of these mechanisms are risc factors for the development of post-transplant ARF.

Still, it remains unclear what specifically increases the propensity for the cadaveric renal graft to develop ARF after engraftment. To gain further insights in the ongoing biological mechanisms, we analyzed the results of a gene expression analysis comparing living and cadaveric donor kidney biopsies in context of functional dependencies of differentially regulated genes, represented in a dependency graph of protein interactions.

1.4.2 Biomarkers

As kidney diseases are major health problems with increasing incidence, the need for identification of novel biomarkers for early diagnosis and prognosis is rising. A frequently used marker for ARF in clinical practice is creatinine, but creatinine levels do not change until loss of renal function has far progressed. Functional markers for CKD are again creatinine or Cystatin C [24], both used to estimate the glomerular filtration rate.

Over the last years, the pool of biomarker candidates could be extended by explorative analysis approaches, enabled by omics technologies. Due to the limitation in number of samples, the false positive rate of potential markers is rather high and requires further experimental verification steps.

Based on a literature review, Perco et al. [25] summarized protein markers reported as associated with ARF and CKD. The underlying detection methods were partly

hypothesis driven, executed as immunohistochemistry or western blot analysis, but also results from microarray experiments were reported.

In view of the fact that biomarkers for early prognosis of renal failure are still missing, further studies are essential to also accomplish the development of improved therapeutic approaches.

1.5 Thesis Goals

The main purpose of this thesis was to identify relevant genes and potential biomarkers in context of kidney diseases by means of microarray technology and subsequent bioinformatic analysis with focus on the mapping of gene expression data on protein networks. In particular two different approaches were applied, namely a sequential and an integrated analysis workflow. The aim of the sequential workflow was to detect genes differentially regulated between healthy and passive heyman nephritis induced rats. This workflow mainly rests on statistical data analysis. The integrated workflow in contrast is based on a dependency graph approach representing a human protein interaction network, and was applied to a set of differentially expressed genes comparing living and cadaveric renal transplants.

MATERIAL

2.1 Experimental Membranous Nephropathy - The PHN Dataset

Passive Heymann Nephritis (PHN) is a rat model that shows similarity to human membranous nephropathy. It remains a valuable experimental tool because the functional and immunohistological features closely resemble the human disease[26]. To determine which known genes are transcriptionally up or down-regulated in PHN, a gene expression analysis was performed by the group of Prof. Shankland at the University of Washington, Seattle, WA, USA.

To induce passive Heymann nephritis twenty male Sprague-Dawley rats (Simson, Gilroy, CA, USA) received a single intraperitoneal injection of sheep anti FX1A antibody as described by Shankland et al.[27]. Twenty additional control animals were injected with normal sheep serum. In order to assess proteinuria and renal function, urine was collected by placing the animals in metabolic cages for 12 hours, during which time water was supplied without restriction. Protein and creatinine excretions were measured using the sulfosalicylic acid turbidity method[28] and a colorimetric microplate assay based on the Jaffe reaction [29] (Oxford Biomedical Research, MI, USA) respectively.

In each case, half of the animals were sacrificed after three and the other half after six days. For glomeruli isolation, the kidney cortex was removed, minced and pressed through sieves. Than, the glomeruli were collected and pelleted in phosphate buffered saline by centrifugation. Subsequently, the total RNA was isolated, using the TRIZOL method (Invitrogene Corp, Carlsbad, CA, USA). Quantity and OD 260/280 of total RNA and cRNA was assessed by UV spectrophotometry and cRNA was labeled with Biotin according the Affymetrix eukaryotic target labeling protocol.

The RNA of two animals was pooled and hybridized to one microarray, resulting in five biological replicates for disease and control animals each at days 3 and 6. The samples were hybridized on Affymetrix R230A GeneChip arrays according to standard Affymetrix protocol.

The resulting raw data of the images was stored in .DAT files, the analysis in .CHP files and probe set information in .CEL files. The latter, each holding 15924 probes and their corresponding intensity values, were further analyzed as described in section3.1.

2.2 Renal Transplants - The LIV/CAD Dataset

A high percentage of cadaveric, but rarely living donor renal transplant recipients develop postischemic acute renal failure (ARF). ARF patients tend to reduced long-term allograft survival and acute rejection occurs more frequently [30]. To identify regulatory pathways of ARF, a genome-wide expression analysis was performed by the group of Prof. Oberbauer at the Medical University of Vienna.

The gene-expression pattern was determined in three classes of 32 donor kidney biopsies: 12 living donor kidneys with primary function, 12 cadaveric donor kidneys with primary function and 8 cadaveric donor kidneys with biopsy proven acute renal failure. Immediately before transplantation, the wedge biopsies were obtained, instantly submerged in RNAlater^{TM} (Ambion, Austin, TX, USA) and homogenized [31]. Total RNA was isolated and purified with RNeasy columns (Qiagen, Hilden, Germany). The RNA yield and quality was checked with the Agilent 2100 Bioanalyzer and RNA6000 LabChip® kit (Agilent, Palo Alto, CA, USA). Stratagene universal human reference RNA was used as standard (Stratagene, La Jolla, CA, USA). Since that the total amount of isolated RNA was very small, a T7 RNA amplification step using the RiboAmp RNA amplification kit (Arcturus, Mountain View, CA, USA) was necessary [32].

The samples were hybridized on cDNA arrays, obtained from the Stanford University Functional Genomics core facility, each holding 26338 genes and 14783 ESTs. Preprocessing and further analysis steps were performed as described in sections 3.1 and 3.2.

METHODS

The following chapter describes the methods used for the analysis of the datasets, generated as outlined in chapter 2. The first section reports the sequential analysis workflow applied on the PHN, as well as on the LIV/CAD dataset. Section 2 provides an overview of the dependency graph approach that was carried out to complement the initially identified, differentially expressed genes of the LIV/CAD set.

3.1 Sequential Analysis Workflow

3.1.1 DNA Microarrays

The majority of DNA microarrays are classified as oligonucleotide arrays and a cDNA arrays, according to the type of probes immobilized thereon. Both work on the principal of base-pairing, allowing probes to hybridize to the targets on the microarray. The major difference between these two arrays is the fact that cDNA arrays are usually used to represent a relative measurement of gene expression, whereas oligonucleotide array results contain absolute values on mRNA concentration. This section provides an overview of design and techniques of cDNA arrays used for analysis of living and cadaveric donor kidney biopsies (see section 2.2), as well as of oligonucleotide arrays which were used for analysis of gene expression in experimental PHN (see section 2.1).

cDNA Arrays Probes used for cDNA arrays are chemically synthesized complementary DNA (cDNA) strands, containing only fragments of the coding part of the sequence, complementary to its corresponding mRNA transcript. The total RNA is first extracted from the experimental samples to be fluorescently labeled in a single round of reverse transcription. In case of two-color experiments, cy3-dUTP (green) and cy5-dUTP (red) are preferentially used because they are readily incorporated

by reverse transcription; they exhibit good photostability and most importantly, are widely separated in terms of their excitation and emission spectra. The fluorescently labeled cDNA probes are hybridized to a single array in a competitive hybridization reaction. Detection of hybridized probes is achieved by laser excitation of the individual fluorescent markers, followed by scanning using a confocal scanning laser microscope. The raw data is represented as a ratio of cy3:cy5 and digitally color coded such that red represents genes transcriptionally upregulated in the test versus the reference, green represents genes downregulated and yellow represents those genes that exhibit no difference between test and reference samples [33]. Besides the two-color approach, samples can also be hybridized to seperate arrays, resulting in absolute values mRNA concentration. Popular one-color systems are Affymetrix GeneChips wich are described in the next paragaph.

Affymetrix GeneChips The Affymetrix GeneChip is one of the most widely used oligonucleotide array. Whereas cDNA microarrays use long strands of DNA ($\sim$300 nucleotides) as fixed probes, oligo-chips use oligonucleotide sequences ($<$25 nucleotides) as their probes. A great part of the genome of an organism can be placed on a single microarray as oligonucleotide probes, where each sequence is usually around 25 base pairs in length. These probes are synthesized onto a glass wafer by a combination of semiconductor-based photolithography and solid phase chemical synthesis technologies.

One component of each probe, referred to as the perfect match probe (PM) is a sequence, perfectly complementary to a unique region at the 3 end of the particular gene. Previous versions of Affymetrix GeneChips were produced with an additional component, the mismatch probes (MM), which were created by changing the middle base with the intention to determine the background and nonspecific hybridization that contributes to the signal for the perfect match oligo. The arrays are scanned and images are produced and analyzed to obtain an intensity value for each probe. These intensities represent how much hybridization occurred for each oligonucleotide probe [34].

3.1.2 Data Preprocessing

Preprocessing procedures depend on the microarray technology in use. In general, preprocessing of Affymetrix chips involves 4 steps, summarized in Table 3.1 [35].

Steps	**Description**
Background correction	For each array, an estimate of the background signal, due to unspecific binding of probes, or chip surface autofluorescence, is generated and removed.
Normalization	Signal intensities are normalized to compare data from one chip to another.
PM correction	To adjust the perfect match signal intensities. To account for nonspecific signals, the information derived from the mismatch signal, is taken.
Expression summary	The probe intensities within a given probe set are combined to a single value.

Table 3.1: Affymetrix data preprocessing Steps

In the last years, a multitude of normalization methods have evolved. Four of the commonly used procedures are listed below:

- RMA - Robust Multi-chip Analysis [34]
- MAS5 - Affymetrix Microarray Suite
- VSN - Variance Stabilizing Normalization [36]
- MBEI - Model Based Expression Index [37]

Preprocessing of the raw data derived from the PHN animal model was carried out in CARMAweb [38] (comprehensive R- and bioconductor-based web service for microarray data analysis). The application provides different normalization methods, including those techniques mentioned above, for all current microarray platforms. For preprocessing of the PHN data, RMA was used and will be described in more detail.

Background Correction

One characteristic of RMA is to exclusively use PM probes and to ignore MM signals. The method is based upon the assumption that the observed PM probe signal O consists of a normally distributed background or noise component N and an exponentially distributed true signal S:

$$O = N + S, \quad N \sim N(\mu, \sigma^2), \quad S \sim Exp(\alpha)$$

where μ is the mean and σ^2 is the variance of N and α is the rate of the exponential. All three parameters are assumed to be equal for all PM probes on a chip and can therefore be estimated from the data. Eventual negative values for N are truncated [39].

The corrected intensities are given by

$$E(s|O = o) = a + b\frac{\phi(\frac{a}{b}) - \phi(\frac{o-a}{b})}{\Phi(\frac{a}{b}) + \Phi(\frac{o-a}{b}) - 1}$$

where $a = o - \mu - \sigma^2\alpha$ and ϕ and Φ are the standard normal density and distribution functions, respectively.

Normalization

Due to small differences in RNA quantities and fluctuations generated by the applied technique, the intensity levels may vary from one replicate to the other due to effects which are unrelated to the genes, requiring data normalization before they can be compared. In case of RMA, quantil normalization, described by Bolstad et al. [40] is used to remove non-biological variability between all arrays.

Goal of this method is to unify the distribution of probe intensities for each array in a set of arrays. The rational behind is that a quantile - quantile plot shows that the distribution of two data vectors is the same if the plot is a straight diagonal line. Usually, the quantiles of two arrays do not lie on the diagonal. To regain the same distributions the quantiles can be projected to the diagonal of the quantile-quantile

plot which is equivalent to replacing every component of a quantile vector by the mean of that vector. Hence, the quantile normalization method is a specific case of the transformation

$$x\prime_i = F^{-1}(G(x_i))$$

where G is estimated by the empirical distribution of each array and F by using the empirical distribution of the averaged sample quantiles.

PM Correction

As mentioned above, MM intensities are ignored within the RMA framework. For this reason RMA skips this preprocessing step.

Expression Summary

Typically, each gene on an Affymetrix GeneChip is represented by 16-20 oligonucleotide pairs. The purpose of the summarization step is to establish a single expression value for each gene on the chip. RMA uses the Median Polish method which is based on an additive linear model. The following formula indicates that the observed, background-corrected and quantil-normalized intensity value y_{ij} is assumed to be the summ of the probe affinity effect α_i of probe i, plus the real hybridization intensity μ_j for array j and the independent, identically distributed error term ϵ_{ij} with a mean value of zero [34].

$$\log_2(y_{ij}) = \alpha_i + \mu_j + \epsilon_{ij}$$

The expression value for the respective probe set is the estimated $\hat{\mu}$ for an array j. This effect is estimated in a robust way by using the median polish algorithm.

The expression values are placed in a matrix where rows represent the probes and columns the arrays. The fitted matrix is obtained by alternately subtracting the row and column medians from the matrix elements. Row and column vectors are updated during each iteration and this process is repeated until the matrix changes

by less than a small margin or until a pre-defined maximum number of iterations is reached. This residual matrix is subsequently subtracted from the original, resulting in a matrix holding the fitted expression values. An estimate for a specific probe value is obtained by calculating the corresponding row average.

This algorithm provides robust estimates for two reasons. First, using medians rather than means makes it less sensitive against outliers, and second, estimations are based on the entire set of arrays.

Preprocessing - cDNA Arrays

Normalization of the cDNA arrays, in this thesis used to derive a gene-expression profile of the donor kidneys, was done with the microarray image analysis software integrated in GenePix [41]. The procedure includes background correction and adjustment between the red and the green channel.

To reduce the number of redundant genes, and genes with missing values (initially each of the 32 investigated arrays hold 41121 partially redundant genes), a filter was applied to every gene in the dataset. Only those genes with values in at least 80% of the experiments were included in further analysis. Thus, genes with expression levels similar to the background intensity are eliminated by this step along with basically unexpressed genes.

The remaining missing values were substituted applying a k-nearest-neighbor algorithm (KNN). The algorithm finds the k genes that are most similar in expression to the gene with the MV as determined by a distance metric. The missing value is then estimated as the average of these k neighbor genes for the same array, weighted according to the inverse of their distance [42]. In case of the donor kidney experiment k was set to 10.

Comparison of two microarray datasets usually leads to systematic biases arising from variability in experimental conditions. To prevent an erroneous detection of differences in the gene expression pattern, several methods have been developed, including singular value decomposition (SVD), which was used to correct biases in the donor kidney experiment. SVD is a linear transformation of the expression data

from the genes × arrays space to the reduced eigengenes × eigenarrays space. In this space the data are diagonalized, such that each eigengene is expressed only in the corresponding eigenarray, with the corresponding eigenexpression level indicating their relative significance. The eigengenes and eigenarrays are unique, and therefore also data-driven, orthonormal superpositions of the genes and arrays, respectively [43].

3.1.3 Clustering

The idea of the clustering step is to group similar data objects and discover patterns in a given pool of data. Objects are grouped based on their proximity to each other via a distance metric. In case of the analysis of the donor kidney experiment, the correlation distance

$$d(X,Y) = 1 - r_{XY}$$

was used, where r_{XY} is the Pearson correlation coefficient between two vectors X and Y, and

$$d(X,Y) = 1 - \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

where n is the length of the vectors and $\bar{X}$ and $\bar{Y}$ are their mean values. Since the Pearson correlation coefficient r_{XY} takes values between -1 and 1, the distance $1-r_{XY}$ will vary between 0 and 2. The Pearson correlation finds whether two objects vary in the same way. The correlation will be high if the corresponding expression levels increase or decrease at the same time, otherwise the correlation will be low.

Once the distances between all objects are calculated, a linkage rule, determining the inter-cluster distances, has to be defined. In general, there are three possibilities:

Single linkage calculates the distance between clusters as the distance between the nearest neighbors by measuring the distance between each member of one cluster to each member of the other cluster and taking the minimum of these.

Average linkage measures the average distance between each member of one cluster to each member of the other cluster.

Complete linkage defines the distance between the furthest neighbors by taking the maximum of distance measures between each member of one cluster to each member of the other cluster.

For further analysis a hierarchical cluster algorithm within Statistica 6 (Statsoft Inc., Tulsa, OK, USA), using the complete linkage rule, was computed. Figure 3.1 gives a graphical representation of the algorithms function.

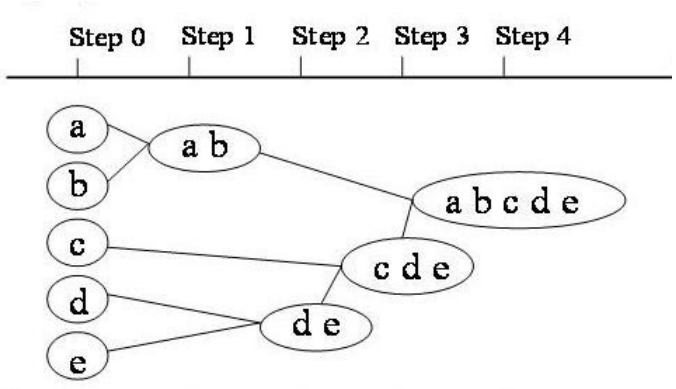


Figure 3.1: The basic agglomerative clustering steps.

The applied steps are:

1. Calculate the correlation distance between all data points.
2. Cluster the data points to the initial clusters, each consisting of one element.
3. Calculate the inter-clusters distances.
4. Repeatedly cluster most similar clusters into a higher level cluster with respect to step 3.
5. Repeat steps 3 and 4 for the most high-level clusters.

Besides this agglomerative (bottom up) procedure, clusters can also be build divisive (top down), starting with a super-cluster containing all elements, followed by splitting iteratively into sub-classes.

3.1.4 Statistical Analysis

The next step is to discover and identify genes that are differentially expressed between the groups of samples, as in the PHN experiment, between healthy and diseased rats. Such microarray experiments generate large multiplicity problems in which thousands of hypotheses have to be tested simultaneously. For every single gene the following decision has to be made:

1. the observed gene is over/under expressed in comparison to the healthy group
2. the gene is **not** differentially expressed in comparison to the healthy group

For the decision a t-test, assuming unequal sample sizes and unequal variance, is calculated according to the equation below:

$$t = \frac{X - Y}{\sqrt{\frac{(\sigma_x)^2}{n_x} + \frac{(\sigma_y)^2}{n_y}}}$$

where X refers to group1, Y to group2, σ to the corresponding standard deviation, and n to the size of the sample.

The test results in either rejecting the null hypothesis H_0 which is equivalent to case 1 or accepting H_0 which is the same as described in case 2. The latter could also be described as rejecting the alternative hypothesis H_A.

A statistical decision process usually comes along with possible errors. The two sources of error are listed below:

type I error: the error of rejecting a hypothesis that should have been accepted ($\Rightarrow$ false positive)

type II error: the error of accepting a hypothesis that should have been rejected ($\Rightarrow$ false negative)

When multiple hypotheses are tested, as in case of microarray analysis, the probability that a type I error is committed increases sharply with the number of hypotheses. There are some standard type I error rates [44]:

FWER - Family Wise Error Rate: The FWER is defined as the probability that the analysis yields any false positive findings.

$$FWER = Pr(V \geq 1)$$

where V is the number of false positives.

PFER - Per Family Error Rate: The PFER measures the expected count E of false positives.

$$PFER = E(V)$$

FDR - False Discovery Rate: The FDR can be interpreted as the expected proportion Q of significant findings that are indeed false positives.

$$FDR = E(Q)$$

PCER - Per-Comparison Error Rate: The PCER is defined as the ratio of the expected false positives to the number of hypothesis G.

$$PCER = \frac{E(V)}{G}$$

maxT Multiple Testing

Statistical significance of genes separating arrays of the donor kidney biopsies in distinct clusters was performed by calculating p-values, which were corrected for multiple testing using a maxT step-down procedure [45] with the aim to control the FWER. In step-down procedures, the hypotheses corresponding to the most significant test statistics are considered successively, with further tests depending on the outcomes

of earlier ones. As soon as one hypothesis is accepted, all remaining hypotheses are accepted.

The algorithm estimates the joint distribution of the test statistics $t_1, ..., t_m$, where m is the number of tested hypothesis, under the complete null hypothesis H_0 by permuting the samples. For the b^{th} permutation, $b = 1, \ldots, B$

1. Permute the n columns of the data matrix X.

2. Compute the test statistics $t_{1,b}, \ldots, t_{m,b}$ for each hypothesis

3. Next, compute successive maxima of the test statistics

$$u_{m,b} = |t_{r_m}, b|$$

$$u_{j,b} = max\left(u_{j+1,b}, |t_{r_j}, b|\right) \quad for\, j = m-1, \ldots, 1,$$

where r_j are such that $|t_{r_1}| \geq |t_{r_1}| \geq \ldots \geq |t_{r_m}|$ for the original data.

The adjusted p-values are estimated by

$$\tilde{p}_{r_1} = \frac{\sum_{b=1}^{B} I\left(u_{j,b} \geq |t_{r_j}|\right)}{B}$$

with the monotonicity constraints enforced by setting

$$\tilde{p}_{r_1} \leftarrow \tilde{p}_{r_1}, \qquad \tilde{p}_{r_1} \leftarrow max\left(\tilde{p}_{r_1}, \tilde{p}_{r_1-1}\right) \qquad for\, j = 2, \ldots, m.$$

SAM - Significance Analysis of Microarrays

Tusher et al. [46] developed a method, Significance Analysis of Microarrays (SAM), addressing the FDR. On the basis of t-tests, it assigns a score to each gene according to the change in gene expression relative to the standard deviation of repeated measurements. For genes with scores greater than an adjustable threshold, SAM uses permutations of the repeated measurements to estimate the percentage of genes

identified by chance. This method was applied in the course of analyzing the PHN rat model.

Since the signal-to-noise ratio is decreasing with decreasing gene expression and the fluctuations are gene specific, SAM defines the relative difference $d(i)$ in gene expression as follows:

$$d(i) = \frac{\bar{X}_I(i) - \bar{X}_U(i)}{s(i) + s_0}$$

where $\bar{X}_I(i)$ and $\bar{X}_U(i)$ are defined as the average levels of expression for gene i in states I and U (in our case healthy and diseased), respectively. The gene-specific scatter $s(i)$ is the standard deviation of repeated expression measurements:

$$s(i) = \sqrt{a \left\{ \sum_m [X_m(i) - \bar{X}_U(i)]^2 + \sum_n [X_n(i) - \bar{X}_I(i)]^2 \right\}}$$

where $\sum_m$ and $\sum_n$ are sums of the expression values in states I and U, respectively, $a = \frac{(\frac{1}{n_1} + \frac{1}{n_2})}{(n_1 + n_2 - 2)}$, and n_1 and n_2 are the numbers of measurements in states I and U.

For each balanced permutation, relative differences $d_p(i)$ are calculated. The expected relative difference, $d_E(i)$, is defined as the average over all permutations, $d_E(i) = \sum_p \frac{d_p(i)}{n}$.

3.1.5 Functional Analysis

The next step was to discover and interpret the biological function of the genes deemed to be of interest following the analysis of the microarray experiment. Different tools and databases exist that facilitate and fasten the information search.

The GO project[47] has developed three structured, controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. The terms are structured as an acyclic directed graph, providing a hierarchial functional annotation of genes.

Another ontology, used for analysis of differentially expressed genes, is similar in structure: the protein analysis through evolutionary relationships (PANTHER) classification system [6]. Using this tool, enriched and depleted functional categories were identified using the PANTHER data set covering the whole rat genome of interest as reference dataset. The ratio of expected to observed frequencies of genes assigned to certain ontology categories were compared using the χ^2 test to derive significance of differences.

Next to gene ontologies, pathway databases like KEGG (Kyoto Encyclopedia of Genes and Genomes) [7], holding information on metabolic networks and signaling transduction cascades, were searched.

Besides the manual search in the described databases, the statistically significant genes in the donor kidneys were further characterized using GenMAPP [48] and High-density Array Pattern Interpreter (HAPI) [49]. GenMAPP is a computer application designed to visualize gene expression data on maps representing biological pathways and groupings of genes. HAPI provides a data mining method that uses keywords from the published literature linked to specific genes to present a view of the similarity of genes within a group of interest.

3.1.6 Network Analysis

Grouping genes with similar functions or genes interacting is usually the next step in interpreting the data. Starting with the set of differentially expressed genes in the PHN model, a protein-protein interaction network following the nearest neighbour expansion method [50] was generated, using data from the Online Predicted Human Interaction Database (OPHID) [10] on interactions of rat proteins.

To identify possible molecular complexes the graph theoretical clustering algorithm MCODE (Molecular Complex Detection) [51] was applied. It operates in three stages, vertex weighting, complex prediction and optionally post-processing to filter or add proteins in the resulting complexes by certain connectivity criteria.

Vertex weighting is based on the local network density. The weight given to a vertex i is the product of the vertex core-clustering coefficient C_i and the highest k-core level k_{max} of the immediate neighborhood of the vertex. A k-core is a graph of minimal degree k, hence k_{max} is the central most densely connected subgraph. Differing from the definition of the clustering coefficient C_i

$$C_i = \frac{2n}{k_i(k_i - 1)}$$

where k_i is the vertex size of the neighborhood of vertex i, excluding i itself, and n is the number of edges in the neighborhood, the core-clustering coefficient includes the vertex i. This results in amplifying the weighting of heavily interconnected graph regions while removing the many less connected vertices that are usually part of a biomolecular interaction network, known to be scale-free.

Complex prediction takes as input the vertex weighted graph. Given a specified vertex weight percentage (VWP), the algorithm seeds a complex with the highest weighted vertex and recursively moves outward including vertices in the complex whose weight is above the given threshold. This way, the most dense regions of the network can be identified.

This sequential workflow rests on transcripts and proteins which show statistically significant differences in expression levels when compared to a control sample. However, measured differences in mRNA abundance do not necessarily correlate with biological relevance. The following section describes an integrated workflow which is aimed at the expansion of an existing core set of differentially expressed genes in renal transplants on the basis of a dependency graph.

3.2 Integrated Analysis Workflow

The following section describes an analysis workflow, based on the concept of a dependency graph. In contrast to the previously outlined sequential analysis, this approach integrates data on genes, RNA and proteins to objects and augment them with information, derived from different levels of observation. Represented as nodes, these objects are linked by edges which are weighted according to estimated functional dependencies.

3.2.1 Object Annotation

The annotation of the molecular objects in the dependency graph integrates given data from different sources. Object definition includes a genes functional annotation, its reference gene expression profile determined for 32 tissues, the interaction data of encoded proteins, and a consensus on each proteins subcellular location. Table 3.2 summarizes the data sources, used for object annotation. A detailed description of the different contributions is given in the following paragraphs.

Data Source	Level of Observation	Data Family
OPHID KEGG PANTHER IntAct	interactome	INT
GSE7905	transcriptome	DGE
KEGG GO PANTHER	phenome	SEM
WPSORT SWISSPROT Omenn et al.	localisome	LOC

Table 3.2: Data Sources

Functional Interaction (INT) The Functional Interactions data family constitutes on the basis of four different data sources: The OPHID database, the IntAct

database [52] and parsed interactions from the KEGG- [7] and PANTHER [6] pathways.

The annotated interactions in OPHID (The Online Predicted Human Interaction Database) are mainly collected from BIND [11], HPRD [12, 13], MIPS [14], DIP [53], and MINT database [15], while another part of interactions are derived by mapping high-throughput model organism data from S. cerevisiae, C. elegans, D. melanogaster, and M. musculus to human proteins. Furthermore OPHID contains PPI predictions based upon gene co-expression in expression profiles from GeneAtlas [54], domain co-occurrence and similarity measures in GO.

The data available in the IntAct database originate entirely from published literature and is manually annotated by expert biologists. IntAct was built by the EBI-EMBL (European Bioinformatics Institute of the European Molecular Biology Laboratory) and made publicly available using the PSI-MI (Proteomics Standards Initiative Molecular Interaction format) XML Standard [55].

KEGG (Kyoto Encyclopedia of Genes and Genomes) and PANTHER (Protein ANalysis THrough Evolutionary Relationships) are databases consisting of pathways which can be used for extracting PPIs. The KEGG database contains a collection of manually drawn molecular pathway maps, each map representing KEGG's knowledge on the molecular interaction and reaction networks for metabolism, genetic information processing, environmental information processing, cellular processes and human diseases, whereas PANTHER primarily describes signaling pathways.

The overall number of functional protein-protein interactions derived from OPHID, IntAct, KEGG and PANTHER resulted in 86012 unique interactions which were further augmented with information from data families as described in the following paragraphs.

Gene Expression Patterns (DGE) According to a public domain GEO dataset GSE7905 [56], holding normalized gene expression profiles of 32 human tissues, vectors with tissue specific expression levels could be assigned to 15382 molecular objects.

Semantic Annotation (SEM) The semantic annotation terms were collected from the Gene Ontology (GO) [47], KEGG and PANTHER databases.

Semantic annotation terms from KEGG and PANTHER pathways were extracted through parsing each pathway for the underlying set of genes. Each pathway's name can be considered to be a semantic annotation for the genes contained in this pathway. Because of the hierarchial nature of the GO annotation as described in the previous section, different annotation terms are more or less specific resulting in the fact that different GO annotation terms span from annotating very few up to some thousand genes. Hence, a normalization step was implemented to make terms from the three different sources comparable. Finally, 586 unique semantic annotation terms were integrated in the object annotation.

Subcellular Location (LOC) To enrich the molecular entities with information about the subcellular location for each underlying protein, the prediction algorithm WPSORT (WoLF PSORT) [57] was used and combined with experimentally available subcellular location information as given in the SWISSPROT subcellular location comment block [3], and the plasma proteome as reported by Omenn et al [58].

WPSORT is an extension of the PSORT II program for protein subcellular location prediction. It converts protein amino acid sequences into numerical localization features, based on sorting signals, amino acid composition and functional motifs such as DNA-binding motifs. The algorithm results in a numeric vector of length ten where each numeric value represents the occurrence probability in the following classes of organelles in percent: cytosol, cytoskeleton, endoplasmic reticulum, extracellular, golgi body, lysosome, mitochondria, nuclear, peroxisome, plasma membrane.

The SWISSPROT protein knowledgebase connects amino acid sequences with an overview of relevant information, including experimental results and computed features.

For refinement and correction, the experimentally validated data from SWISSPROT and Omenn et al. were normalized to the WPSORT format. To give consideration to the more accurate information derived from experiments, the corresponding vector entries (v_{SWI},v_O) were exclusively set to 0% and 100%.

Data integration followed the principle:

$$v_{LOC_i} = \frac{v_{WPSORT_i} + v_{SWI_i} + v_{O_i}}{\sum_{i=1}^{10} v_{WPSORT_i} + v_{SWI_i} + v_{O_i}}$$

where v_{WPSORT_i}, v_{SWI_i} and v_{O_i} denote the localization probabilities and i denotes the position in the numeric vectors for the respective organelle. The sum of each position in the numeric vectors is divided by the sum of all vectors probability results in v_{LOC_i}. Thus, 18138 molecular objects could be annotated by the means of their proteins subcellular location.

It has to be mentioned that the described annotation of graph objects does not consider the different splice variant products which can differ substantially in function. A future goal is to overcome this shortcomings but this issue is nontrivial since that the number of splice products is enormous.

3.2.2 Graph Construction

For construction of the dependency graph only those 18572 objects were considered, showing to have a corresponding identifier to their NCBI Gene Symbol in the NCBI Reference Sequence collection (RefSeq) [5]. Each object consists of 4 data entries o_{INT}, o_{DEG}, o_{SEM}, o_{LOC}, corresponding to the data families described in the previous section.

The data entries were realized as vectors, like for an entry $o_{i_{INT}}$, the corresponding string vector consists of all objects o_j, where $1 \leq j \leq 18575$ and $i \neq j$, having an interaction with o_i. The $o_{i_{DEG}}$ entry is a numeric vector where each coordinate in the vector represents a quantile normalized signal to noise expression value for 32 human tissues. Annotation terms, according to an entry $o_{i_{SEM}}$, are represented as coordinates of a string vector and the $o_{i_{LOC}}$ entry is a probability vector, containing probabilities for the occurrence of o_i in each of the 10 subcellular locations as listed in the previous section.

Dependency Matrix

The pairwise distances d_{ij} between any two objects o_i and o_j were stored in a dependency matrix D. Definitions for the 4 contributions to d_{ij} are given below:

- $f_{INT_{ij}}$ is a binary function:

$$f_{INT_{ij}} = \begin{cases} 1 & if \quad o_{i_{INT}} \bigcap o_{j_{INT}} \neq \{\emptyset\}, \\ 0 & if \quad o_{i_{INT}} \bigcap o_{j_{INT}} = \{\emptyset\}, \end{cases}$$

- $f_{DEG_{ij}}$ is the Pearson coefficient of correlation between two objects $o_{i_{INT}}$ and $o_{j_{INT}}$, μ_i and μ_j their expected values and σ_i and σ_j their standard deviations:

$$f_{DEG_{ij}} = \frac{[(o_{i_{DEG}} - \mu_i)(o_{j_{DEG}} - \mu_j)]}{\sigma_i \sigma_j}$$

- $f_{SEM_{ij}}$ calculates the Dice coefficient (taken as a string similarity measure) between two objects:

$$f_{SEM_{ij}} = \frac{2|o_{i_{SEM}} \bigcap o_{j_{SEM}}|}{|o_{i_{SEM}}| + |o_{j_{SEM}}|}$$

- $f_{LOC_{ij}}$ is defined as the difference between the probability vectors $o_{i_{LOC}}$ and $o_{j_{LOC}}$:

$$f_{LOC_{ij}} = 1 - \sum_{x=1}^{10} |o_{i_{LOC}} - o_{j_{LOC}}|$$

To obtain a pairwise dependency score, the following function was computed for each entry d_{ij} in D:

$$d_{ij} = f_{INT_{ij}} + 2\frac{f_{DEG_{ij}} + f_{SEM_{ij}} + f_{LOC_{ij}}}{3}$$

Following this function, object dependencies are biased for physical or logic protein interactions (contributing with an dependency score of +1), whereas the other three contributions are equally weighted with a total contribution in the interval [-1,1]. The overall dependency score d_{ij} consequently scales in the interval [-1,2].

Graph Characteristics

The previously described matrix represents a weighted functional dependency network in form of an undirected complete graph where each entry describes an edge weight in the network. An excerpt of this reference network is shown in Figure 3.2.

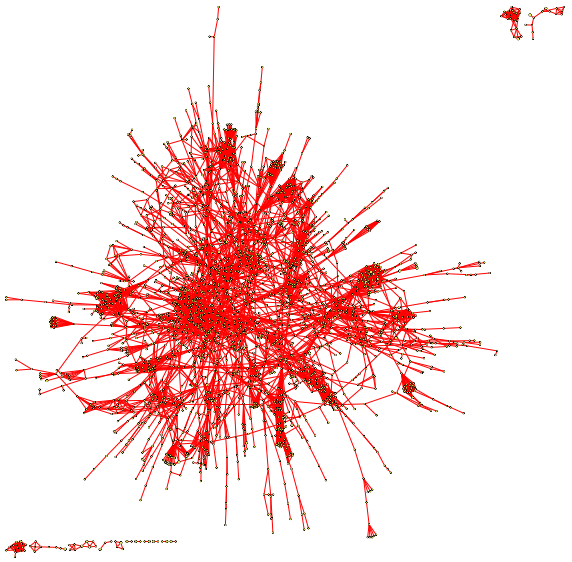


Figure 3.2: Visualization of an excerpt of the dependency graph. Nodes represent objects; only nodes with at least one edge with a weight > 1.0 are shown. At this particular cutoff one main sub-graph is found, complemented by various small subgraphs. The graph view was generated by Cytoscape using the organic layout option [59, 60].

The cutoff for edge weights can be varied in discrete steps, thus enabling the analysis of subnetworks, only including edges, having a higher weight than the introduced cutoff. Figure 3.3 shows characteristics of this reference network, namely number of edges, number of vertices, number of subgraphs, as well as the Index of Aggregation,

depending on the cutoff.

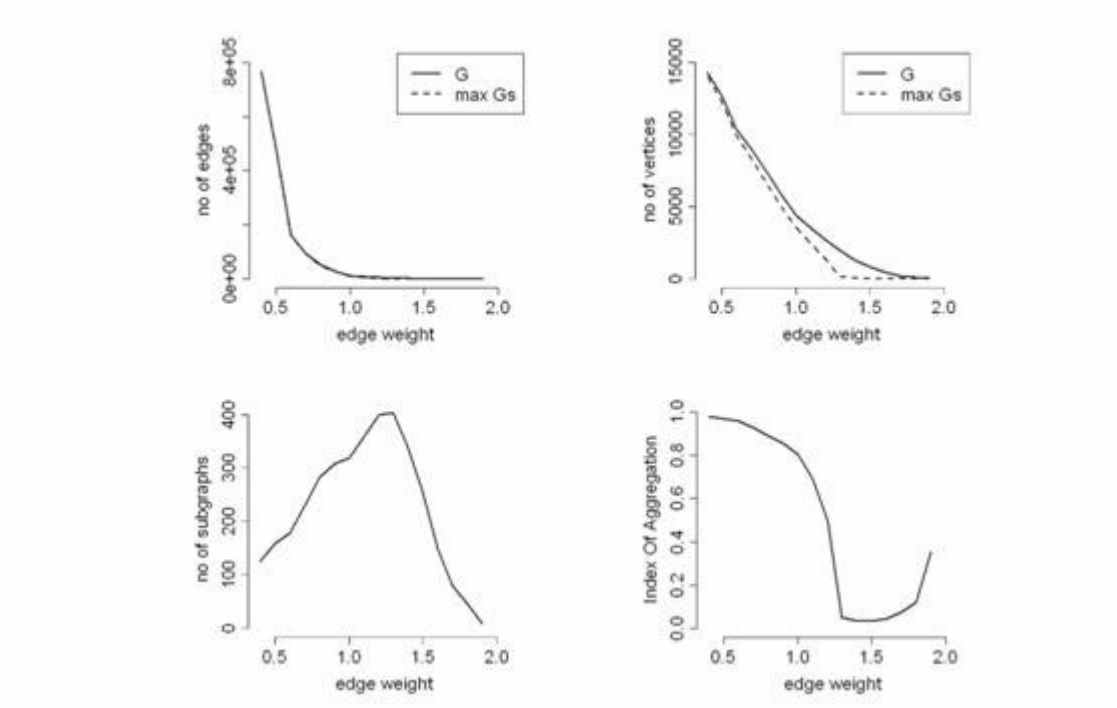


Figure 3.3: Graph characteristics were calculated for defined cutoffs, ranging from 0.5 to 2 for the complete graph (G) as well as for the largest subgraph (maxGs). (A) Shows the number of edges, (B) the number of vertices, (C) the number of subgraphs and (D) the Index of Aggregation.

The Index of Aggregation (IoA) defines the ratio of the total number of vertices in a subgraph to the total number of all given vertices in the graph [61].

In contrast to a sequential analysis this integrated procedure allows the complementation of a core set by genes that were initially considered as insignificant in a statistical sense. However, not only differences in transcript abundance have impact on cellular processes. This dependency graph approach was applied on the LIV/CAD dataset, aimed to identify functionally relevant objects and subgraphs.

RESULTS AND DISCUSSION

The following sections provide results derived by applying the sequential analysis workflow in context of the animal model of PHN, and the dependency graph approach applied on the LIV/CAD dataset. Finally the results are discussed.

4.1 PHN Dataset

4.1.1 Results

The preprocessed raw data, comprising information of 20 arrays, each holding expression values of 15924 transcripts, was further analyzed to identify differentially regulated genes in healthy and diseased rats.

Clustering

After preprocessing, the 20 arrays underwent a hierarchical clustering step, resulting in the pattern represented as a dendrogram in figure 4.1.

As shown in figure 4.1 the control and the diseased group form almost two separate clusters, with the exception of samples 1Pd3, 7Pd3 and 10Cd3 and 5Pd3 respectively. The latter two cluster separately. Furthermore, arrays of the d3 group cluster together very closely, indicating that the expression patterns are similar, and the same can be observed within group d6. This is due to the fact that the development of subepithelial immune deposits in the glomerular capillary walls and proteinuria in PHN commences 5 to 7 days after injection of the antibody.

Statistical Analysis

The preprocessed data were statistically analyzed using two different methods, namely maxT adjustment and SAM analysis. A description of the results and a comparison of the methods is given in the next sections.

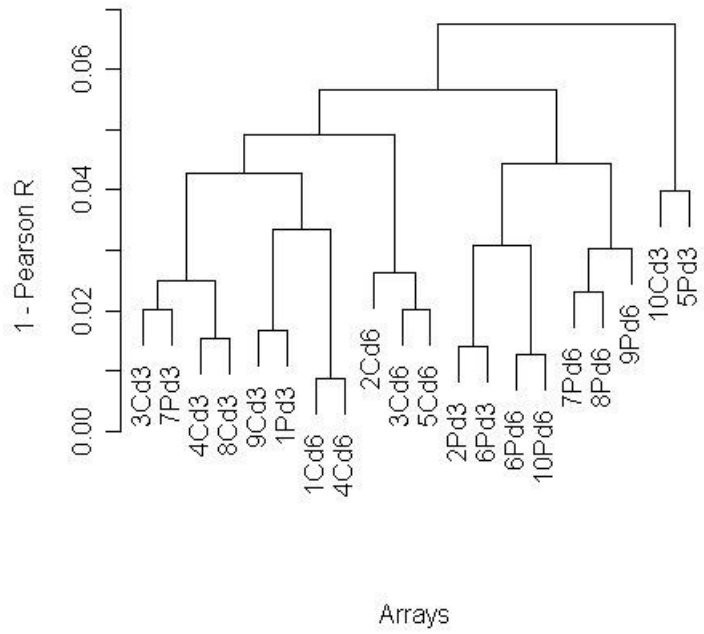


Figure 4.1: The dendrogram shows the clustering results for gene expression profiles in glomeruli of healthy and Fx1A induced rats (C and P respectively), 3 as well as 6 days (3d and 6d respectively) after injection.

maxT Adjustment By setting the overall p-value to 0.05 (indicating that at least 5% of the observed differences in expression between genes of the diseased and the control group occured by chance), the t-test adjusted by the maxT method revealed 33 unique genes in case of group d3 and 186 unique genes in case of group d6, which were differentially regulated in the control and the diseased glomeruli probes.

SAM Analysis SAM analysis was performed setting the false discovery rate to $<5\%$, resulting in 108 and 580 unique differentially expressed genes in group d3 and d6 respectively. Figure 4.2(a) and figure 4.2(b) show the scatter plots of the observed relative difference d(i) vs. the expected relative difference dE(i) for the two time points.

All genes but one revealed by the t-test with maxT adjustment were also found in the data set resulting from SAM analysis. This is true for both groups d3 and d6, except the one before mentioned gene which is missing in the SAM data set for d3.

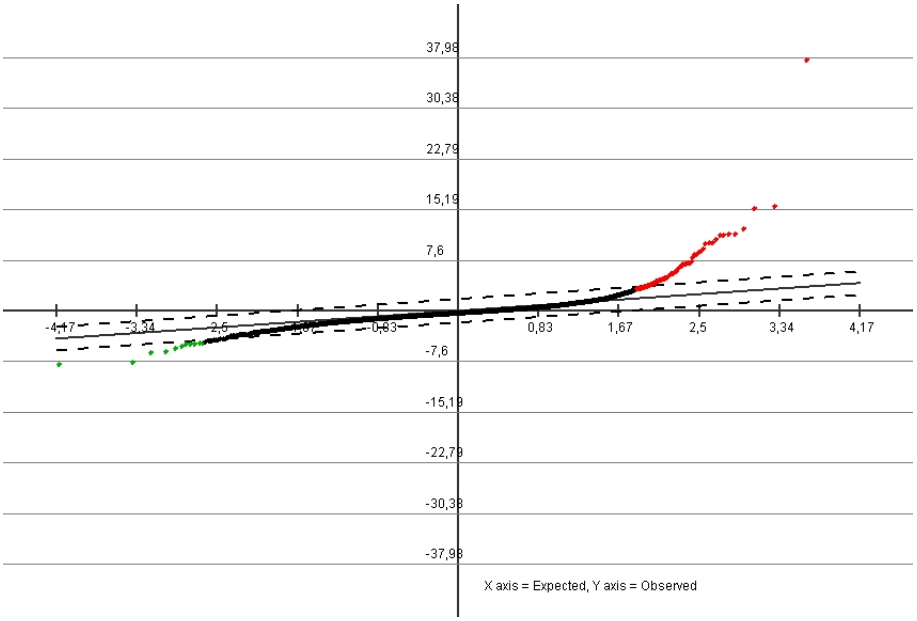
The focus of further analysis steps was on the genes identified as significant after SAM analysis. Genes showing a fold change <1.5 were not cosidered as significantly differentially expressed and were filtered out, resulting in two final sets, holding 54 genes in case of group d3 and 226 genes in case of group d6. All but 9 genes upregulated three days after injection of Fx1A were also expressed significantly differential six days after treatment. Altogether, 235 unique genes (PHN core set), all of them being upregulated in diseased rats, could be identified as differentially expressed when comparing control and immunized rats.

Functional Analysis

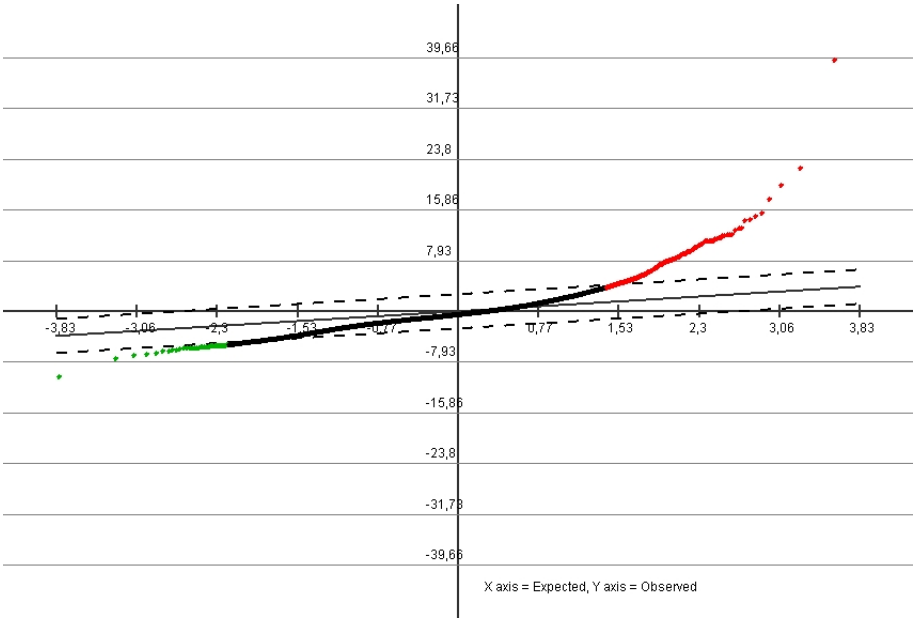
The 235 differentially regulated genes were categorized according to the biological process (GO term) they are involved in, as listed in tables 4.1 - 4.8. This categorization is not exclusive since some of the identified genes play roles in more than one process.

As can be seen in tables 4.1 and 4.2, a multitude of differentially regulated genes encode structural proteins of the cytoskeleton or proteins involved in cell adhesion, including integrin beta 1 (Itgb1), secreted phosphoprotein 1 (SPP1), lamin A (Lmna), desmin (Des), nestin (Nes), or tubulin beta 2 (Tubb2). Another strongly represented process is cell cycle. This category holds several genes encoding prominent proteins like the transforming growth factors beta 2 and 3 (Tgfb2, Tgfb3), the two cyclins B1 and B2 (Ccnb1, Ccnb2), or the proliferating cell nuclear antigen (Pcna).

The mentioned categories not only show an absolute enrichment of identified genes but are also significantly overrepresented in at least one time point in comparison to the PANTHER reference data set of the whole rat genome. All over/underrepresented processes are listed in table 4.9. Figure 4.3(a) and 4.3(b) respectively demonstrate the number of genes assigned to a certain category compared to all genes categorized to an overrepresented process.



(a) 3 days after immunization



(b) 6 days after immunization

Figure 4.2: Scatterplot of the SAM analysis : The two dotted lines represent the region within +/- delta units (set to 1.8) from the observed = expected line. The genes whose plot values are represented by black dots are considered non-significant, those colored red are significantly upregulated, and the green ones are significantly downregulated.

	Symbol	Gene Name	Fold Change	
			Day 3	Day 6
Apoptosis				
BE112895	Pea15 *	phosphoprotein enriched in astrocytes 15		2.55
AI411997	Adamtsl4	ADAMTS-like 4		1.9
NM_012935	Cryab	crystallin, alpha B		1.75
NM_022546	Dapk3	death-associated protein kinase 3	1.51	1.54
BF282636	RGD1305457	similar to RIKEN cDNA 1700023M03		1.54
NM_017180	Phlda1	pleckstrin homology-like domain, family A, member 1		1.5
Cell Adhesion				
Z78279	Col1a1	collagen, type 1, alpha 1	7.68	14.09
AF056034	Nexn	nexilin	4.77	4.55
BI303379	Tnfrsf12	tumor necrosis factor receptor superfamily, member 12a	3.7	4.21
BF407194	Itgb1bp2 *	integrin beta 1 binding protein 2	2.14	2.09
NM_030828	Gpc1	glypican 1		2.02
NM_017022	Itgb1	integrin beta 1		1.82
NM_012811	Mfge8	milk fat globule-EGF factor 8 protein		1.79
BG379319	Tgfbi	transforming growth factor, beta induced		1.72
NM_022266	Ctgf	connective tissue growth factor		1.66
AW433888	Vcl *	vinculin		1.63
NM_133409	Ilk	integrin linked kinase	1.52	1.61
BI275904	Lims2 *	LIM and senescent cell antigen like domains 2	1.63	1.59
AI008975	LOC311772	similar to nidogen 2		1.57
NM_022523	Cd151	CD151 antigen		1.51
NM_019237	Pcolce	procollagen C-proteinase enhancer protein		1.5
AB001382	Spp1	secreted phosphoprotein 1	2.1	
NM_012774	Gpc3	glypican 3	1.87	
Cell Structure and Motility				
U22520	Cxcl10	chemokine (C-X-C motif) ligand 10		3.66
BE111697	Kif20a *	kinesin family member 20A		3.55
NM_022531	Des	desmin	1.85	3.29
NM_031005	Actn1	actinin, alpha 1	3	3.19
BI283060	Flna *	filamin, alpha	2.6	2.96
NM_019131	Tpm1	tropomyosin 1, alpha	3.33	2.9
BI279044	Myl9 *	myosin, light polypeptide 9, regulatory	5.58	2.81
AI179391	Enh	enigma homolog	2.04	2.58

Table 4.1: Differentially expressed genes in glomeruli of rats with induced PHN.

NM_017148	Csrp1	cysteine and glycine-rich protein 1	1.92	2.42
X03369	Tubb2b	tubulin, beta 2b		2.34
BI274903	RGD1305887 *	similar to RIKEN cDNA 2310057H16	1.8	2.27
AA012755	MGC109519	similar to tropomyosin 1, embryonic fibroblast - rat		2.23
NM_019361	Arc	activity regulated cytoskeletal-associated protein		2.18
NM_012987	Nes	nestin		2.07
AI598442	RGD1564875 *	similar to mKIAA0613 protein		2.04
AW919109	Cap2	CAP, adenylate cyclase-associated protein, 2 (S. cerevisiae)	1.88	2.03
BI285440	Tubb5	tubulin, beta 5		1.91
AI103106	Lmnb1	lamin B1		1.86
AA892044	Tubb2	tubulin, beta, 2		1.8
NM_013194	Myh9	myosin, heavy polypeptide 9		1.76
NM_031970	Hspb1	heat shock 27kDa protein 1		1.74
BG381583	RGD1565118 *	similar to mKIAA0843 protein		1.68
X70706	Pls3	plastin 3 (T-isoform)		1.67
BM391364	LOC290704		1.6	1.67
BI296011	Cfl2 *	cofilin 2, muscle		1.65
NM_021755	Lmna	lamin A		1.63
NM_130411	Coro1a	coronin, actin binding protein 1A		1.63
BI278813	Ckap4 *	cytoskeleton-associated protein 4		1.6
BM383953	LOC367171	microtubule-associated protein 4		1.57
NM_031675	Actn4	actinin alpha 4	1.51	1.56
AA891834	Col4a5 *	collagen, type IV, alpha 5		1.56
AI407239	Myom2	myomesin 2		1.55
NM_031140	Vim	vimentin		1.54
NM_134452	Col5a1	collagen, type V, alpha 1		1.52
AI180161	Mapre1	microtubule-associated protein, RP/EB family, member 1		1.52
AW252250	Nebl *	nebulette		1.51
NM_019212	Acta1	actin, alpha 1, skeletal muscle	6.03	
Cell Cycle				
BG379338	Rrm2	ribonucleotide reductase M2		6.87
BE113362	Cdkn3 *	cyclin-dependent kinase inhibitor 3		5.53
AI409259	Racgap1 *	Rac GTPase-activating protein 1		5.02
AA944180	RGD1562047 *	similar to Cyclin-dependent kinases regulatory subunit 2 (CKS-2)		5
NM_019296	Cdc2a	cell division cycle 2 homolog A (S. pombe)		4.4
AW253821	Ccnb2 *	cyclin B2		3.47

Table 4.2: Differentially expressed genes in glomeruli of rats with induced PHN.

BI296084	Ube2c *	ubiquitin-conjugating enzyme E2C		3.15
NM_031131	Tgfb2	transforming growth factor, beta 2	2.06	3.06
NM_021989	Timp2	tissue inhibitor of metalloproteinase 2	1.58	2.44
BF417638	RGD:1359093	similar to cell division cycle associated 3		2.27
BG380355	Cdca8	cell division cycle associated 8		2.24
U05341	Cdc20	cell division cycle 20 homolog (S. cerevisiae)		2.2
BE117002	LOC362021			2.12
AI408269	Spbc25	spindle pole body component 25 homolog (S. cerevisiae)		2.06
NM_053483	Kpna2	karyopherin (importin) alpha 2		2.04
AA874827	Dlg7 *	discs, large homolog 7 (Drosophila)		1.84
AA996882	Stk6	serine/threonine kinase 6		1.82
AW920000	LOC362587	similar to microfilament and actin filament cross-linker protein isoform b		1.81
NM_022381	Pcna	proliferating cell nuclear antigen		1.78
X64589	Ccnb1	cyclin B1		1.72
AI407985	LOC686524	hypothetical protein LOC686524		1.7
AF140232	S100a6	S100 calcium binding protein A6 (calcyclin)		1.69
BM386384	Nap1l1	nucleosome assembly protein 1-like 1		1.65
NM_053819	Timp1	Tissue inhibitor of metalloproteinase 1		2.3
AA957183	Cit	citron		1.57
NM_013174	Tgfb3	transforming growth factor, beta 3		1.55
Immunity and Defense				
AI233530	C1qtnf3 *	C1q and tumor necrosis factor related protein 3		2.67
BI284441	Colec12	collectin sub-family member 12		2.59
BI278802	Prnp	prion protein		2.51
BF389535	LOC299339	similar to Tumor necrosis factor, alpha-induced protein 2		2.14
NM_053843	Fcgr3	Fc receptor, IgG, low affinity III		2.1
AI176519	Ier3	immediate early response 3		2.08
L12458	Lyz	lysozyme		2.05
AW918311	C1qtnf4 *	C1q and tumor necrosis factor related protein 4		2.02
NM_012620	Serpine1	serine (or cysteine) proteinase inhibitor, clade E, member 1		1.66
AI228623	Nptx2 *	neuronal pentraxin II		1.64
NM_031971	Hspa1a	heat shock 70kD protein 1A		1.54
Transport (membrane)				
BI293600	Slc35b2	solute carrier family 35, member B2	2.08	2.64
NM_019354	Ucp2	uncoupling protein 2		1.5

Table 4.3: Differentially expressed genes in glomeruli of rats with induced PHN.

Transport (intracellular)				
NM_022959	Pamci	peptidylglycine alpha-amidating monooxygenase COOH-terminal interactor		1.57
BG381589	Stx6	syntaxin 6		1.5
Transport (other)				
AI170609	RGD1560252 *	similar to hypothetical protein MGC31967	4.02	3.1
AI407838	Ecm1	extracellular matrix protein 1		2.04
NM_022278	Glrx1	glutaredoxin 1 (thioltransferase)		1.52
BE111722	Ndufs2 *	NADH dehydrogenase (ubiquinone) Fe-S protein 2		1.52
Metabolism (lipid)				
NM_031043	Gyg1	glycogenin 1	1.59	2.64
NM_013200	Cpt1b	carnitine palmitoyltransferase 1b	1.6	2.43
AF248543	A3galt2	alpha 1,3-galactosyltransferase 2 (isoglobotriaosylceramide synthase)		1.74
NM_017235	Hsd17b7	hydroxysteroid (17-beta) dehydrogenase 7		1.67
NM_012941	Cyp51	cytochrome P450, subfamily 51		1.67
NM_031840	Fdps	farensyl diphosphate synthase		1.51
Metabolism (glycogen)				
AW919180	Pygm	muscle glycogen phosphorylase	1.85	1.76
Metabolism (other)				
NM_022674	H2afz	H2A histone family, member Z		1.61
AA891760	RGD1308350 *	similar to hypothetical protein MGC13251		1.68
BG381486	Large *	like-glycosyltransferase		1.64
BM385390	Uxs1	UDP-glucuronate decarboxylase 1		1.58
BI282076	Prdx4	peroxiredoxin 4		1.52
Protein Modification				
AI236997	Dusp14 *	dual specificity phosphatase 14		2.24
NM_130403	Ppp1r14a	protein phosphatase 1, regulatory (inhibitor) subunit 14A		1.82
AW531714	Ube2t *	ubiquitin-conjugating enzyme E2T (putative)		1.76
BI276525	Ate1 *	arginine-tRNA-protein transferase 1		1.7
BI283703	Mapkapk2	MAP kinase-activated protein kinase 2		1.69
AF106659	Usp2	ubiquitin specific protease 2		1.68
NM_053323	Degs	degenerative spermatocyte homolog (Drosophila)		1.64
AI010241	Usp1 *	ubiquitin specific protease 1		1.63
AA799400	B3galt3 *	UDP-Gal:betaGlcNAc beta 1,3-galactosyltransferase, polypeptide 3		1.59
AA849399	Ctsz	cathepsin Z		1.54
BI279788	Ube2s *	ubiquitin-conjugating enzyme E2S		1.51
NM_024135	Limk2	LIM motif-containing protein kinase 2		1.5

Table 4.4: Differentially expressed genes in glomeruli of rats with induced PHN.

Protein Folding				
AI175031	Dnajb4 *	DnaJ (Hsp40) homolog, subfamily B, member 4		1.53
BG671521	Hspca	heat shock protein 1, alpha		1.54
Signal Transduction				
NM_019904	Lgals1	lectin, galactose binding, soluble 1	1.61	3.71
NM_012715	Adm	adrenomedullin	1.78	2.86
BF405151	Gpr39 *	G protein-coupled receptor 39	1.72	2.8
NM_053634	Fcnb	ficolin B	1.91	2.7
NM_033099	Ptprv	protein tyrosine phosphatase, receptor type, V		2.27
BE117002	RGD1560967 *	similar to Pins		2.12
AW253242	Magi1 *	membrane associated guanylate kinase interacting protein-like 1		1.74
X78595	Npr3	natriuretic peptide receptor 3		1.74
BG378926	S100a11	S100 calcium binding protein A11 (calizzarin)		1.73
BI276015	RGD1559882 *	similar to hypothetical protein E130310N06		1.66
BM386204	Ran	RAN, member RAS oncogene family		1.65
BI295991	Rab2l	RAB2, member RAS oncogene family-like		1.6
NM_022236	Pde10a	phosphodiesterase 10A	1.56	1.56
NM_012823	Anxa3	annexin A3		1.51
NM_053299	Ubd	ubiquitin D	1.72	
M35297	Mrgprf	MAS-related GPR, member F	1.79	
Transcription				
BM385445	Top2a	topoisomerase (DNA) 2 alpha		3.37
L81174	Ankrd1	ankyrin repeat domain 1 (cardiac muscle)	2.62	3.02
NM_031628	Nr4a3	nuclear receptor subfamily 4, group A, member 3		2.34
NM_017187	Hmgb2	high mobility group box 2		2.04
NM_131902	Cdkn2c	cyclin-dependent kinase inhibitor 2C (p18, inhibits CDK4)		2.01
NM_017365	Pdlim1	PDZ and LIM domain 1 (elfin)		2
AI170362	Nfkb2	nuclear factor of kappa light polypeptide gene enhancer in B-cells 2, p49/p100		1.95
U17565	Mcmd6	mini chromosome maintenance deficient 6 (S. cerevisiae)		1.88
NM_053583	Zfp423	zinc finger protein 423		1.87
BM387864	Lrrfip1	similar to FLI-LRR associated protein-1		1.76
BG664147	Ptrf *	polymerase I and transcript release factor		1.65
BG380385	Srf *	serum response factor	1.95	1.64
BF403027	Hdac5	histone deacetylase 5		1.64
AI179264	Creb3	cAMP responsive element binding protein 3		1.5

Table 4.5: Differentially expressed genes in glomeruli of rats with induced PHN.

Homeostasis				
BI285437	Nxn *	nucleoredoxin		1.62
Nucleus				
BE104102	RGD1306774 *	similar to SPT3-associated factor 42		1.83
Membrane				
BM385031	Plp2	proteolipid protein 2	1.76	2.55
BM388441	RGD1311946 *	similar to RIKEN cDNA 1810055G02		1.97
NM_030847	Emp3	epithelial membrane protein 3		1.89
AW917760	RGD1564216 *	similar to Myoferlin (Fer-1 like protein 3)		1.88
AI009530	MGC72614	hypothetical LOC310540		1.86
BI296048	Myadm	myeloid-associated differentiation marker	1.55	1.77
J03867	Dia1	diaphorase 1		1.6
BM385463	Tmem43	transmembrane protein 43		1.55
AI230273	RGD:735199	Unknown (protein for MGC:72987)		1.55
BF283798	Nipsnap3a	nipsnap homolog 3A (C. elegans)		1.53
AA850909	Pvrl2 *	poliovirus receptor-related 2 (herpesvirus entry mediator B)		1.5
BI290029	RGD1562920 *	similar to Aig1 protein		1.5
AI407016	RGD1307736 *	similar to Hypothetical protein KIAA0152	1.55	
BI294974	Ldlr	low density lipoprotein receptor	1.69	
Developmental Processes				
NM_031549	Tagln	transgelin	69.77	38.66
NM_012636	Pthlh	parathyroid hormone-like peptide		3.13
NM_030584	Sost	sclerostin		3.12
AW141680	Bmp6	bone morphogenetic protein 6		2.36
NM_019242	Ifrd1	interferon-related developmental regulator 1	2.05	2.32
AW251450	Mustn1	musculoskeletal, embryonic nuclear protein 1		2.13
AI235465	Ssg1	steroid sensitive gene 1		2.07
AW435036	Smtn *	smoothelin		1.65
BI290551	Fnbp1	Formin binding protein 1		1.6
BI275485	Sema3b *	semaphorin 3B, immunoglobulin domain, secreted		1.57
AW144216	Enpep	glutamyl aminopeptidase	2.11	1.57
BG666787	Gmfg	glia maturation factor, gamma		1.56
BM384088	Socs2	suppressor of cytokine signaling 2		1.54
NM_031114	S100a10	S100 calcium binding protein A10 (calpactin)		1.51
Other				
AI229404	RGD1566097 *	similar to Anillin	2.47	8.41

Table 4.6: Differentially expressed genes in glomeruli of rats with induced PHN.

BI295828			2.28	3.31
BI279587			2.12	3.01
BI283695			1.73	2.48
AW531909				2.44
BF419834			1.62	2.29
BF415061	RGD1307034 *	similar to hypothetical protein CG003		2.21
BF408518	RGD1305081 *	similar to ionized calcium binding adapter molecule 2 (Iba2)	1.98	2.16
AI712694	RGD1308747 *	similar to hypothetical protein FLJ10156		2.15
BI296728	RGD1564957 *	similar to RIKEN cDNA 3110007P09		2.04
AI176172				2.03
BM387112			1.71	2.03
AI071000				2
AA799328	RGD1560913 *	similar to expressed sequence AW413625		1.93
BE096535		transcribed locus, strongly similar to XP_574462.1 similar to hypoth. protein C230069C04		1.89
BG378155	RGD1565079 *	similar to hypothetical protein MGC17839		1.88
AA943808	RGD1307215	similar to protein phosphatase 1, inhibitory subunit 1C; thymocyte ARPP		1.82
AW143197				1.79
AW529960				1.78
AI177743	LOC498261			1.72
AI317841	Gramd3	GRAM domain containing 3		1.69
BI303106				1.64
BF561368	RGD1306959 *	similar to C11orf17 protein		1.63
AW253004		CDNA clone IMAGE:7317367		1.62
BF398756				1.62
AI009167	Zfp451 *	zinc finger protein 451		1.62
AI412389				1.61
BE111057				1.6
BI282694	RGD1565037 *	similar to selenoprotein SelM		1.6
AI231225				1.58
AA942716	Hn1	hematological and neurological expressed sequence 1		1.56
BF284519				1.55
BG671786				1.53
AW914928				1.53
AI113146	Acpl2	acid phosphatase-like 2		1.51
AI170820	RGD1310383 *	similar to T-cell activation protein phosphatase 2C		1.5
AA800892	RGD1563599 *	similar to putative SH3BGR protein	1.81	

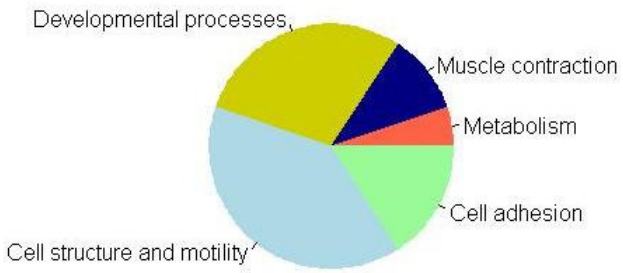
Table 4.7: Differentially expressed genes in glomeruli of rats with induced PHN.

BG380430	RGD1564105 *	similar to RIKEN cDNA B130052G07	1.54	
NM_021584	Ania4	activity and neurotransmitter-induced early gene protein 4 (ania-4)		2.53
AA997359	Serpinb6	serine (or cysteine) peptidase inhibitor, clade B, member 6		1.54
NM_012618	S100a4	S100 calcium-binding protein A4		1.8
NM_022382	Pde4dip	phosphodiesterase 4D interacting protein (myomegalin)		1.7
AI112962	Rcn *	reticulocalbin		1.92
AI232065	Arhgap18 *	Rho GTPase activating protein 18		1.55

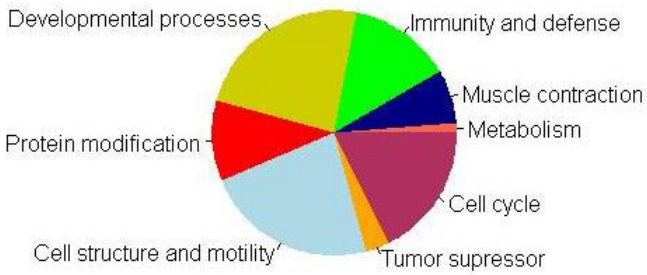
Table 4.8: Differentially expressed genes in glomeruli of rats with induced PHN.

	Day 3			Day 6		
Biological Processes	**Number of Genes**	**over/under**	**p-Value**	**Number of Genes**	**over/under**	**p-Value**
Cell structure and motility	15	+	4,20E-010	35	+	1,02E-013
-Cell structure	12	+	1,33E-009	27	+	3,08E-013
-Cell motility	6	+	3,74E-005	17	+	9,07E-010
Developmental processes	11	+	1,46E-003	36	+	9,29E-007
-Mesoderm development	4	/	/	15	+	2,68E-005
Cell cycle	3	/	/	27	+	2,29E-010
-Mitosis	3	/	/	10	+	2,53E-004
-Cell cycle control	1	/	/	10	+	1,01E-003
Muscle contraction	4	+	4,29E-004	11	+	2,30E-007
Immunity and defense	2	/	/	21	+	4,34E-004
-Macrophage-mediated immunity	0	/	/	5	+	1,81E-003
-Stress response	0	/	/	8	+	1,86E-004
Tumor supressor	1	/	/	5	+	4,73E-004
Metabolism			/			
- sulfur redox	0	/	/	2	+	1,23E-002
- glycogen	2	+	3,84E-003	2	/	/
Cell adhesion	6	+	6,59E-004	9	/	/
G-protein mediated signaling	1	/	/	2	-	1,11E-002
Protein modification	2	/	/	16	+	1,14E-002

Table 4.9: Biological processes that are significantly enriched or depleted in rats 3, as well as 6 days after induction of PHN.



(a) 3 days after immunization



(b) 6 days after immunization

Figure 4.3: Biological processes that are significantly enriched by genes of the PHN core set.

The overrepresented functional groups before onset of proteinuria (d3) diverge from those after proteinuria has commenced (d6). The relative amount of cell cycle associated genes is 21.3% of all differentially regulated genes after the onset of proteinuria compared to 5.6% before. An even larger change was found for the genes with a function in cellular immunity and defense. They are not enriched (3.7%) at day 3, but at day 6 they make up 16.7% of all abundant genes.

A total of 30 genes in case of the dataset of identified genes in group d3 and 82 genes in group d6 could be assigned to KEGG pathways. An overview of these pathways and number of involved genes is given in table 4.10 .

	Day 3	**Day 6**
Pathways	**Number of Genes**	**Number of Genes**
Focal adhesion	6	9
Cell cycle	1	9
Regulation of actin cytoskeleton	3	8
MAPK signaling pathway	3	7
Cell communication	2	7
Leukocyte transendothelial migration	3	5
Gap junction	2	5
axon guidance	0	5
Tight junction	3	4
Adherens junction	2	4
p53 signaling pathway	0	4
Adipocytokine signaling pathway	2	3
ECM-receptor interaction	1	3
TGF-beta signaling pathway	1	3
Cytokine-cytokine receptor interaction	1	2
Cell adhesion molecules	0	2
Toll-like receptor signaling pathway	0	2

Table 4.10: KEGG pathways and the number of genes according to each category.

Network Analysis

To detect coherences between genes of the PHN core set and physical interaction partners, initially not identified as differentially regulated, a protein network analysis

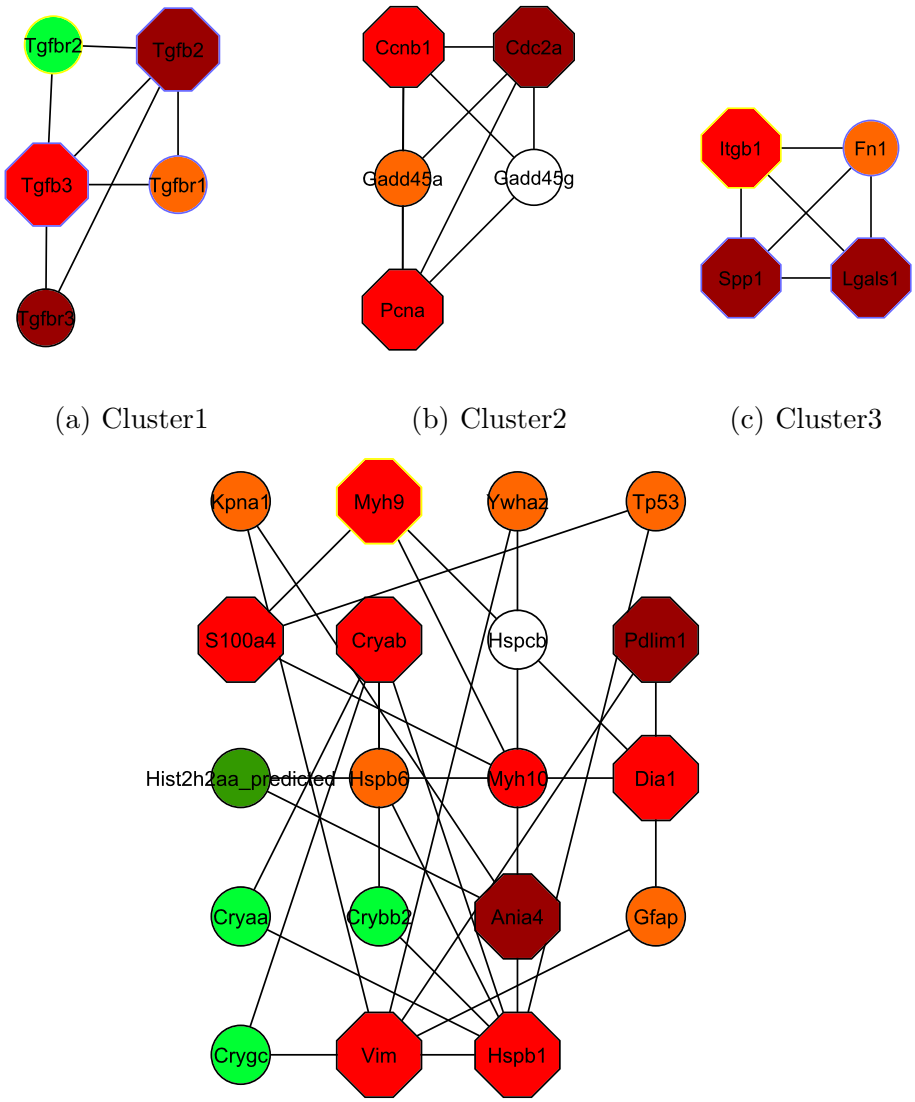
was performed. Based on information derived from the Online Predicted Human Interaction Database (OPHID), protein-protein interactions were visualized using the Cytoscape bioinformatic platform for visualizing molecular interaction networks. The network includes the genes of the PHN core set and their interaction partners, having at least 2 edges. To detect densely connected regions, the graph theoretic clustering algorithm "Molecular Complex Detection" (MCODE) was used. Running the algorithm with a k-Core of 2 and a maximal Depth of 100 resulted in detection of 4 clusters shown in figure 4.4.

Components of each of the 4 clusters can be assigned to a certain biological process. All genes in cluster1 play a role in cell cycle. Except of Lgals, genes in cluster2 are associated with cell adhesion. Cluster3 completely represents growth factors and most of the genes in cluster4 are involved in structure and motility.

4.1.2 Discussion

Passive Heyman nephritis (PHN) is a rat model of immune complex glomerular disease that closely resembles human membranous glomerulonephritis. PHN is induced by an injection of anti-Fx1A antibodies which are raised against rat proximal tubular brush-border antigens. Binding of the antibodies leads to activation of the complement cascade and insertion of the C5b-9 membrane attack complex into the glomerular basement membrane, accompanied by podocyte effacement and a conspicuous lack of inflammatory cells [20]. Within 5 to 7 days after injection, abnormal proteinuria occurs. These findings are consistent with the results of the clustering analysis, showing almost two separate clusters for arrays holding probes 3 and 6 days after immunization, respectively.

The statistical analysis procedure resulted in 235 significantly upregulated genes. Notably significant is the gene Tagln, also known as SM22, which shows a fold change of 69.77 at day 3 and 38.66 6 days after induction of PHN. The protein encoded by Tagln is transgelin, a cytoskeletal protein that is exclusively expressed in smooth muscle cells. It has been shown previously that Tagln is one of the genes highly expressed



(a) Cluster1 (b) Cluster2 (c) Cluster3

(d) Cluster4

Figure 4.4: Clusters detected by MCODE. Node colors indicate the measured fold change, green for under-expressed and red for over-expressed genes (orange <1,5, red >1,5, dark red >2) in rats with induced PHN when compared to healthy rats. Hexagons represent those genes that were considered to be significantly upregulated in previous analysis.

in kidneys with anti-glomerular basement membrane nephritis (caused by antibody response against antigens in the glomerular basement membrane), providing an indication that injured glomerular epithelial cells undergo structural alterations [62].

Furthermore, Tagln has been identified as a repressor of the matrix metalloproteinase MMP-9, a member of a family of neutral proteinases that can degrade extracellular matrix components [63]. MMP-9 is produced by podocytes, which therefore are able to enzymatically modify the glomerular basement membrane they adhere to. The high expression of Tagln could indicate the activation of an cellular emergency program, trying to prevent cell damage caused by MMP-9. Two further genes, identified to be upregulated in PHN, encode well known inhibitors of several members of the MMP family, namely the tissue inhibitors of metalloproteinases Timp1 and Timp2.

The dense cluster4 4.4(d), identified by network analysis, also reflects the structural changes and cytoskeletal rearrangements of the glomerulus as a result of the experimental injury. One of the interacting proteins is vimentin, a class-III intermediate filament of the cytoskeleton and usually found in mesenchymal tissues. The strong upregulation after the onset of proteinuria (d6) might suggest a phenotypic transformation of parts of the glomerulus tissue to a mesenchymal morphology as can be observed during diverse chronic kidney diseases [64].

Another gene in cluster4 is Myh9, which encodes for the nonmuscle myosin heavy chain protein myosin-IIA, a part of the actinomyosin complex. Previous studies demonstrated that myosin-IIA is involved in changes of cell morphology in many cell types [65].

The great number of differentially expressed genes in PHN induced rats, involved in cell structure and motility (Figure 4.3), suggest the critical relevance of an intact cytoskeleton.

Proteinuria emerges from a loss of the permeability barrier in the kidney. For an intact barrier, podocyte-podocyte as well as podocyte-glomerular basement membrane junctions are essential but usually damaged by changes in the extracellular matrix in

case of many nephropathies. Fibronectin1 (Fn1) for example is involved in mechanisms like cell adhesion and fibrosis and it was shown that its expression is induced by high-glucose in human renal cells [66], suggesting similarity between the PHN model and development of diabetic nephropathy. The gene is not differentially expressed in the PHN dataset but is part of cluster3 4.4(c).

One further interacting protein in cluster3 is osteopontin. The encoding gene is Spp1 which is upregulated before the onset of proteinuria. Osteopontin is expressed in distal tubular cells of the kidney and it is jointly responsible for urinary stone formation [67]. Furthermore, it has been shown that upregulation of osteospotin in the kidney influences monocyte migration into renal compartments and aggravates the immune response after the initial injury [68].

Another phenomenon in experimental membranous nephropathy is that C5-b9 induced injury to podocytes comes along with DNA synthesis but not cytokinesis. Previous studies demonstrated the role of the transforming growth factor Tgfb1 to limit DNA synthesis in renal cells. The PHN core set does not involve Tgfb1 but the isoforms Tgfb2 and Tgfb3, a fact that also seems to be indicated by Shankland et al. [69]. This leads to the assumption that Tgfb2 and Tgfb3 have different biological effects than Tgfb1 and might inhibit the proliferation of injured podocytes. The interaction network between Tgfb isoforms and their receptors represented as cluster1 in figure 4.4(a) additionally demonstrates their coherence with podocyte injury.

As indicated by the high expression levels of cyclins B1 and B2 (Ccnb1,Ccnb2) and the cell division cycle 2 homolog Cdc2a, all involved in the M-phase of cell cycle, podocytes have the ability to increase cell cycle proteins required for mitosis. The lack of proliferation might be due to a regulatory disturbance in cytokinesis [70].

Cell cycle arrest can be the consequence of upregulated Gadd45 (growth arrest and DNA damage inducible 45) which was shown to be expressed after sublytic injury activated by C5b-9 [71], suggesting that DNA damage occurs in the course of injury. The isoforms Gadd45a and Gadd45g, besides Ccnb2, Cdc2a and the proliferating cell nuclear antigen Pcna, are part of cluster2 4.4(b). Pcna also plays an important role in DNA repair and influences cell survival. The upregulation of Pcna 6 days after

immunization of the rats may be due to an activation of repair processes in the injured tissue [72].

The results of the sequential analysis of the PHN dataset clearly demonstrate the crucial role of an intact permeability barrier in the kidney, strongly associated with glomerulus and podocyte interaction. The main important involved biological processes are cell structure and motility, cell adhesion, cell cycle and immunity. Failure in podocyte structure or proliferation as a consequence of complement mediated injury is critical and may lead to further damage of the glomerulus and proteinuria.

4.2 LIV/CAD Dataset

The following section describes results and discusses usage of the dependency graph approach applied on 105 genes (LIV/CAD core set) initially identified as differentially regulated between living and cadaveric donor kidneys, as identified by a genome-wide expression analysis of 32 kidney biopsies [32].

4.2.1 Results

Out of 132 statistically significantly up- or downregulated transcripts, 105 could be functionally annotated and assigned to PANTHER biological processes. In particular, they belong to the functional categories proteolysis, immunity and defense/complement-mediated immunity and metabolism. Results derived from GenMAPP, MAPPFinder and HAPI emphasize the complement system as a critical part in differentiation between living and cadaveric donor kidneys. The most frequently found MeSH term was "Complement" and the only identified pathway by GenMAPP and MAPPFinder was again the complement system.

Among the 105 genes of the LIV/CAD core set, 96 could be mapped on our dependency graph. For interpreting the differentially regulated genes on the level of the graph we calculated those subgraphs showing edge weights $\geq$1.3 and holding at least one member of the core set. The resulting subgraphs consist of a total of 166 nodes, including 17 members (see Figure 4.5).

Although the number of differentially regulated genes found in the dependency graph when using an edge weight cutoff of 1.3 is rare, a further enrichment of genes according to initially overrepresented biological processes can be observed. 61 of the identified dependency graph objects are involved in immunity and defense while this group is represented by 18 in case of the set of 105 differentially expressed genes. A similar outcome can be found for the category proteolysis. In contrast, complement-mediated immunity lost significance. Table 4.11 gives an overview of the overrepresented processes.

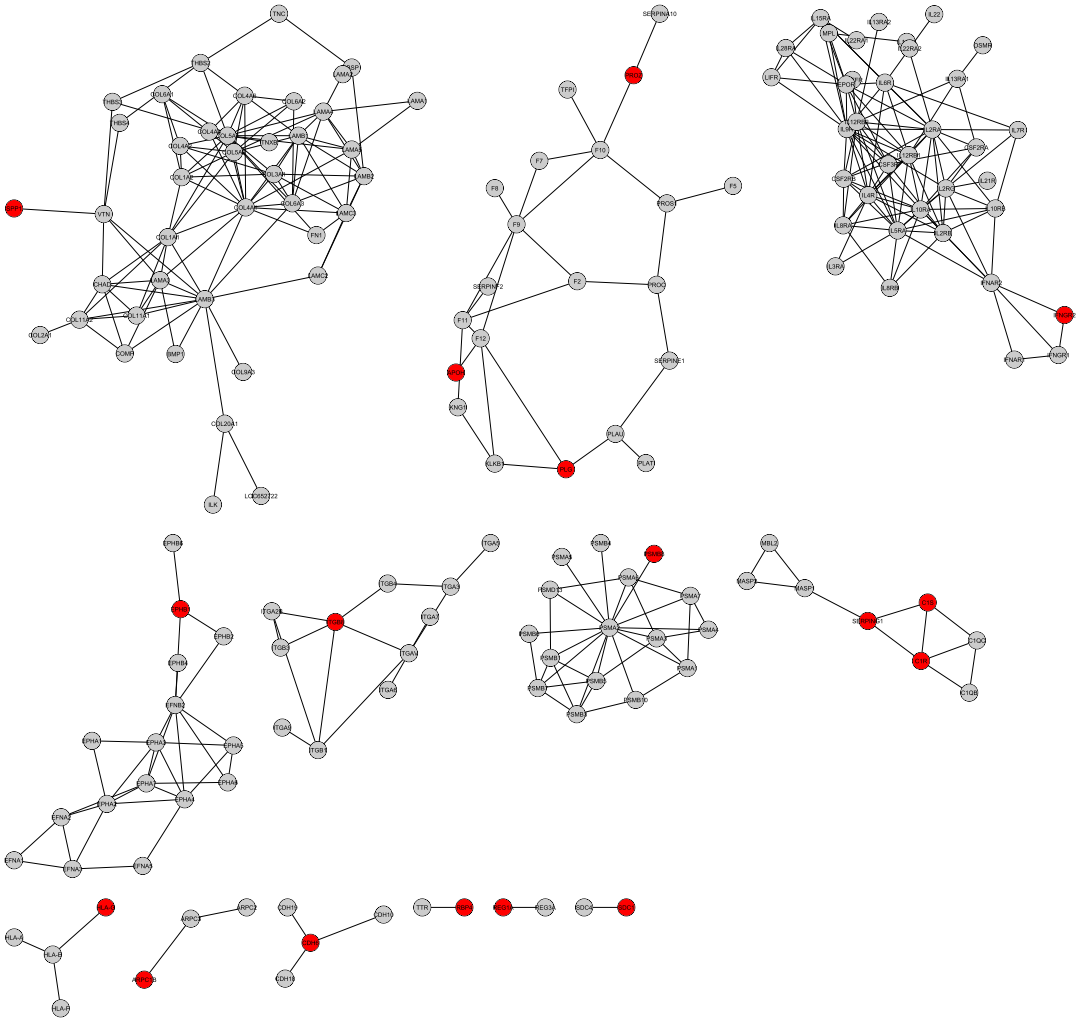


Figure 4.5: The figure shows subgraphs identified by introducing an edge weight cutoff of 1.3. Each subgraph holds at least one gene of the 105 differentially regulated genes found when comparing living and cadaveric donor kidneys on the level of gene expression (shown in red).

Statistical Analysis		Dependency Graph Analysis	
Biological process	**p-value**	**Biological Process**	**p-value**
Immunity and defense	1.33E-04	Immunity and defense	5.41E-34
Lipid metabolism	3.76E-04	Cytokine and chemokine mediated signaling pathway	5.68E-29
Proteolysis	8.96E-04	Signal transduction	1.76E-27
Coenzyme metabolism	2.31E-03	Blood Clotting	4.81E-27
Vitamin/cofactor transport	6.49E-03	Cell communication	5.42E-25
Amino acid metabolism	1.03E-02	Cell adhesion	2.88E-24
Complement-mediated Immunity	1.17E-02	Proteolysis	6.72E-20

Table 4.11: PANTHER biological processes and their significance of population expressed as p-value for statistical analysis of the LIV/CAD gene expression data and analysis in the context of the dependency graph.

In terms of the number of differentially regulated genes the resulting subgraphs are rather sparse. For further analysis the weighted shortest paths (Dijkstra algorithm [73]) between all pairs of the core set, showing at least one edge with a weight exceeding 1.0, were computed. Figure 4.6 displays the graph component, holding 32 genes of the LIV/CAD core set and 230 additional nodes that build the links between the members of the core set. Out of these 230 genes, 25.7% (59 genes) are involved in immunity and defense. Table 4.12 lists the main overrepresented biological processes resulting from the shortest path analysis.

4.2.2 Discussion

Kidneys from living donors hardly ever exhibit acute renal failure and display longer allograft survival compared to cadaveric organs [21]. Clinical observations have shown alterations in brain-dead donor organs that can cause organ injury, which has been suggested to alter the immunological or inflammatory status of the organ after transplantation [74]. The main findings of the sequential, statistical analysis workflow were increased activations of immunity and defense mechanisms in cadaveric donor kidneys compared to living transplants [32].

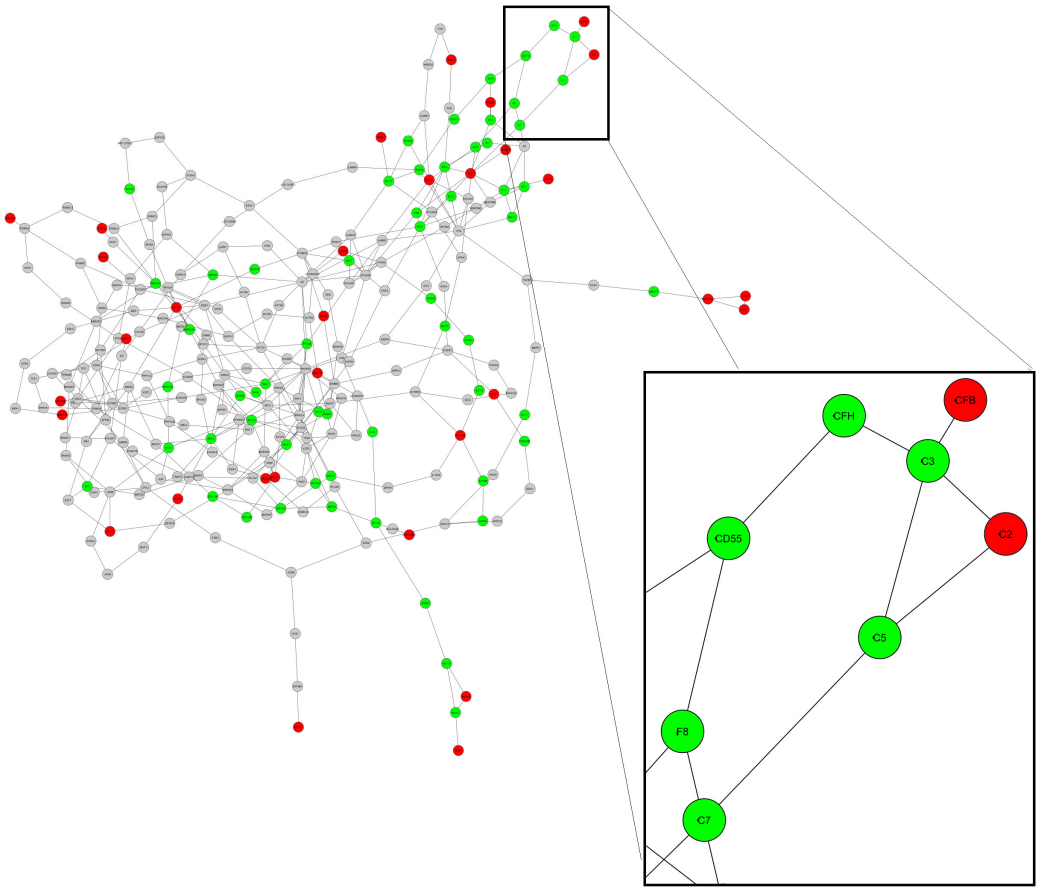


Figure 4.6: The figure displays the graph component consisting of the shortest paths between 32 members of the LIV/CAD core set. Red nodes represent members and green nodes additional genes associated with immunity and defense. The detailed graph area shows objects associated with complement-mediated immunity.

Shortest Paths Analysis	
Biological process	**p-value**
Immunity and defense	1.94E-35
Signal transduction	2.95E-32
Cell proliferation and differentiation	3.18E-32
Intracellular signaling cascade	6.04E-22
Developmental processes	4.39E-21
Protein phosphorylation	1.15E-20
Cell cycle control	2.62E-18

Table 4.12: Overrepresented PANTHER biological processes for genes identified by shortest path analysis of the LIV/CAD gene expression data.

The aim of the dependency graph approach was to identify additional and relevant genes previously not considered by statistical analysis but linked to members of this core set. Following the construction principle of the dependency graph, interlinking paths may indicate functional dependencies between objects. Indeed, a further enrichment of genes involved in corresponding biological processes could be identified.

Previous studies have implicated complement activation in the pathogenesis of Ischemia reperfusion injury (I/R) in the kidney which is a common cause of acute renal failure and impacts short- as well as long-term graft survival after kidney transplantation [75]. The subgraph, consisting of the shortest paths between 32 members of the LIV/CAD core set, includes 10 components of the complement system. Early components of the complement cascade namely C1r, C1s, C2, and factor B (CFB) are upregulated in cadaveric donor kidneys. C3, C5, C7, the complement factor H (CFH), the decay accelerating factor for complement CD55 and the mannan-binding lectin serine peptidase MASP1 could be found by the shortest path approach.

For example, shortest paths between CFB and any of the remaining 31 LIV/CAD core set members in the graph include C3, indicating a strong dependency of the two objects. A similar network topology led to the detection of C5 as a relevant gene in context of the core set of differentially expressed genes (see figure 4.6). In a model of

renal I/R injury by Zhou et al. [76], C3 and C5 deficient mice were protected from renal failure. Furthermore, this study showed that mice with isolated deficiency of terminal pathway activation (C6) exhibited a similar degree of protection to those with more proximal interruption of the complement cascade (early complement component C3 and intermediate component C5). This leads to the assumption, that formation of the membrane attack complex ,the end product of the complement cascade to which C6 and C7 contribute, is a critical mechanism through which complement mediates renal postischemic injury.

An immediate neighbor of the complement component C3 is CFH, a regulator of complement activation. Alexander and colleagues performed renal transplants between wildtype and CFH deficient mice and showed that CFH prevents the generation of proinflammatory complement activation products [77]. Similar outcome could be found when analyzing mice that lack a functional CD55, another complement regulatory protein and neighbor of CFH [78].

Two central nodes in the analysed graph component represent the transcription factor NFKB1 and the glucocorticoid receptor NR3C1. Both of these proteins are targets for steroids as prednisolone for reducing systemic inflammation. Renal grafts from marginal donors treated with prednisolone demonstrated a significantly improved graft function [79]. At present Oberbauer et al. evaluate the impact of steroid treatment of cadaveric donor kidneys on graft survival [80].

Among the LIV/CAD core set, holding 105 differentially regulated genes in living and cadaveric donor kidneys, 96 genes could be mapped on the dependency graph. Although the graph does not include all known proteins and is incomplete in terms of splice variants, the interpretation of the LIV/CAD core set on the level of the dependency graph offers further insights into the ongoing biological processes. The results reinforce the findings that activation of the inflammation cascade is a main cause of delayed graft function. Furthermore, they support donor immunosuppression with steroids as an indication for therapeutic strategies. By all means, the depen-

dency approach has the potential for triggering subsequent functional analysis on the experimental level.

CONCLUSION AND OUTLOOK

This thesis presents two complementary analysis procedures of gene expression data. Both approaches make use of statistical and functional classification methods, but also provide aspects of protein-protein interactions. As input, data derived from experimental analysis of PHN induced rat glomeruli, as well as of transplant donor kidney biopsies was used.

The PHN dataset was analyzed following the sequential analysis workflow. The procedure involved data pre-processing, clustering, statistical, functional, and protein network analysis and resulted in a gene expression pattern that mainly goes inline with previous work. Major findings were upregulated genes involved in cell structure, cell motility, cell cycle, and in immunity and defense.

Transgelin, also known as SM22, showed the highest expression value associated with PHN. The protein was found to be a repressor of the matrix metalloproteinase 9 (Mmp9) that is jointly responsible for degradation of extracellular matrix components. Further inhibitors of metalloproteinases, Timp1 and Timp2, as well as Myh9 which is involved in changes of cell morphology were shown to be upregulated in immunized rats. Expression levels of cyclines as Ccnb1 or Ccnb2 and growthfactors like Tgfb also separate healthy from diseased rats, thereby indicating a dysregulation of podocyte proliferation.

The analysis of physical interactions between members of the core set of differentially expressed genes supports the hypothesis of the critical relevance of the before mentioned biological processes. The four identified clusters describe the main processes in the glomerulus during the early phase of PHN that will lead to proteinuria. These four processes are: (i) DNA damage and repair as seen in cluster 1; (ii) Changes in the extracellular matrix; (iii) Deregulation of cytokines and growth factors and (iv) Re-arrangement of the cytoskeleton.

Subject of the integrated analysis workflow was the mapping of genes previously identified as differentially regulated between living and cadaveric renal transplants, on the level of a dependency graph. Dependencies between the objects were determined by four parameters, namely protein interaction, gene expression, function, and subcellular location. Calculating shortest paths between members of the LIV/CAD core set led to an extended set of genes potentially indicative for the development of ARF and delayed graft function of transplants from cadaveric donors.

Most important results relate to members and regulators of the complement cascade. The formation of a membrane attack complex and subsequent cellular lysis can lead to ischemia reperfusion injury, a major cause of ARF. Another interesting finding was the detection of strong functional dependencies of genes of the LIV/CAD core set and targets for steroids, namely the transcription factor NFKB1 and the glucocorticoid receptor NR3C1. Treatment of marginal donors with steroids is a clinical approach to improve graft function.

In conclusion, an analysis of gene expression data on the level of a dependency graph presents an improved expansion of the traditional sequential analysis workflow in terms of deciphering functional relationships. Compared to solely statistical procedures, the dependency graph approach follows the principal of integrating data from different levels of observation, thereby enabling the elucidation of the overall state of cells. These procedure promises the identification of potential biomarkers worth of further experimental verification.

APPENDIX

REFERENCES

[1] K. Aggarwal and K.H. Lee. Functional genomics and proteomics as a foundation for systems biology. *Brief Funct Genomic Proteomic.*, 2(3):175–184, 2003.

[2] K. Strange. The end of "naive reductionism": rise of systems biology or renaissance of physiology? *Am J Physiol Cell Physiol.*, 288(5):C968–C974, 2005.

[3] B. Boeckmann, A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic Acids Research*, 31(1):365–370, 2003.

[4] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic Acids Res.*, 28(1):235–242, 2000.

[5] K.D. Pruitt, T. Tatusova, and D.R. Maglott. Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 35(Database Issue):D61–D65, 2007.

[6] P.D. Thomas, M.J. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, K. Diemer, A. Muruganujan, and A. Narechania. Panther: a library of protein families and subfamilies indexed by function. *Genome Res*, 13(9):2129–2141, 2003.

[7] M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, 2000.

[8] S.P. Shah, Y. Huang, T. Xu, M.M. Yuen, J. Ling, and B.F. Ouellette. Atlas - a data warehouse for integrative bioinformatics. *BMC Bioinformatics.*, 6:34, 2005.

[9] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet, and Y. Moreau. Gene prioritization through genomic data fusion. *Nat Biotechnol.*, 24(5):537–544, 2006.

[10] K.R. Brown and I. Jurisica. Online predicted human interaction database. *Bioinformatic*, 21(9):2076–2082, 2005.

[11] C. Alfarano, C.E. Andrade, K. Anthony, N. Bahroos, M. Bajec, K. Bantoft, D. Betel, B. Bobechko, K. Boutilier, and E. Burgess. The biomolecular interaction network database and related tools 2005 update. *Nucleic Acids Research*, 33(Database Issue):D418–D423, 2005.

[12] S. Peri, J.D. Navarro, R. Amanchy, T.Z. Kristiansen, C.K. Jonnalagadda, V. Surendranath, V. Niranjan, B. Muthusamy, T.K. Gandhi, and M. Gronborg. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research*, 13(10):2363–2371, 2003.

[13] G.R. Mishra, M. Suresh, K. Kumaran, N. Kannabiran, S. Suresh, P. Bala, K. Shivakumar, N. Anuradha, R. Reddy, and T.M. Raghavan. Human protein reference database–2006 update. *Nucleic Acids Research*, 34(Database Issue):D411–D414, 2006.

[14] P. Pagel, S. Kovac, M. Oesterheld, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, P. Mark, V. Stumpflen, and H.W. Mewes. The mips mammalian protein-protein interaction database. *Bioinformatics*, 21(6):832–834, 2005.

[15] A. Chatr-aryamontri, A. Ceol, L.M. Palazzi, G. Nardelli, M.V. Schneider, L. Castagnoli, and G. Cesareni. Mint: the molecular interaction database. *Nucleic Acids Research*, 35(Database Issue):D572–D574, 2007.

[16] C. von Mering, L.J. Jensen, B. Snel, S.D. Hooper, M. Krupp, M. Foglierini, N. Jouffre, M.A. Huynen, and P. Bork. String: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.*, 33(Database issue):D433–D437, 2005.

[17] C.L. Myers and O.G. Troyanskaya. Context-sensitive data integration and prediction of biological networks. *Bioinformatics.*, 23(17):2322–2330, 2007.

[18] S.J. Shankland. New insights into the pathogenesis of membranous nephropathy. *Kidney International.*, 57:1204–1205, 2000.

[19] D. Kerjaschki. Megalin/gp330 and pathogenetic concepts of membranous glomerulopathy (mgn). *Kidney Blood Press Res.*, 23(3-5):163–166, 2000.

[20] P.N. Cunningham, B.K. Hack, G. Ren, A.W. Minto, B.P. Morgan, and R.J. Quigg. Glomerular complement regulation is overwhelmed in passive heymann nephritis. *Kidney Int.*, 60(3):900–909, 2001.

[21] S. Hariharan, C.P. Johnson, B.A. Bresnahan, S.E. Taranto, M.J. McIntosh, and D. Stablein. Improved graft survival after renal transplantation in the united states, 1988 to 1996. *N Engl J Med.*, 342(9):605–612, 2000.

[22] O.H. Koning, R.J. Ploeg, J.H. van Bockel, M. Groenewegen, F.J. van der Woude, G.G. Persijn, and J. Hermans. Risk factors for delayed graft function in cadaveric kidney transplantation: a prospective study of renal function and graft survival after preservation with university of wisconsin solution in multi-organ donors. european multicenter study group. *Transplantation.*, 65(5):1620–1628, 1998.

[23] P. Perco, C. Pleban, A. Kainz, A. Lukas, B. Mayer, and R. Oberbauer. Gene expression and biomarkers in renal transplant ischemia reperfusion injury. *Transplant Int.*, 20(1):2–11, 2007.

[24] C. White, A. Akbari, N. Hussain, L. Dinh, G. Filler, N. Lepage, and G.A. Knoll. Estimating glomerular filtration rate in kidney transplantation: a comparison between serum creatinine and cystatin c-based methods. *J Am Soc Nephrol.*, 16(12):3763–3770, 2005.

[25] P. Perco, C. Pleban, A. Kainz, A. Lukas, G. Mayer, B. Mayer, and R. Oberbauer. Protein biomarkers associated with acute renal failure and chronic kidney disease. *Eur J Clin Invest.*, 36(11):753–763, 2006.

[26] A.V. Cybulsky, R.J. Quigg, and D.J. Salant. Experimental membranous nephropathy redux. *Am J Physiol Renal Physiol*, 289(4):F660–F671, 2005.

[27] S.J. Shankland, J. Floege, S.E. Thomas, M. Nangaku, C. Hugo, J. Pippin, K. Henne, DM. Hockenberry, R.J. Johnson, and W.G. Couser. Cyclin kinase inhibitors are increased during experimental membranous nephropathy: potential role in limiting glomerular epithelial cell proliferation in vivo. *Kidney Int.*, 52(2):404–413, 1997.

[28] I. Davidson and J.B. Henry. *Clinical Diagnosis by Laboratory Methods.* Philadelphia, PA: Saunders, 19th edition, 1969.

[29] C. Slot. Plasma creatinine determination. a new and specific jaffe reaction method. *Scand J Clin Lab Invest.*, 17(4):381–387, 1965.

[30] A.O. Ojo, R.A. Wolfe, P.J. Held, F.K. Port, and R.L. Schmouder. Delayed graft function: risk factors and implications for renal allograft survival. *Transplantation.*, 63(7):968–974, 1997.

[31] W.H. Wang, L.G. McNatt, A.R. Shepard, N. Jacobson, D.Y. Nishimura, E.M. Stone, V.C. Sheffield, and A.F. Clark. Optimal procedure for extracting rna from human ocular tissues and expression profiling of the congenital glaucoma gene foxc1 using quantitative rt-pcr. *Mol Vis.*, 7:89–94, 2001.

[32] P. Hauser, C. Schwarz, C. Mitterbauer, H.M. Regele, F. Mhlbacher, G. Mayer, P. Perco, B. Mayer, T.W. Meyer, and R. Oberbauer. Genome-wide gene-expression patterns of donor kidney biopsies distinguish primary allograft function. *Lab Invest.*, 84(3):353–361, 2004.

[33] D.P. Harkin. Uncovering functionally relevant signaling pathways using microarray-based expression profiling. *Oncologist.*, 5(6):501–507, 2000.

[34] R.A. Irizarry, B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf, and T.P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[35] S.E. Choe, M. Boutros, A.M. Michelson, G.M. Church, and M.S. Halfon. Preferred analysis methods for affymetrix genechips revealed by a wholly defined control dataset. *Genome Biol.*, 6(2):R16, 2005.

[36] W. Huber, A. von Heydebreck, H. Sltmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics.*, 18:96–104, 2002.

[37] C. Li and W.H. Wong. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A.*, 98(1):31–36, 2001.

[38] J. Rainer, F. Sanchez-Cabo, G. Stocker, A. Sturn, and Z. Trajanoski. Carmaweb: comprehensive r- and bioconductor-based web service for microarray data analysis. *Nucleic Acids Res.*, 34(Web Server Issue):W498–W503, 2006.

[39] B.M. Bolstad. affy: Built-in processing methods. 2005.

[40] B.M. Bolstad, R.A. Irizarry, M. Astrand, and T.P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics.*, 19(2):185–193, 2003.

[41] Genepix software.

[42] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R.B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics.*, 17(6):520–525, 2001.

[43] O. Alter and G.H. Golub. Singular value decomposition of genome-scale mrna lengths distribution reveals asymmetry in rna gel electrophoresis band broadening. *Proc Natl Acad Sci U S A.*, 103(32):11828–11833, 2006.

[44] S.B. Pounds. Estimation and control of multiple testing error rates for microarray studies. *Brief Bioinform.*, 7(1):25–36, 2006.

[45] P.H. Westfall and S.S. Young. *Resampling-based Multiple Testing: Examples and Methods for P-Value Adjustment.* New York, Wiley, 1993.

[46] V.G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A.*, 98(9):5116–5121, 2001.

[47] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–29, 2000.

[48] K.D. Dahlquist, N. Salomonis, K. Vranizan, S.C. Lawlor, and B.R. Conklin. Genmapp, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet.*, 31(1):19–20, 2002.

[49] D.R. Masys, J.B. Welsh, J. Lynn Fink, M. Gribskov, I. Klacansky, and J. Corbeil. Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics.*, 17(4):319–326, 2001.

[50] J.Y. Chen, C. Shen, and A.Y. Sivachenko. Mining alzheimer disease relevant proteins from integrated protein interactome data. *Pac Symp Biocomput.*, pages 367–378, 2006.

[51] G.D. Bader and C.W. Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics.*, 4:2, 2003.

[52] S. Kerrien, Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, E. Dimmer, M. Feuermann, A. Friedrichsen, and R. Huntley. Intact–open source resource for molecular interaction data. *Nucleic Acids Res*, 35(Database Issue):D561–5, 2007.

[53] L. Salwinski, C.S. Miller, A.J. Smith, F.K. Pettit, J.U. Bowie, and D. Eisenberg. The database of interacting proteins: 2004 update. *Nucleic Acids Research*, 32(Database Issue):D449–D451, 2004.

[54] A.I. Su, T. Wiltshire, S. Batalov, H. Lapp, K.A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, and G. Kreiman. A gene atlas of the mouse and human protein-encoding transcriptomes. *PNAS*, 101(16):6062–6067, 2004.

[55] H. Hermjakob, L. Montecchi-Palazzi, G. Bader, J. Wojcik, L. Salwinski, A. Ceol, S. Moore, S. Orchard, U. Sarkans, and C. von Mering. The hupo psis molecular interaction formata community standard for the representation of protein interaction data. *Nat Biotechnol*, 22(2):177–183, 2004.

[56] T. Barrett, D.B. Troup, S.E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I.F. Kim, A. Soboleva, M. Tomashevsky, and R. Edgar. Ncbi geo: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Research*, 35(Database Issue):D760–D765, 2007.

[57] P. Horton, K.J. Park, T. Obayashi, N. Fujita, H. Harada, C.J. Adams-Collier, and K. Nakai. Wolf psort: protein localization predictor. *Nucleic Acids Research*, 35(Web Server Issue):W585–W587, 2007.

[58] G.S. Omenn, D.J. States, M. Adamski, T.W. Blackwell, R. Menon, H. Hermjakob, R. Apweiler, B.B. Haab, R.J. Simpson, J.S. Eddes, E.A. Kapp, R.L. Moritz, D.W. Chan, A.J. Rai, A. Admon, R. Aebersold, J. Eng, W.S. Hancock, S.A. Hefta, H. Meyer, Y.K. Paik, J.S. Yoo, P. Ping, J. Pounds, J. Adkins, X. Qian, R. Wang, V. Wasinger, C.Y. Wu, X. Zhao, R. Zeng, A. Archakov, A. Tsugita, I. Beer, A. Pandey, M. Pisano, P. Andrews, H. Tammen, D.W. Speicher, and S.M. Hanash. Overview of the hupo plasma proteome project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics*, 5(13):3226–3245, 2005.

[59] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11):2498–2504, 2003.

[60] M.S. Cline, M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A.R. Pico, A. Vailaya, P.L. Wang, A. Adler, B.R. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G.J. Warner, T. Ideker, and G.D. Bader. Integration of biological networks and gene expression data using cytoscape. *Nat Protoc*, 2(10):2366–2382, 2007.

[61] A. Platzer, P. Perco, A. Lukas, and B. Mayer. Characterization of protein-interaction networks in tumors. *Bioinformatics*, 8:224, 2007.

[62] A. Ogawa, M. Sakatsume, X. Wang, Y. Sakamaki, Y. Tsubata, B. Alchi, T. Kuroda, H. Kawachi, I. Narita, F. Shimizu, and F. Gejyo. Sm22alpha: the novel phenotype marker of injured glomerular epithelial cells in anti-glomerular basement membrane nephritis. *Nephron Exp Nephrol.*, 106(3):77–87, 2007.

[63] R.R. Nair, J. Solway, and D.D. Boyd. Expression cloning identifies transgelin (sm22) as a novel repressor of 92-kda type iv collagenase (mmp-9) expression. *J Biol Chem.*, 281(36):26424–264236, 2006.

[64] L. Ivanova, M.J. Butt, and D.G. Matsell. Mesenchymal transition in kidney collecting duct epithelial cells. *Am J Physiol Renal Physiol.*, 294(5):F1238–F1248, 2008.

[65] F.N. van Leeuwen, S. van Delft, H.E. Kain, R.A. van der Kammen, and J.G. Collard. Rac regulates phosphorylation of the myosin-ii heavy chain, actinomyosin disassembly and cell spreading. *Nat Cell Biol.*, 1(4):242–248, 1999.

[66] S. Lam, N.A. Verhagen, F. Strutz, J.W. van der Pijl, M.R. Daha, and C. van Kooten. Glucose-induced fibronectin and collagen type iii expression in renal fibroblasts can occur independent of tgf-beta1. *Kidney Int.*, 63(3):878–888, 2003.

[67] K. Kohri, S. Nomura, Y. Kitamura, T. Nagata, K. Yoshioka, M. Iguchi, T. Yamate, T. Umekawa, Y. Suzuki, and H. Sinohara. Structure and expression of the mrna encoding urinary stone protein (osteopontin). *J Biol Chem.*, 268(20):15180–15184, 1993.

[68] X.Q. Yu, D.J. Nikolic-Paterson, W. Mu, C.M. Giachelli, R.C. Atkins, R.J. Johnson, and H.Y. Lan. A functional role for osteopontin in experimental crescentic glomerulonephritis in the rat. *Proc Assoc Am Physicians.*, 110(1):50–64, 1998.

[69] S.J. Shankland, J. Pippin, R.H. Pichler, K.L. Gordon, S. Friedman, L.I. Gold, R.J. Johnson, and W.G. Couser. Differential expression of transforming growth factor-beta isoforms and receptors in experimental membranous nephropathy. *Kidney Int.*, 50(1):116–124, 1996.

[70] A.T. Petermann, J. Pippin, K. Hiromura, T. Monkawa, R. Durvasula, W.G. Couser, J. Kopp, and S.J. Shankland. Mitotic cell cycle proteins increase in podocytes despite lack of proliferation. *Kidney Int.*, 36(1):113–122, 2003.

[71] J.W. Pippin, R. Durvasula, A. Petermann, K. Hiromura, W.G. Couser, and S.J. Shankland. Dna damage is a novel response to sublytic complement c5b-9-induced injury in podocytes. *J Clin Invest.*, 111(6):877–885, 2003.

[72] L. Xia, L. Zheng, H.W. Lee, S.E. Bates, L. Federico, B. Shen, and T.R. O'Connor. Human 3-methyladenine-dna glycosylase: effect of sequence context on excision, association with pcna, and stimulation by ap endonuclease. *J Mol Biol.*, 346(5):1259–1274, 2005.

[73] A. Frank. *Connectivity and networkflows.* Elsevier, 1995.

[74] J.A. Van der Hoeven, S. Lindell, R. van Schilfgaarde, G. Molema, G.J. Ter Horst, J.H. Southard, and R.J. Ploeg. Donor brain death reduces survival after transplantation in rat livers preserved for 20 hr. *Transplantation.*, 72(10):1632–1636, 2001.

[75] J.M. Thurman, D. Ljubanovic, C.L. Edelstein, G.S. Gilkeson, and V.M. Holers. Lack of a functional alternative complement pathway ameliorates ischemic acute renal failure in mice. *J Immunol.*, 170(3):1517–1523, 2003.

[76] W. Zhou, C.A. Farrar, K. Abe, J.R. Pratt, J.E. Marsh, Y. Wang, G.L. Stahl, and S.H. Sacks. Predominant role for c5b-9 in renal ischemia/reperfusion injury. *J Clin Invest.*, 105(10):1363–1371, 2000.

[77] J.J. Alexander, Y. Wang, A. Chang, A. Jacob, A.W. Minto, M. Karmegam, M. Haas, and R.J. Quigg. Mouse podocyte complement factor h: the functional analog to human complement receptor 1. *J Am Soc Nephrol.*, 18(4):1157–1166, 2007.

[78] K. Yamada, T. Miwa, J. Liu, M. Nangaku, and W.C. Song. Critical protection from renal ischemia reperfusion injury by cd55 and cd59. *J Immunol.*, 172(6):3869–3875, 2004.

[79] A. Reutzel-Selke, T. Zschockelt, C. Denecke, U. Bachmann, A. Jurisch, J. Pratschke, G. Schmidbauer, H.D. Volk, P. Neuhaus, and S.G. Tullius. Short-term immunosuppressive treatment of the donor ameliorates consequences of ischemia/ reperfusion injury and long-term graft function in renal allografts from older donors. *Transplantation.*, 75(11):1786–1792, 2003.

[80] J. Wilflingseder, A. Kainz, C. Mitterbauer, G. Gyri, P. Perco, T. Meyer, B. Mayer, R. Langer, R. Magreiter, and R. Oberbauer. A mulitcenter rct of deceased organ donor pre-treatment with corticosteroids for the prevention of postischemic acute renal failure. *J Am Soc Nephrol.*, 18:31A, 2007.

ABSTRACT

Background: Large scale differential gene expression profiling has provided a major contribution to an extended analysis of cellular processes. However, interpretation of such data for reaching a functional understanding of ongoing processes is a major challenge, in particular when thinking in terms of a systems biology-driven analysis. Following this concept, an integrated interpretation of expression profiles utilizing broadly available additional 'omics' sources promises a route for improved analysis procedures.

Goals: This thesis exemplifies two workflows within the scope of computational systems biology, applied on gene expression data characterizing the molecular basis of diseases of the kidney. One workflow follows a sequential analysis procedure, centered around a core set of differentially regulated genes derived on a purely statistical basis.

The second workflow in contrast uses an interaction network-based procedure. Within this model intracellular interactions are represented as dependency graph generated by integrating 'omics' sources describing biological categories, tissue specific gene expression, protein sub-cellular location, and known protein interactions. Gene expression profiles are then interpreted on the level of this dependency graph.

Systems studied: Diseases of the kidney are a prevalent health issue, and understanding the pathophysiology might trigger improved diagnosis and therapy options. This thesis focuses on an animal model for human membraneous nephropathy, and second on mechanisms involved in kidney transplant failure. For both systems gene expression data provided the basis for utilizing the data analysis workflows.

Results: For gene expression data characterizing membraneous nephropathy the sequential analysis workflow was applied leading to the identification of 235 differentially expressed genes. A functional classification of these features indicated the disease associated involvement of changes in cell structure, cell cycle, as well as in immunity and defense.

For expression profiles linked to transplant failure protein sub-networks linked to immunity and defense, and here in particular members of the complement cascade were found as key players.

Conclusions: Major challenges along 'omics' based procedures have moved from experimental data generation to their functional interpretation. Diverse analysis workflow concepts are presently under development; their first data, however, clearly indicate the potential of computational systems biology for providing an understanding of complex cellular processes.

ZUSAMMENFASSUNG

Hintergrund: Das Erstellen von umfangreichen Genexpressionsprofilen lieferte einen wesentlichen Beitrag zu erweiterten Analysen von zellulären Prozessen. Die Interpretation dieser Daten zielt auf ein funktionales Verständnis der biologischen Prozesse ab und ist nach wie vor, insbesondere in Hinsicht auf eine systembiologische Analyse, eine große Herausforderung. Diesem Konzept folgend verspricht eine Interpretation der Expressionsprofile unter Einbeziehen von vorhandenen 'omics' Daten einen guten Ansatz für verbesserte Analysen.

Ziele: Diese Diplomarbeit erläutert zwei Arbeitsabläufe aus dem Bereich der computeruntersttzten Systembiologie, welche auf Genexpressionsdaten zur Charakterisierung der molekularen Prozesse in Nierenerkrankungen angewandt wurden. Einer der Abläufe ist eine sequentielle Analyse Prozedur, die sich mit differential regulierten Genen aus einem rein statistischen Ansatz beschäftigt.

Im Gegensatz dazu verwendet der zweite Ablauf Prozeduren basierend auf Interaktionsnetzwerken. Innerhalb dieses Modells werden die intrazellulären Interaktionen durch Kanten eines Abhängigkeitsgraphen repräsentiert. Die Gewichte dieser Abhängigkeiten werden durch Integration mehrerer 'omics' Datenquellen berechnet, die die Beschreibung biologischer Prozesse, gewebsspezifischer Genexpression, subzellulärer Lokalisation von Proteinen und Proteininteraktionen inkludieren. Anhand dieses Graphens wurden die Genexpresionsprofile interpretiert.

Untersuchte Systeme: Nierenerkrankungen stellen ein weitverbreitetes medizinisches Problem dar. Das Verständnis ihrer Pathophysiologie könnte zu einer verbesserten Diagnose und zur Eröffnung weiterer Therapieoptionen führen. Diese Diplomarbeit beschäftigt sich zum einen mit einem Tiermodell der menschlichen membranösen Nephropathie und zum anderen mit Mechanismen, die zum Versagen von

Nierentransplantaten führen. Beide angewandten Arbeitsabläufe basieren auf Genexpressionsdaten der beschriebenen Krankheiten.

Resultate: Die sequentielle Analyse der Genexpressionsdaten, welche membranöse Nephrities charakterisieren, führte zur Identifikation von 235 differential regulierten Genen. Eine funktionale Analyse der Gene zeigte, dass Veränderungen in der Zellstruktur, im Zellzyklus und auch in der Immunabwehr mit der Krankheit in Verbindung stehen.

Die Subgraphen, die basierend auf den Genexpressionsprofil der Nierentransplantate identifiziert werden konnten, waren stark mit Prozessen der Immunabwehr und im speziellen, mit Komponenten des Komplementsystems, assoziiert.

Schlussfolgerungen: Die großen Herausforderungen die sich mit 'omics' basierenden Prozeduren entwickelten, haben sich von Generation der Daten hin zu deren funktionaler Interpretationen verschoben. Momentan entwickeln sich unterschiedliche Analysekonzepte aber auch die schon vorhandenen Daten zeigen, dass die computerunterstützte Systembiologie das Potential besitzt, das Verständnis komplexer biologischer Prozesse voranzutreiben.

CURRICULUM VITAE

IRMGARD MÜHLBERGER

Parhamerplatz 17
1170 Vienna, Austria
e-mail: a0105478@unet.univie.ac.at

Personal data

Name	Irmgard Mühlberger
Date of birth	10-09-1981
Place of birth	Vienna, Austria
Nationality	Austria

Basic education

Sep 1987 - Jun 1991	Primary School (Rötzergasse Wien XVII)
Sep 1991 - Jun 1995	Secondary School (GRGXVII, Geblergasse)
Sep 1995 - Jun 2000	Secondary School (GRGXV, Auf der Schmelz)
Jun 2000	A-level

Higher education

since Oct 2001	Study of Molecular Biology at the University of Vienna, Austria
Mar 2006 - Jun 2007	Study of informatics at the Technical University of Vienna, Austria
since Sep 2007	Diploma work at the Institute of Theoretical Chemistry at the University of Vienna, Austria

Poster Presentations

{1} 6th International Conference on Pathways, Networks and Systems Medicine, Chania, Crete, Greece, June 16-21, 2008
A dependency graph approach for analyzing differential gene expression data of B-cell lymphomas
I. Mühlberger, P. Perco, A. Bernthaler, R. Fechete, A. Lukas, and B. Mayer

Paper drafts

{1} P. Perco, A. Bernthaler, I. Mühlberger, M. Haiduk, C. Stadler, R. Fechete, A. Lukas, R. Oberbauer and B. Mayer. Context analysis of differential gene expression data.

{2} P.V. Hauser, P. Perco, I. Mühlberger, J. Pippin, M. Blonski, B. Mayer, C.E. Alpers, R. Oberbauer and S.J. Shankland. Microarray and bioinformatics analysis of gene expression in experimental membranous nephropathy.

{3} A. Bernthaler, I. Mühlberger, R. Fechete, P. Perco, A. Lukas, B. Mayer. A dependency graph approach for analysis of differential gene expression profiles.