Master Thesis

Identification of functional modules in human protein interaction networks

Konstantin Pentchev[*]

03.03.2010 – 02.09.2010

1. Supervisor:	Dr. Ralf Herwig Max Planck Institute for Molecular Genetics Dept. Vertebrate Genomics Ihnestr. 63-73 14195 Berlin
2. Supervisor:	Prof. Dr. Alexander Bockmayr Freie Universität Berlin DFG-Research Center Matheon FB Mathematik und Informatik Arnimallee 6 14195 Berlin
Additional Supervisor:	Atanas Kamburov

[*] Matrikelnummer: 4083097, Freie Universität Berlin, Studiengang Bioinformatik

Contents

1 Introduction

In recent years, cancer has become the leading cause of death in the developed world, accounting for more than 13% of all deaths [1]. While many of its characteristics have been identified [2], the exact molecular processes, which drive its progression and distinguish one cancer type from another remain elusive.

In the post-*Human Genome Project* era, the focus of molecular biology has shifted from identifying the functional units of cells (genes, proteins etc) to analyzing their interactions and their dynamic behavior in the context of larger systems [3]. In order to fully comprehend the role of molecular entities, one has to study how they react to signals from outside and within the cell, what effect these changes have on their interactors and ultimately how the system accommodates the perturbations. Recently, the research on expression levels and molecular interactions has yielded large amounts of data, providing for a more thorough analysis of the molecular mechanisms underlying human disease [3, 4]. However, the vast amount of proteins and their interactions, makes it difficult to determine the molecular abnormalities, which underlie human disease and cancer in particular. This imposes the need for identifying the functional modules within the network.

With this thesis I propose a method for identifying the parts of the human protein interaction network, which functionally describe a given cancer cell line. The specificity is hereby derived from the expression profile data measured for a set of different cancer cell lines.

In the context of this thesis, a module is defined as a group of proteins, whose function can be separated from those of other sets of proteins and is defining for the presentation of a certain phenotype. The members of the module share molecular interactions, or are components of the same protein complex.

1.1 Motivation

The availability of genome-wide expression profiles has allowed researchers to link genes to diseases and cancer types in particular [5, 6, 7]. These so-called markers are selected according to how well their expression patterns discriminate between different pathologies. However, it was shown that these marker sets, while able to successfully classify diseases, are highly specific for the patients from the corresponding study and have a very small overlap [8]. It has been proposed, that this is due to strong variations in downstream effector genes' expression, whereas the expression of the genes actually causing the disease remain constant [8, 9].

Following the criticism of using expression profiles alone for the identification of markers, attempts have been made to address the issue by combining microarray data with the available networks of protein interactions in order to extract the functionally relevant modules [10, 11, 12]. The key concept behind these developments is that phenotypes are rarely defined by single genes or proteins but rather by subnetworks of interacting proteins derived from the entire human interactome. This approach allows the identification of genes with known disease related mutations, which typically remain undetected by classic expression profile analysis. Moreover, the identified subnetworks have been shown to be more reproducible between different studies [12].

The method proposed in this thesis also combines expression profiles with protein interaction networks. The goal was to develop a method capable of extracting subnetworks of functionally related entities, which are determinant for the featured phenotype. The approach is demonstrated for different cancer cell lines, presenting in essence different pathological phenotypes. Because different cancer types vary strongly in their molecular abnormalities, application of the method should highlight modules, which are specific for a given malignant cell line. The expression data was used to determine the significance of interactions for the specific phenotype. Information theory and graph methods were then used to outline characteristic genes/proteins and the modules in which these participate. What outlines this approach from others [10, 11, 12, 13], is the filtering step, which reduces the interaction graph to its functional core and the identification of functionally specific proteins, from which the modules are expanded. This makes the approach applicable to genome-wide and interactome-wide data.

1.2 Cancer

The term *cancer* describes a class of diseases caused by abnormal behavior of autologous cells. The latter are characterized by two major properties: *uncontrolled growth* and the *invasion* of foreign tissues [14]. The first ability allows the cancer cells to proliferate infinitely, without regard of control mechanisms - a tumor will develop. However, it will only be considered malignant if it is able to invade surrounding tissue. If the cancer cells enter the bloodstream or lymphatic system they can form secondary tumors called *metastases*.

Cancers are classified according to the cell type from which they arise. Malignancies originating from epithelial cells are called *carcinomas*, those from connective or muscle tissue *sarcomas* and from hemopoietic cells arise various *leukemias* [14]. These are actually very different diseases, with varying progression speed and degrees of invasiveness. Currently there are more than 100 known distinct types of cancer. A comprehensive list of known cancers is available at `http://www.cancer.gov/cancertopics/alphalist`.

Cancer usually develops due to somatic mutations and rarely due to epigenetic changes. However, a single genetic aberration is not enough to transform a healthy cell into a cancer cell. Oncogenesis is a multi-step process, involving several mutations in the lifespan of a cell, ranging from simple point mutations to changes in chromosome complement. Thereby, a number of regulatory circuits are affected [14]. However, it is unclear whether these pathways differ from cancer type to cancer type and if there are any, which are deregulated in most malignancies. It has been shown that the majority of cancers acquire six essential alterations in cell physiology (See Figure 1) [2].

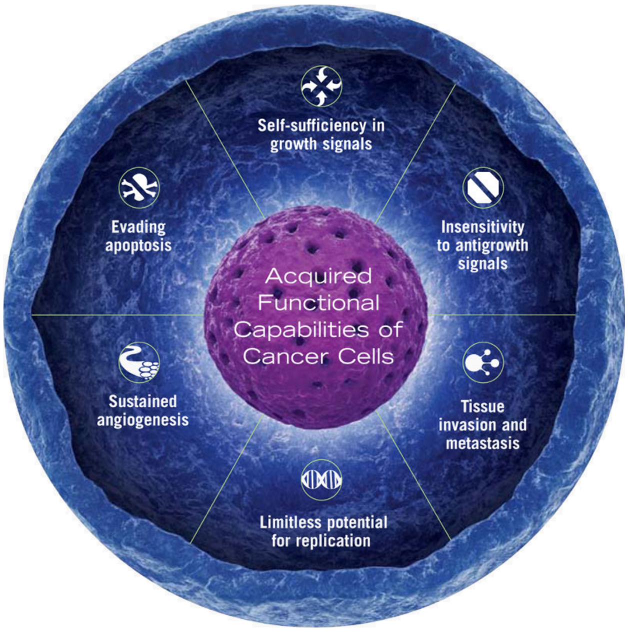
Self-sufficiency in growth signals
Evading apoptosis
Insensitivity to antigrowth signals
Acquired Functional Capabilities of Cancer Cells
Sustained angiogenesis
Tissue invasion and metastasis
Limitless potential for replication

Figure 1: The hallmarks of cancer. The six basic functional capabilities common to most malignancies are *self-sufficiency in growth signals, insensitivity to anti-growth signals, evading apoptosis, sustained angiogenesis, limitless replicative potential* and *tissue invasion and metastasis.* The order, in which these are acquired, differs from cancer to cancer. [*Genentech BioOncology*]

The first attribute a potential cancer cell usually acquires is the *self-sufficiency in growth signals.* Under normal conditions, a cell requires a range of stimuli – growth signals – in order to proliferate [14]. These stimuli come in the form of diffusible growth factors, e.g. EGF, VEGF, NGF etc., extracellular matrix components and cell-cell adhesion/interaction molecules. Three main strategies for achieving growth signal independence have been identified [2]. The first involves production of growth signals by the cell, to which it is itself responsive. This creates a positive feed-back loop called autocrine stimulation. For example glioblastomas produce PDGF[1] and sarcomas TGF-α [2]. The second strategy is based on the alteration of transcellular transducers of those signals – usually the extracellular receptors for the growth factors. This can be achieved by overexpression of the receptor proteins, allowing the cancer cell to be stimulated by molecule concentrations, which would not be sufficient otherwise. For example, the *EGFR* is overexpressed in stomach, brain and breast tumors, while the *HER2/neu* receptor is overexpressed in mammary carcinomas. In addition, structural changes in the receptors due to mutations can cause the receptors to be in a permanent stimulated state, without binding to their ligands. Finally, growth signal autonomy can be achieved by deregulations of the downstream pathways, which transduce the signals to the cell nucleus. One such major pathway, which is deregulated in about 25% of human cancers is the *SOS-Ras-Raf-MAPK* cascade [2].

Cell proliferation is governed not only by growth stimulating factors, but rather an interplay by the latter and so called *antigrowth signals.* These can either force the cell into the quiescent G_0 state of the cell cycle, or initiate differentiation, which causes the cell to lose its proliferative potential [2]. Therefore, potential cancer cells have to develop some sort of resistance to these signals if they are to prosper. One crucial stage of the cell cycle is the transition from G_1 to S phase. The transcription factor $E2F$ governs the expression of a set of genes, required for cell proliferation to start. Its functionality is regulated by the *retinoblastoma protein* (pRb). If the latter is phosphorylated, it releases $E2F$ and the cell can enter the S phase. One common external antigrowth signal comes in the form of *TGF-β*, which through *p21* blocks the phosphorylation of *pRb*, thus activating it [14]. Most tumors lose their responsiveness to TGF-β due to downregulation of the receptors. Mutations in the latter or in downstream regulators, including *Smad4*, *p21* and *pRb* itself, also render the cell immune to this antigrowth pathway. How cancer cells avoid post-mitotic differentiation is not fully understood, but it has been shown that it involves the oncogenes *c-myc* and *erbA*. Deregulations of the *APC/β-catenin pathway* have also been recorded to block differentiation [2].

As discussed above, cells acquire several mutations before exhibiting malignant properties. However, there exists a built in program, which monitors a cell's DNA for the accumulation of abnormalities. Upon detection, the

[1] platelet-derived growth factor

[2] tumor growth factor alpha

protein *ATM* is activated, which can either force cell cycle arrest through *p21* or initiate the cell's self-destruct programm - *apoptosis* [14]. Thereby, the cell is disposed of in a controlled manner for the benefit of the organism. Clearly then, *evading apoptosis* is a necessity for every cancer cell. In about 50% of all known malignancies this is done by deregulating the function of the tumor-suppressor *p53*. This protein mediates the signal from the nucleus to pro-apoptotic members of the Bcl-2 protein family[3], which form pores in the mitochondrial membrane. Thereby, *cytochrome c* is released, which is essential for the initiation of the cell death programm. Upregulating the expression of anti-apoptotic members of the Bcl-2 family[4] is another common strategy for cancer cells to block apoptosis [2].

The three acquired capabilities discussed above, should suffice to enable the generation of vast tumor cell populations. However, this presumption has been negated by the discovery that mammalian cells have an intrinsic program, which limits their replicative potential [14]. This program operates independently of cell-cell signaling pathways. Indeed, it has been proven that cells in culture have a limited replicative potential [2, 14]. The state they enter after reaching their doubling maximum is termed *senescence*. In certain cell types it can be circumvented by disabling the tumor suppressors p53 and pRb, allowing additional generations to be created until the cells enter the *crisis* state. It is characterized by massive cell death and karyotypic abnormalities. It has been observed that cancer cells become immortalized during oncogenesis, allowing them to gain limitless replicative potential [2]. The key to acquiring this ability is the maintenance of telomeres – the ends of chromosomes, composed of several thousand repeats of a 6 bp sequence. At each replication cycle part of the telomeres are lost, due to the inability of DNA polymerases to completely replicate the 3'-ends of chromosomal DNA. Eventually, this leads to the latter becoming unprotected, causing chromosomal fusions, i.e. karyotypic disarray. This is achieved mainly by upregulating the expression of telomerases – enzymes, which add nucleotide repeats onto the ends of telomeres. However, this does not explain how the senescences stage is circumvented.

Cells in higher organisms are dependent on the oxygen and nutrients supplied by the vascular system. New blood vessels are growing actively during organogenesis – a process called *angiogenesis*. However, as soon as the tissue is formed, angiogenesis enters a quiescencent state and is strictly regulated. Because the maximal distance from a capillary for a cell to survive is approximately 100 μm, tumors can grow only to a certain size, before

[3] e.g. Bak, Bax

[4] e.g. Bcl-2, Bcl-X_L

undersupplied cells start to necroticly die out [2]. Therefore, it is critical for macro-tumors to force the growth of blood vessels towards them. Angiogenesis is controlled by a complex interplay of positive and negative signals. The *vascular endothelial growth factor* (VEGF) and the *fibroblast growth factors* (FGF1/2) are examples of pro-angiogenic signals, which bind to the transmembrane tyrosine kinase receptors of endothelial cells. On the other hand, *thrombospondin-1* is a prototypic angiogenesis inhibitor. *Integrins*, mediating cell-cell and cell-matrix association also play a critical role, as quiescent vessels express one class of them and sprouting capillaries another.Indeed, many tumors activate an *angiogenic switch* by shifting the balance of pro- and anti-angiogenic factors. This involves the upregulation of VEGF and/or FGFs expression and the downregulation of thrombospondin-1 [2]. Consequently, tumor angiogenesis is an attractive therapy target.

Approximately 90% of cancer deaths are due to *metastases*. The ability to invade foreign tissue and travel to distant sites in the organism provides neoplastic cells with new terrain, where initially nutrients and space are not a limiting factor [2, 14]. Thus, the metastases, having usually acquired the other hallmark capabilities, rapidliy grow to form secondary tumors. The processes of invasion and metastasis involve changes on two levels – the uncoupling of the cell's physical attachment to its environment and the activation of extracellular proteases, enabling it to migrate accross blood vessel walls and epithelial cell layers [2]. The former property is achieved through changes in proteins mediating cell-cell and cell-matrix contacts – *CAMs* and *integrins*. One commonly deregulated protein is *E-cadherin*, a homotypic cell-to-cell interaction molecule. It is also coupled to *β-catenin*, thus regulating several intracellular signaling pathways. Its function is impaired in cancer cells by means of mutational inactivation of its gene, transcriptional repression, or proteolysis of its extracellular domain. CAMs from the immunoglobulin superfamily, e.g. N-CAM, also undergo changes in their expression, switching from a highly adhesive isoform to a poorly adhesive. The expression pattern of integrins is also altered in cancer cells, shifting the composition of integrin α and β subunits displayed on the membrane surface. The new integrin composition has been shown to bind preferentially to the degraded stromal components, produced by proteases [2]. The genes for the latter are also upregulated and the enzymes are transformed to their active form. The alteration in activity of all these proteins is clearly essential in acquiring the invasiveness and metastatic ability. However, the signaling pathways and molecular mechanisms, which regulate these shifts remain loosely understood. An overview of the current knowledge on signaling pathways involved in oncogenesis is shown in Figure 2.

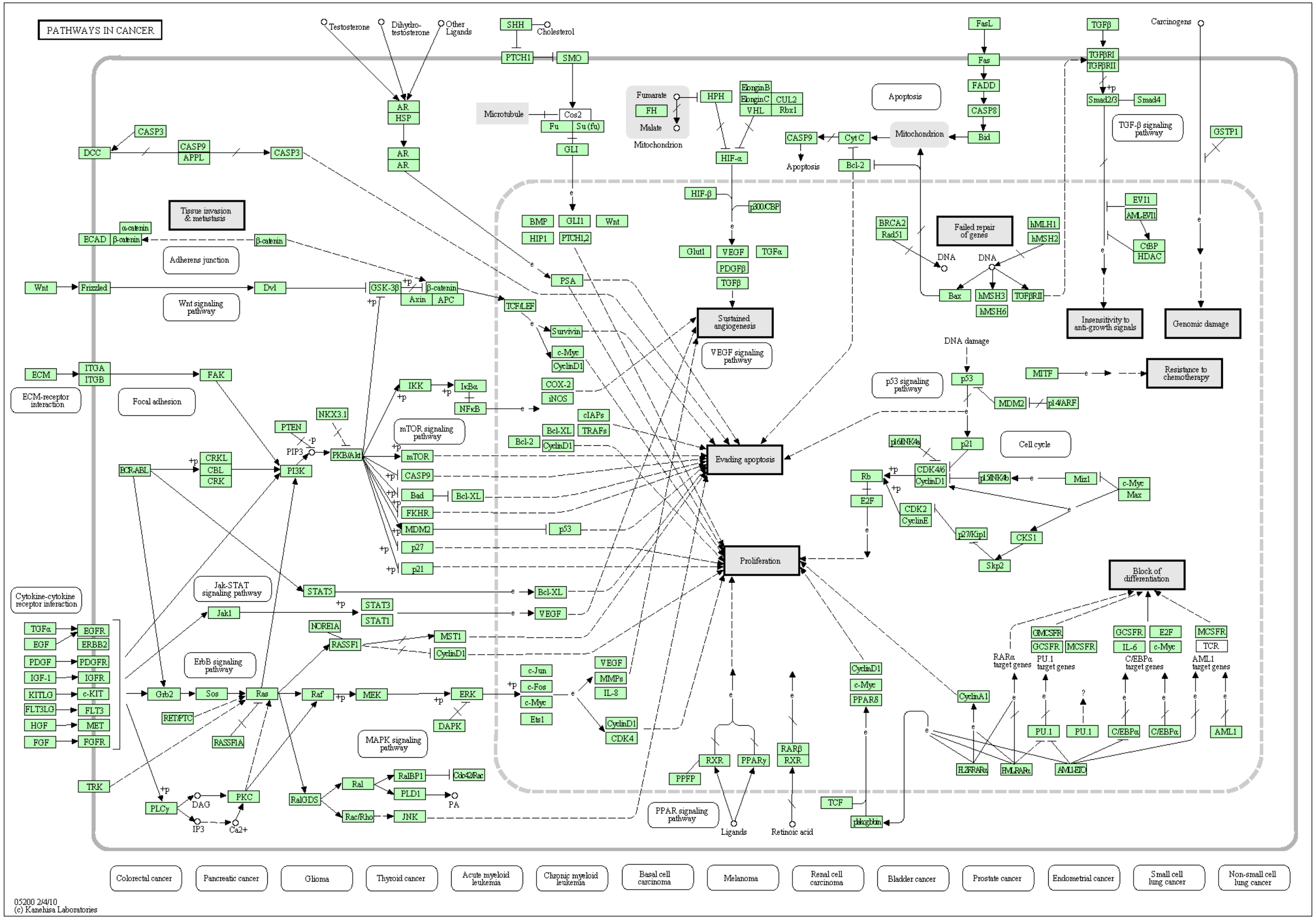

Figure 2: Cancer signaling pathways. A global map of signaling events related to oncogenesis as proposed by the Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway database. The green boxes represent entities linked to entries from KEGG Genes. Arrows stand for activation interactions, while edges ending with a "T" symbolize inhibitions. Dashed lines indicate indirect effects. [*KEGG http://www.genome.jp/kegg-bin/show_pathway?map05200*]

1.3 Protein interaction networks

In the past few years a wide range of methods have been developed for the identification of biomolecular networks, e.g *yeast two-hybrid assays*[5], for the detection of protein protein interactions [15, 16], *tandem affinity purification* coupled with *mass spectrometry*[6], for determining protein complexes [17, 18] and *chromatin immunoprecipitation* for unraveling protein-DNA interactions among others. This has spurred the study of how biological entities function in the context of each other, resulting in a variety of biological networks – gene regulatory networks, signal transduction networks, metabolic pathways and kinase-substrate interaction maps (See Figure 3). The adaption of Y2H for large-scale experiments yielded large-amounts of human protein-protein interaction data [19, 20]. This, combined with their simplicity promoted *protein interaction networks* as the most commonly studied network class.

Protein interaction networks consist of binary relations between proteins, usually representing some physical interaction on molecular level – binding, phosphorylation, cleavage etc. These can formally be described as undirected graphs of the type $G < V, E >$, where the set of nodes V is equal to the set of interacting proteins. An undirected edge $e_{p_1,p_2} \in E$ exists if and only if the proteins p_1 and p_2 interact.

This formal definition allows the study of protein networks using graph-based and statistical approaches. Several global properties have been identified, which describe interaction networks: *average degree, average shortest path, connected components* and *clustering coefficient* among others [21]. However, the structure of the graph depends on the input data used to create it. Inferring a network from *Y2H* is trivial, as it detects binary interactions. However, TAP-MS for example yields information on multi-protein complexes. The exact interaction pattern between the components differs from complex to complex. Therefore, two models have been developed to deal with this issue – the *spoke* model suggests that there exists a central protein connected to the remaining members of the complex. On the other hand, according to the *matrix* model, there exists an interaction between all components of the complex (See Figure 4). Consequently, a complex of n proteins will contribute $n - 1$ interactions to the network after the spoke model and $\frac{n \times (n-1)}{2}$ after the matrix model, possibly altering the statistical properties of the graph.

The average degree[7] and its distribution are two key characteristics of graph models. Based on empirical observations, the average connectivity of

[5] Y2H

[6] TAP-MS

[7] i.e. connectivity

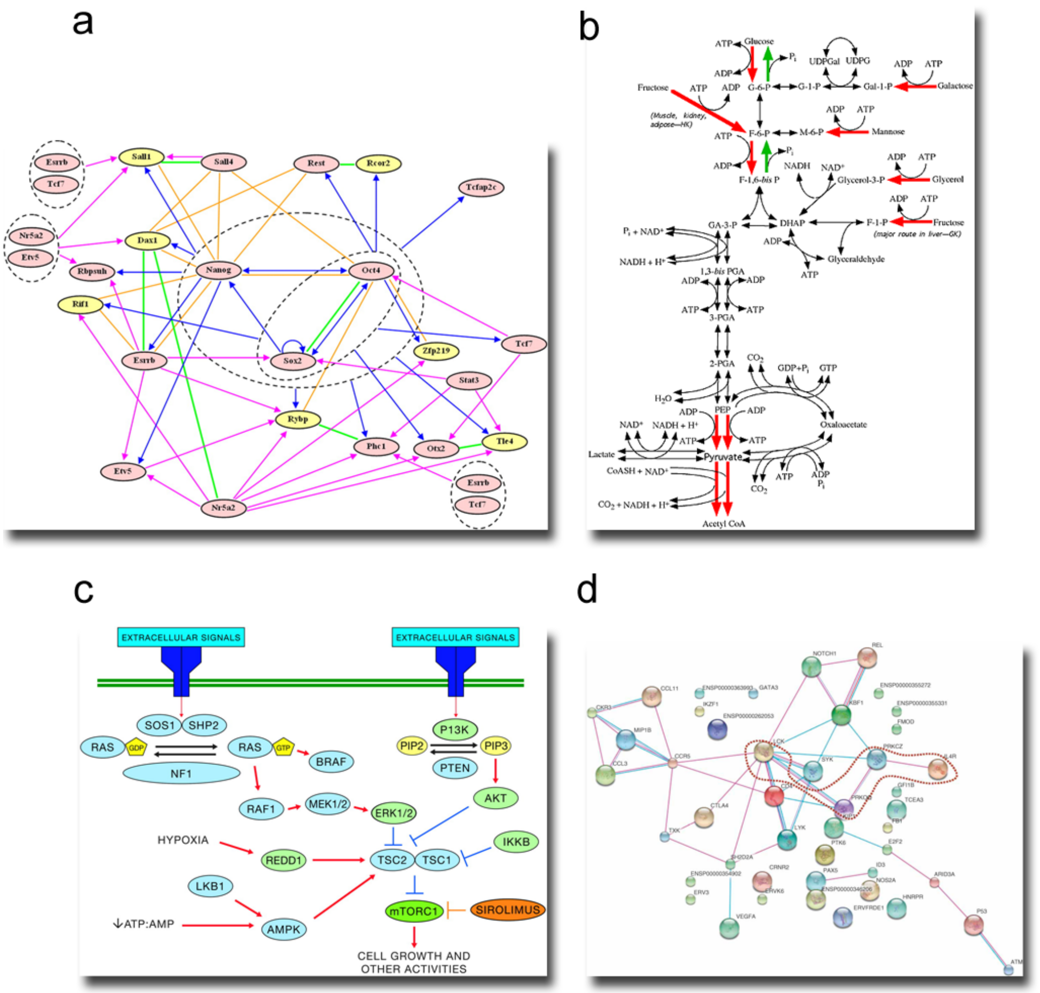

Figure 3: An overview of different network classes. a) A gene regulatory network, describing how a set of genes influence each other's expression [*Zhou, Q. et al]*. ***b) A metabolic network*** *depicting the processes of* Glycolysis *and* Gluconeogenesis *[Reactome].* ***c) A signal transduction network*** *showing how extracellular signals are processed from the cell surface to the nucleus [Davies, D. M. et al].* ***d) A classic protein-protein interaction network****, consisting of binary interactions [Nicolini, C. et al].*

a protein interaction network has been estimated to approximately 2.5. Additionally, it has been postulated that the distribution follows a power law with an exponential cutoff [21]. However, coming from an evolutionary point of view and arguing that only a small fraction of the complete human interactome is resolved, the hypothesis has been proposed that the connectivity

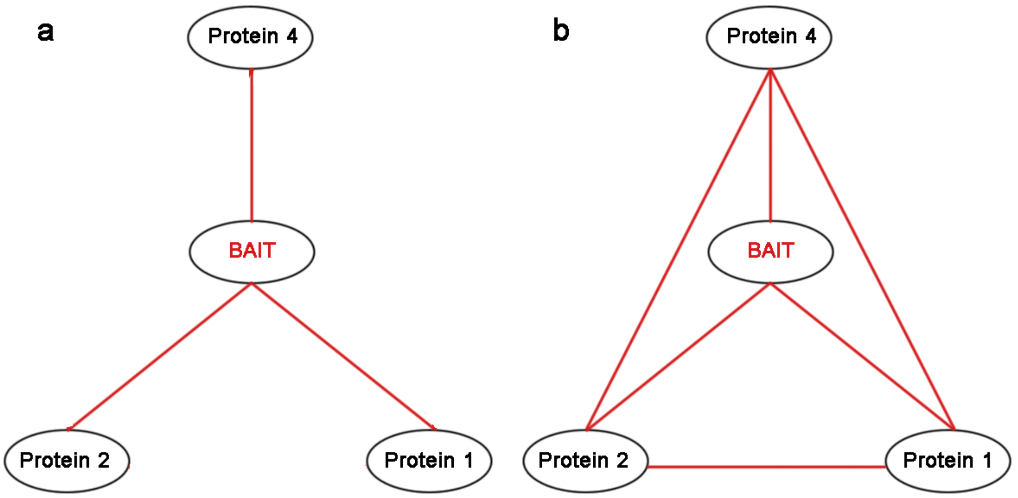

Figure 4: a) The spoke model. A central protein exists within the complex, to which all other components are connected. **b) The matrix model.** No central protein exists. All components of the complex are connected and form a clique.

distribution is actually not scale free[8] [22].

The clustering coefficient is a measure of how strong nodes in a graph tend to cluster together. It can be calculated both for single nodes(local) and for the entire graph (global). The local clustering coefficient c of a node v_i in a graph is a degree of how close that node and its immediate neighborhood N_i are to being a clique, i.e. to being fully connected. If k_i is the degree of v_i and E_{N_i} the set of edges in N_i, for undirected graphs it is defined as:

$$c_i = \frac{2 \times |E_{N_i}|}{k_i(k_i - 1)} \tag{1}$$

The global clustering coefficient is related to the local. It can be defined as:

$$C = \frac{1}{N} \sum_{i=1}^{N} c_i \tag{2}$$

It is also equal to the number of closed triplets over the number of total triplets in the graph. A triplet is a set of three nodes with interactions between them. If these are fully connected the triplet is called closed. It has been shown that protein interaction networks are highly clustered [23].

According to both degree distribution theories, there exist a relatively small number of proteins that are highly connected and a majority with only

[8] i.e. does not follow a power-law

a few interactions. The former are called *hubs*. This structure of the protein network ensures a robustness to random deletions of nodes/proteins, as it is far more likely to remove vertices of low degree. However, if one of the hubs is deleted, it is likely that more connected components arise, which can be linked to higher lethality. Indeed, it has been shown that highly connected protein are more often essential from a biological point of view [24].

1.4 Expression profiles

In principle, all cells in an multicellular organism contain the same genetic information. However, these differ in their shape, structure and function. These differences are due to the variable expression of genes, leading to the accumulation of distinct levels of RNAs and proteins [25].

Gene expression describes the process of reading a DNA fragment – the gene – and translating it into proteins. In eucaryotes this includes the processes of transcribing the gene DNA to RNA (*Transcription*), the excision of Intron sequences from the RNA (*Splicing*), transporting the mRNA out of the nucleus into the cytosol and its subsequent translation into proteins (*Translation*). The ultimate goal is to determine how genes – and the proteins they encode – function and interact in a living organism. Of interest is, which genes are switched on and off, at which point of a cell's life, for which proteins they encode and eventually how much of the proteins is produced [14].

There are several methods for studying gene expression, each providing answers to different questions. For example, *Linkage analysis* determines where a gene is located in the genome, while comparing it to other known genes in search for homology can predict its function. Using *reporter genes* one can reveal when a gene is switched on or off. However, all of these methods are suitable for studying a single gene at a time.

DNA microarrays have revolutionized the way we study gene expression. They use thousands of DNA fragments as probes, which can bind gene-specifically to RNA. Thus, we can measure the amount of RNA produced by the transcription of a gene at a certain time. We can determine which genes are switched on or off as cells grow, divide or respond to external stimuli. Moreover, the amount of RNA gives us a hint how strong a gene is read and relatively how much of its product protein is present in the cell. Most importantly, by examining the expression of so many genes simultaneously, we can study the expression patterns, which underlie cellular physiology. However, one should not forget that gene expression is regulated on several levels – transcription, RNA processing, mRNA transport and localization, mRNA degradation, translation and protein activity [14]. Microarrays give us information on gene expression after just the first control stage.

cDNA microarrays are based on cDNAs isolated from cells and printed on glass or membrane slides [26]. Because cDNA libraries are used, each spot may contain different amounts of cDNA. Therefore, the comparison of measurements between arrays and genes is challenging. Usually, a reference is hybridized to the microarray labeled by a different color, which can then be used to normalize the data.

Another type of microarrays is based on synthetic oligonucleotides, which are densely printed on a surface. The main difference to cDNA microarray is that the amount of oligonucleotides per spot is equal for all spots. After the sample RNA binds to the chip. it is labeled with a fluorophore. The intensity of the signal is assumed to be proportional to the amount of RNA in the sample [27]. For each microarray a background correction has to be performed in order to eliminate signals caused by non-specific binding and auto-fluorescence. In addition, the measurements contain systematic errors due to variations in RNA extraction, reverse transcription, labeling, photodetection etc. Therefore, different normalization strategies are applied, attempting to minimize the effects of these aberrations, e.g. *Robust Multiarray Analysis* [28], *Variance Stabilizing Normalization*[29] etc.

A expression profile describes the specific pattern of gene expressions for a set of genes of a given cell sample at the specific time point. These profiles can be used to describe a given cell type, which is in homeostasis, or to study how the cell type reacts to certain perturbations.

2 Methods

Given a large-scale protein interaction network and a set of expression profiles for several different phenotypes, the method creates an interaction graph for each phenotype. Then the edges in each graph are weighted according to the expression values for each phenotype respectively. In the next step the graphs are reduced to their functional cores. These reduced graphs are then compared to each other in order to identify the functionally specific proteins. Finally, using these proteins as seeds, the functional modules are extracted.

2.1 Data

2.1.1 Human protein interaction network

The recent development of large-scale methods for identifying molecular interactions has yielded large amounts of data. However, most of the information is dispersed between more than one hundred databases. Moreover, because it is coming from different sources and experiments this data is heterogeneous. Consequently, if a detailed analysis of the substructures in the human interactome is to be conducted, the data has to be integrated in order to provide as complete a model as possible.

ConsensusPathDB is a metadatabase on human molecular interactions. Currently[9] it integrates data from 18 different sources. The database supports three classes of interactions – *physical interactions* stand for protein-protein and protein-compound interactions, *biochemical interactions* for metabolic and signaling reactions and *gene regulations* describe how the transcription of genes is regulated. ConsensusPathDB uses a bipartite graph model, where some nodes represent interactions and some physical entities, e.g. proteins, compounds, substrates *etc.* In order to account for redundancies, *physical entities* are mapped to one another based on common accession numbers, e.g. UniProt, Ensembl, Entrez-gene, ChEBI etc. Interactions are also compared, based on their participants (See Figure 5). ConsensusPathDB distinguishes between two types of interactors - primary and secondary participants. Two or more interactions are considered similar if all of their primary participants match. These are the *interactors* for physical interactions, the *products* and *substrates* for biochemical reactions and the *regulated genes* for gene regulations [30].

The integrated protein interaction network from ConsensusPathDB is freely available. It contains data from seven source databases and was used

[9] version 14

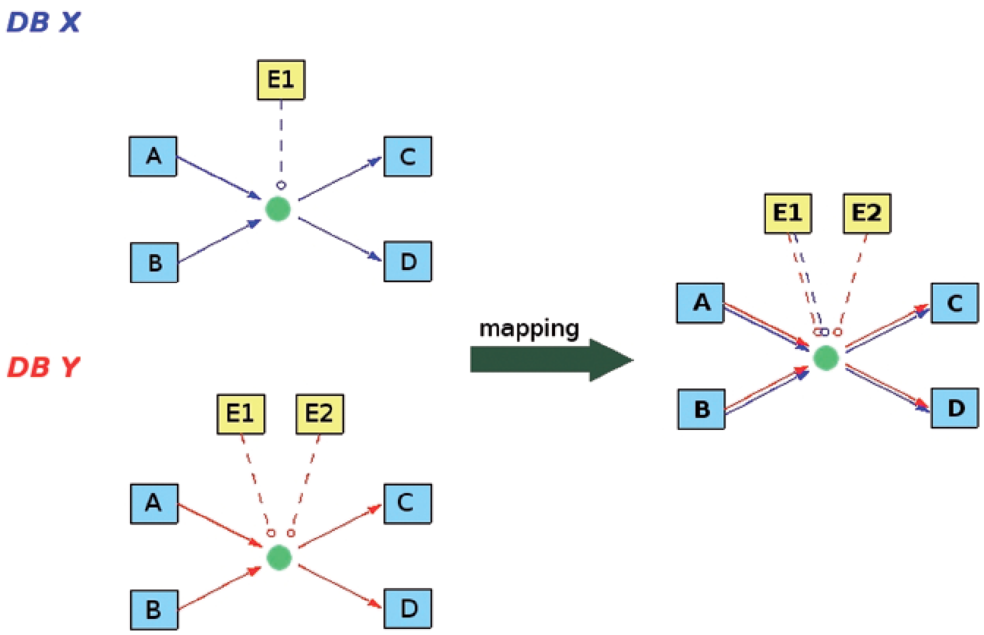

Figure 5: Interaction mapping in ConsensusPathDB. The scheme illustrates how 2 biochemical reactions $A+B->C+D$ from two databases X and Y are mapped to one another based on their primary participants. The enzymes (secondary participants) $E1$ and $E2$ are not regarded when deciding if the two interactions are matching. [*Kamburov, A.s et al*]

as the basis for this thesis. The network consists of 12706 proteins and 170906 interactions/edges. Self-interactions were eliminated for the purpose of the thesis. Complex interactions – between more than two proteins – were resolved to binary interactions according to the matrix model. The graph is not connected, but consists of 234 components. The average degree of the network is 26.9.

2.1.2 Cancer cell lines

For the purpose of the thesis an expression profile on a genome-wide scale was required. Moreover, as the method makes use of correlations of gene expressions, multiple samples per phenotype had to be available. Ideally, the measurements for all phenotypes had to be performed on the same platform and under comparable conditions.

The *Connectivity Map dataset* available at `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5258` contains genome-wide measurements of several cell lines threated with different bioactive compounds. Because a sample exists for each cell line and each compound, a high enough number of cases was provided for calculating meaningful correlations per phenotype. The *Affymetrix Human Genome U133A Array* was used for all measurements.

The raw data was normalized according to the workflow in Figure 6 using the GC-RMA approach [31], an improvement of the RMA method. The sets were grouped according to cell line threated with a certain compound at a certain concentration. Custom CDFs as defined by [32] and available at `http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/genomic_curated_CDF.asp` were used.

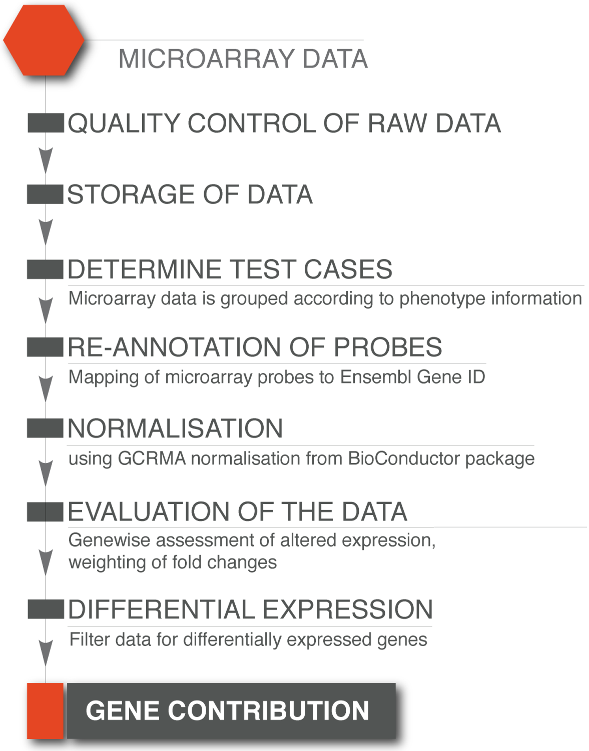
MICROARRAY DATA
QUALITY CONTROL OF RAW DATA
STORAGE OF DATA
DETERMINE TEST CASES
Microarray data is grouped according to phenotype information
RE-ANNOTATION OF PROBES
Mapping of microarray probes to Ensembl Gene ID
NORMALISATION
using GCRMA normalisation from BioConductor package
EVALUATION OF THE DATA
Genewise assessment of altered expression, weighting of fold changes
DIFFERENTIAL EXPRESSION
Filter data for differentially expressed genes
GENE CONTRIBUTION

Figure 6: Workflow for microarray data processing. The raw data is grouped according to test cases. The microarray probes' results are mapped to Ensembl Gene IDs using the custom CDFs. Each test case is then normalized using the GCRMA approach. [*Rasche, A. et al*]

For this thesis four cancer cell lines were selected. Each of these was extracted from a different cancer type (See Table 1). Because the experiment was originally designed to test the effect of bioactive molecules on the expression of genes, the fold-change between the treated and control samples is only indicative of the effect of the compounds used. Ideally, a control sample of healthy tissue would be used to calculate the changes in gene expression due to cancer-related mutations for each cell line. However, such measurements were not available. To solve this issue I made use of the notion, that genes, which correlate strongly over many perturbed conditions, are likely to be involved in the same or similar cellular processes [10]. This assumption fits ideally with the expression data, as for each cell line at least 50 measurements were available. Of course, when calculating the expression correlations in this way, both cancer specific and unspecific relations will be highlighted. The idea behind this approach is that when compared to the correlations from other cell lines, the unspecific interactions[10] can be filtered out.

Table 1: A summary of the cell lines used.

Cell line	Cancer type
MCF7	breast cancer
PC3	prostate cancer
HL60	human promyelocytic leukemia
SKMEL5	melanoma

2.2 Calculation of interaction scores

Different methods have been used to weight a graph of molecular interactions, for example using the number of experiments that support an interaction or the co-occurrence of the interactions in medical texts. However, these approaches provide just a measure of confidence for the interactions, but fail to highlight possible functional association between pairs of proteins.

Clustering of gene expression has been vastly used and showed that the products of coexpressed genes often contribute to a common biological function [33]. However, this notion is over-simplifying the problem as for example two functionally related proteins might be inversely co-regulated.

Pearson's correlation coefficient is the simplest measure of the linear relation between two variables. It is defined as the quotient between the covariance of the two variables and their respective standard deviations:

$$cor_P(x,y) = \frac{cov(x,y)}{sd(x) \times sd(y)}, \text{ where} \tag{3}$$

$$cov(x,y) = \frac{1}{n-1}\sum_i (x_i - mean(x))(y_i - mean(y)) \tag{4}$$

[10] all cancer unrelated and uncharacteristic for the cell line

The correlation coefficient will be close to ±1 if a strong relation between the two variables exists and close to 0 otherwise.

Gene co-expressions often exhibit non-linear relations. For example, the upregulation of a gene, whose product is a transcription factor will cause another gene to also become upregulated. However, the signal will be amplified, causing an exponential increase in the expression of the second gene. Apparently, Pearson's correlation coefficient, which assumes linear dependance, is not suitable for quantifying co-expression.

The *Spearman rank correlation* is a measure of the non-linear relation between two variables. It is based on the simple correlation between the ranked variables:

$$cor_S(x,y) = \frac{cov(rank_x, rank_y)}{sd(rank_x) \times sd(rank_y)} \tag{5}$$

The correlation between the transformed variables(ranks) will also take values between -1 and $+1$. It will be close to $+1$, if the order of the values of x and y is the same. Consequently, it is a measure not of the linear but of the monotone relation between variables.

It was used in this thesis as a measure of co-expression. The rank correlation was calculated for all pairs of proteins, for which an interaction exists in the protein interaction network. If no expression data was available for some protein its interactions were scored with a 0 and thus disregarded from further analysis.

2.3 Filtering

To reduce the interaction graph to its functionally relevant components, a filtering step is performed based on the co-expression scores calculated according to the previous subsection. Two criteria for retaining an interaction were applied - if its participants are significantly coexpressed, or if it lies on a shortest path between two significantly coexpressed proteins.

The first major issue was deciding, whether the expression correlation of two proteins is significantly high. Setting an arbitrary absolute threshold seemed inappropriate, as the values from different experiments would vary due to a handful of conditions. Instead, I computed the distributions of co-expressions for each cell line (see Figure 7). Then for each cell line I calculated a threshold, equal to the 0.9 quantile of the corresponding distribution. The values are listen in Table 2.

As noted before, co-expression provides insights into functional relations between genes and their products. Therefore, the goal being to identify modules of functionally related proteins, it is reasonable to retain proteins,

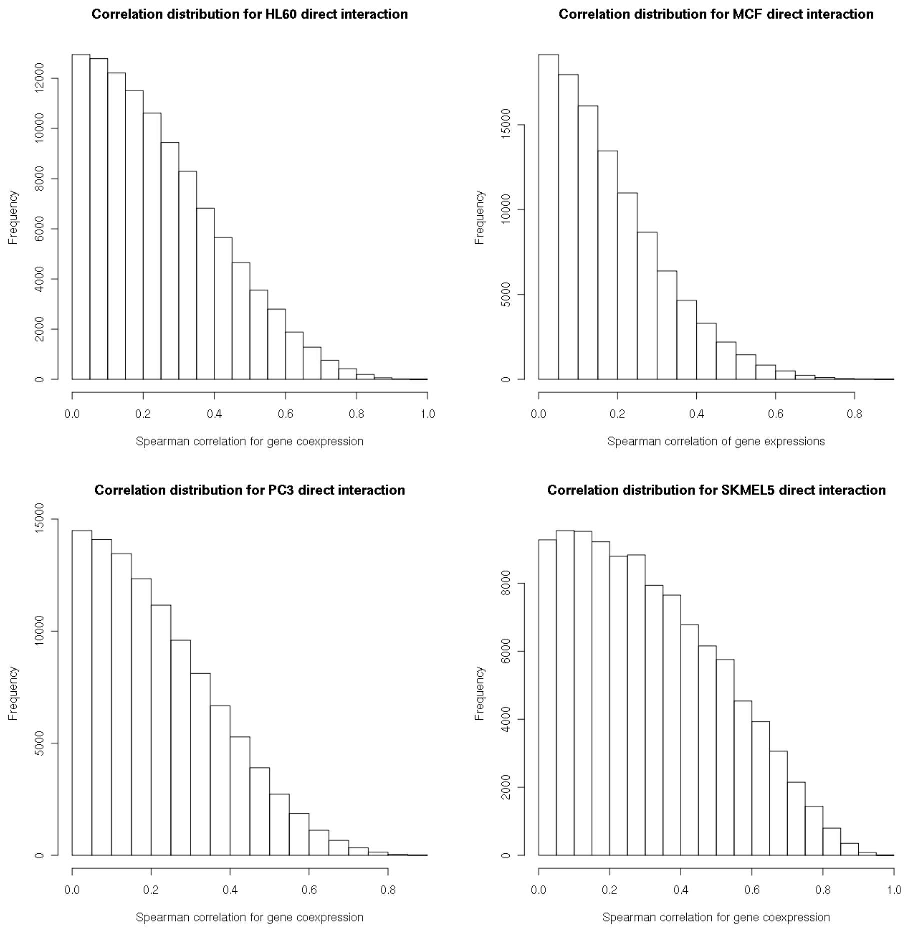

Figure 7: Correlation distributions. The correlation distributions for the four cell lines are shown as histograms.

Table 2: The co-expression significance thresholds equal to the 0.9 quantile for the four cell lines.

Cell line	co-expression significance threshold
MCF7	0.6
PC3	0.66
HL60	0.72
SKMEL5	0.8

which have both a direct interaction and high co-expression. In practice, this means that an interaction is preserved in the graph if the absolute value of its score is greater than the given threshold or it is removed otherwise.

However, there exist pairs of proteins with high co-expression values, which do not have direct interactions. Obviously, these can still be functionally related, transitively interacting with each other. In order to resolve the 'pathway', it will be required to also retain the proteins, which maintain the transient interaction. One possible method for analyzing transitive dependencies is the *shortest path* [34]. Hereby, pairs of co-expression values above 0.9 are regarded. Dijkstra's shortest path algorithm was used, where the edge weights are the absolute co-expression values. Because real networks have a small average shortest path, which is approximately 4 for protein interaction networks, paths where the two seed proteins are separated by more than 3 proteins are unlikely to connect functionally related proteins and are therefore discarded.

2.4 Identification of key genes

Given the reduced graphs obtained for each cell line, the specific proteins need to be identified. Suppose a score is available for each protein j for a given cell line i of the set g. Its specificity can be quantified by an adaptation of *Shannon's entropy* formula, which is a measure of the uncertainty associated with a random random variable:

$$S_j = -\sum_{i=1}^{g} p_{ij} \log_2 p_{ij} \tag{6}$$

S_j will be close to zero, when the protein score is significantly higher in one cell line compared to the others and close to 2, if the scores are equal for all cell lines.

A simple score for the proteins has been used, equal to the sum of the absolute co-expression scores of its interactions:

$$z_{ij} = \sum_{e \in incidence(v_{ij})} |w_e| \tag{7}$$

It is a measure of the importance of the node(protein) for the network – reflected by the number of its incident edges. In addition, the score is also indicative of the functional significance of the node, as the actual weights of the edges are used for calculation. Finally, it is normalized according to the scores for the protein for all cell lines:

$$p_{ij} = \frac{z_{ij}}{\sum_{k \in cl} z_{kj}} \tag{8}$$

If a protein is not present in the reduced graph of a cell line, it receives the score 0 correspondingly.

2.5 Extracting the modules

Finally, the actual extraction of the modules has to be performed. For the task I used a modified version of the *Detect Module from Seed Protein* (DMSP) algorithm [13].

The original algorithm runs on a weighted graph structure. Starting from a *seed* protein, it expands its neighborhood and builds functional modules. The decision, whether a proximity node should be part of the module, is based on several concepts and criteria. Two key observables are the *internal* and *external weighted degree of a node*, defined as the sum of weights of edges between a node x and its neighbors, which are part of a subnetwork G_1 or not, respectively. The sum is divided by the number of neighbors:

$$\beta_{G_1}^{INT}(x) = \frac{\sum_{y \in N_{G_1}^{INT}} w_{xy}}{|N_{G_1}^{INT}|} \tag{9}$$

$$\beta_{G_1}^{EXT}(x) = \frac{\sum_{y \in N_{G_1}^{EXT}} w_{xy}}{|N_{G_1}^{EXT}|} \tag{10}$$

Another important measure is the *density* of a graph $G(V, E)$, defined as the quotient between the number of vertices in the graph and the number of all possible vertices:

$$D_w(G) = \frac{\sum_{x,y \in E} w_{xy}}{|V|(|V| - 1))} \tag{11}$$

Initially, a kernel K_s is selected, consisting of all neighbors of the seed protein. Then some of the nodes are filtered out according to two conditions - a node's external weighted degree has to be greater than its internal weighted degree and the fraction of its internal degree and the sum of the internal and external degree[11] has to be greater than some threshold p_1:

$$\beta_{K_s}^{INT}(x) < \beta_{K_s}^{EXT}(x) \tag{12}$$

$$IO(K_s, u_i) = \frac{|N_{K_s}^{INT}(u_i)|}{|N_{K_s}^{EXT}(u_i)| + |N_{K_s}^{INT}(u_i)|} > p_1 \tag{13}$$

Vertices that pass the above criteria are sorted according to their internal degree. Following this step, vertices from the sorted list are removed one at a time, starting from the most insignificant, until a certain value of weighted density of the kernel is reached. This is done, in order to receive an even more coherent kernel.

[11] i.e. the nodes total degree

In the next stage, the algorithm proceeds iteratively to add adjacent nodes to the kernel. These are again selected based on two criteria. The first one is the same as in equation 13. If it is satisfied, the procedure checks whether the edge's weight connecting the new node is smaller than some percentage p_2 of the node's weighted internal degree:

$$w_{vu_i} \leq p_2 \times \beta_{K_s}^{INT}(v) \tag{14}$$

Thereby, p_2 is required to be between 0.9 and 1.0.

The difference between the original version of the *DMSP* and the implementation in this thesis is the scoring function for the edge weights. In [13] the weights are based on the distance between the centroids of the corresponding expression profile clusters. For the purpose of the thesis, the co-expression scores previously calculated were used. The functionally specific proteins identified in the previous step were used as seed nodes.

3 Results

The method proposed in this thesis has been implemented in *Java*. The *Java Universal Graph/Network* (JUNG) Framework was used for modeling the graphs and calculating shortest paths, degrees etc. For each cell line a co-expression matrix of approximately 70 *Mio* cells was calculated and stored for further use. 11582 out of 11899 genes measured where mapped to 11668 proteins from the network. This was done using the internal *ConsensusPathDB* Ensembl to UniPROT mapping table.

The filtering for each network with the corresponding threshold yielded four reduced graphs, each with a different size and composition. See Table 3 for details.

Table 3: A summary of the reduced graphs for each cell line with the corresponding threshold.

Cell line	co-expression		reduced graph	
	significance threshold	proteins	interactions	average degree
MCF7	0.6	627	909	2.89
PC3	0.66	808	1064	2.63
HL60	0.72	1476	2098	2.84
SKMEL5	0.8	1476	2098	4.43

3.1 Seed proteins

The following analysis was performed for the breast cancer cell line MCF7 as it had the most test cases and is thus best suited for analyzing the proposed method. Using the four reduced graphs, the functionally specific proteins were identified. In Table 4 the 10 lowest entropy-scoring have been listed. Clearly, some of the selected proteins alone present interesting cancer related targets associated to RNA modification (PAPOA_HUMAN), cellular metabolism (PYC_HUMAN), DNA damage and repair (ZN281_HUMAN, DDB2_HUMAN), post-translational modification and localization of proteins (ENPL_HUMAN, COPA_HUMAN) and signaling pathways (IKKA_HUMAN, ABLM1_HUMAN).

Most notably of all, the *Inhibitor of nuclear factor kappa-B kinase subunit alpha* (IKKA) has been shown to be deregulated in a high percentage of breast cancers. This protein phosphorylates *β-catenin* and thus prevents its degradation [35]. Studies have reported, that β-catenin is upregulated in breast cancer and as also explained in previous sections, acts as a proto-oncogene by activating the sustained expression of proliferative genes[2, 35]. Consequently, the identification of IKKA as a functionally specific protein for the breast cancer cell line MCF7 is in agreement with these findings. Moreover, IKKA is required for the stabilization of NF-κB, another proto-oncogene, implying that its upregulation might have additional proto-oncogenic effects.

The selection of the *DNA damage-binding protein 2* (DDB2) is also interesting, as it is a critical signal transducer when genetic aberrations arise [36]. The protein has a high binding affinity for damaged DNA and is required for p21 ubiquitination. It was shown that DDB2 deficient cancer cells are immune to DNA damage induced apoptosis [37]. Considering that DDB2 is shown to be underexpressed in breast cancer in two studies [38, 39], it is an important target protein in the context of this thesis.

Heat shock proteins (HSPs) are molecular chaperones, which are activated upon cell stress, e.g. hyperthermia, viral infection, glucose deprivation, and oxidative stress. Two proteins of this family have been shown to be presented on the plasma membrane of several breast cancer cell lines – *HSP70* and *GP96*. These have been linked to natural killer (NK) cell cytotoxity and may play an important role in the immune response to malignancies [40]. Interestingly, surface GP96 was detected in the breast cancer cell lines but not in healthy tissue. One of its homologs, Endoplasmin[12] (ENPL_HUMAN), was selected as a functionally specific protein (see Table 4). In addition, another study suggests that the GRP family of proteins, and GRP94 in particular, render cancer cells immune to cytotoxic T-lymphocytes when upregulated [41]. This implies that Endoplasmin might be a key protein with regard to cancer immunology.

Malignant cells often exhibit disruptions in their aktin cytoskeleton, which render them immune to *anoikis*, a form of programmed cell death initiated when cell-matrix adhesions are lost. *TPM1_HUMAN*, a key cytoskeletal protein, can induce anoikis in neoplastic cells upon stimulation. However, it has been observed, that the expression of the protein is repressed in breast cancer cells [42]. As it also modulates the activity of integrins, its deregulation might be crucial for cancer cells acquiring tissue invasiveness. In addition, the fodrin/e-cadherin/β-catenin adhesion complex is shown to be deregulated in breast cancers, which leads to loss of cell-cell adhesion [43]. One of the components of fodrin, *SPTB2*[13], is one of the functionally specific proteins identified in this thesis, suggesting it might be involved in the breast cancer cells developing metastatic capabilities.

[12] GRP94, Tumor rejection antigen 1

[13] fodrin beta chain

Another protein interacting with the cytoskeleton and selected as functionally specific is the *Actin-binding LIM protein 1*[14](ABLM1). It possesses both F-actin and DNA binding domains[15]. There are indications that it is phosphorylated by ATM or ATR upon DNA damage, suggesting that it plays a role in deciding whether a cell should enter cell cycle arrest, induce apoptosis or survive. Moreover, its gene is located on the human chromosome region *10q25*, which is often deleted in cancer cells, implying that it may actually function as a tumor suppressor [44].

Though some of the selected seed proteins have not been shown to be cancer related, there are indications for at least some of them, that they play a role in the development and survival of malignancies. The next step is the study of their interactors and functionally related partners. This may reveal if they have oncogenic potential and hint to what their molecular function might be.

3.2 Functional modules

The functionally specific proteins were used as seeds for identifying the functional modules using the adapted DMSP algorithm. The procedure ran over the entire protein interaction network from ConsensusPathDB, annotated with edge weights (co-expression values as previously described) for the *MCF7* cell line. For the parameters p_1 and p_2 the values 0.45 and 0.9 were selected. The weighted density of the kernel graph was required to be 0.7. These values were selected according to the best practices reported in the original paper [13]. A summary of the modules identified for each seed protein is available in Table 5.

It is obvious that some of the identified modules are larger and more densely populated (see Figures 8, 9, 10, 11, 12 and 17), while others consist of a handful of proteins, with a few interactions between them (see Figures 13, 15 and 16). It is likely that these two groups present different forms of functional relationships. My assumption was that the former modules are actually multi-protein complexes, which have been resolved to cliques after the matrix model and that the latter consist of proteins, which are part of a signaling cascade. To test this thesis I manually looked up if the members of the modules are components of a common complex using the ConsensusPathDB web interface. Indeed, three of the modules contained proteins, which all participated in a common complex interaction. The subnetwork around ENPL_HUMAN is a significant part of a complex formation involving

[14] ABLM1_HUMAN

[15] zinc finger domain

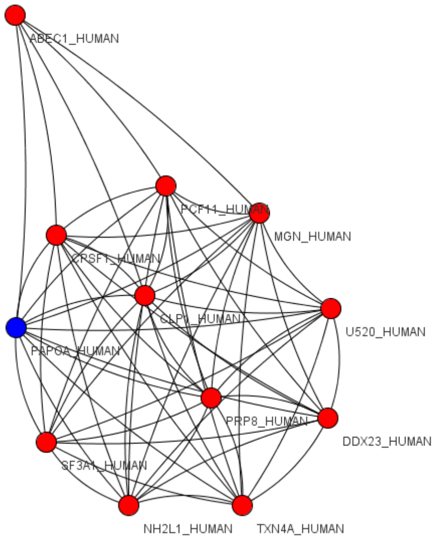

Figure 8: The module extracted from PAPOA_HUMAN. A dense subnetwork of proteins participating in four different complexes. The protein ABEC1_HUMAN is not part of any complex.

MAPK13 and the *protein kinase D*, which transduces signals downstream of *protein kinase C*. The module extracted from ZN281_HUMAN is part of the *transcription mediator complex* (MED) involving *Cyclin C* but without *CDK8*. The PYC_HUMAN module constitutes a part of the PPP2RB/CCT2 complex, involving *TIF1B*[16] and *MCM5*.

However, the results showed that the majority of the modules contain a significant number of components of at least two multi-protein complexes. In addition, single protein-protein interactions, which are not known to form stable complexes, were also contained in some subnetworks. For example, the module extracted from the seed protein PAPOA_HUMAN (see Figure 8) contains proteins that participate in several different complexes: *Cleavage Polyadenylation Complex, Intronless pre-mRNA cleavage Complex, Exon Junction Complex, Spliceosome active C complex* among others. All of these complexes share at least three proteins from the extracted subnetwork. As a result, after resolving the interactions, all these proteins form a clique, although no complex exists, in which they are all contained. The protein ABEC1_HUMAN is also part of the module, although it does not participate in any complex formation with PAPOA_HUMAN. Other modules, which exhibit

[16] Transcription intermediary factor 1-beta, KAP1

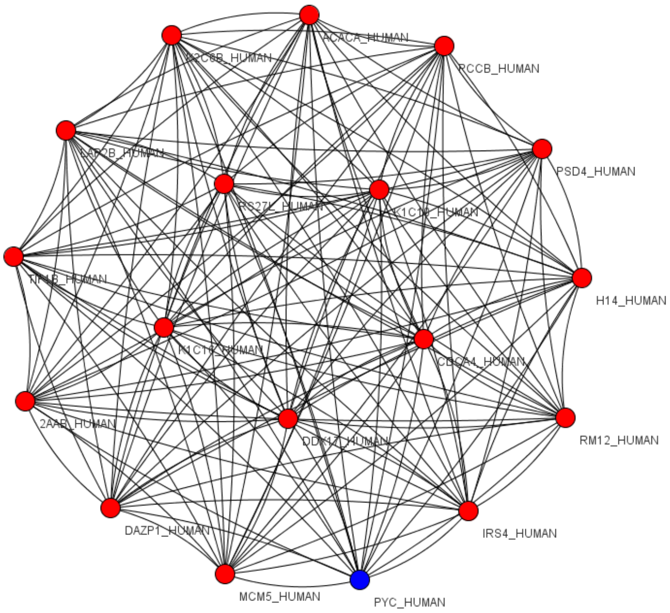

Figure 9: The functional module extracted from PYC_HUMAN. A very dense subnetwork. All participants are components of the PPP2RB/CCT2 complex.

similar composition properties, are the ones derived from IKKA_HUMAN, COPA_HUMAN and DDB2_HUMAN (see Figures 13, 15 and 16).

Of special interest are the modules extracted from the proteins TPM1_HUMAN, SPTB2_HUMAN and ABLM1_HUMAN (see Figures 12, 14 and 17). These subnetworks all contain members of the multiprotein complex involving the *Epidermal growth factor receptor I* (EGFR1) and *SHC1*, among others. This protein complex transduces the signal from the transmembrane receptor down the MAPK pathway [14]. In fact, the three modules share 23 proteins, which are part of this complex. Moreover, the module around ABLM1_HUMAN involves the *MAPK/ERK kinase kinase 2* (MEKK), while the subnetwork originating from TPM1_HUMAN contains *c-Jun* – a target of the MAPK pathway and member of the *AP-1* transcription factor [45]. This suggests a strong functional relationship between the proteins involved and highlights the importance of the MAPK pathway.

The notion that two different kinds of modules were identified appears to be incorrect. It is obvious that the extracted modules are not resolved protein complexes, as there is not one such complex that is fully contained in any module. Indeed some of the subnetworks comprise entirely of proteins that are members of a single complex, but this is due to the functional closeness of the members as defined: a result of the very high topological

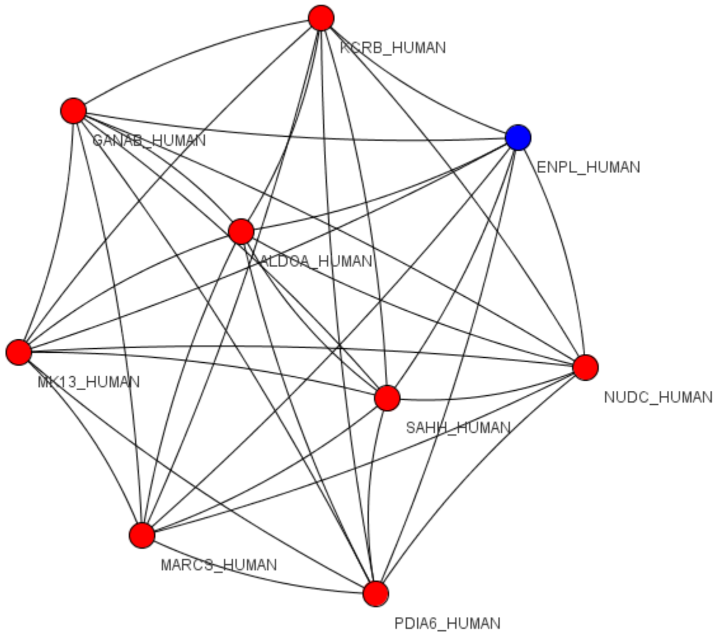

Figure 10: The functional module extracted from ENPL_HUMAN. A dense subnetwork. All participants are components of a complex interaction involving MAPK13 and the protein kinase D.

relation combined with high co-expression relations. However, there exist subnetworks that contain members of several complexes and/or proteins, which participate only in transient interactions. This suggests, that they have actually been selected according to functional relations based on the co-expression of the participants.

To more specifically identify the roles of the extracted modules, a pathway-based over-representation analysis was performed using the ConsensusPathDB web application. A p-value was calculated for each pair of pathway and module according to a hypergeometric test, based on the number of proteins in the predefined set, i.e. the pathway and the user-specified set, i.e. the module. The hypergeometric p-value equals the probability of choosing k_p or more proteins participating in a pathway p of size K_p when randomly drawing n proteins from the protein background universe of ConsensusPathDB of

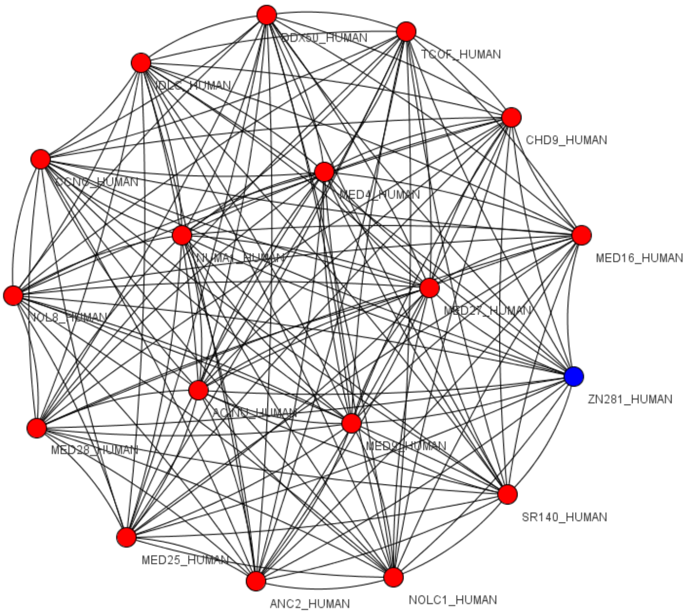

Figure 11: The functional module extracted from ZN281_HUMAN. A dense subnetwork. All participants are components of the MED complex.

size N:

$$\text{p-value} = \sum_{i=k_p}^{min(n,K_p)} \frac{\binom{K_p}{i}\binom{N-K_p}{n-i}}{\binom{N}{n}} \tag{15}$$

The lower the p-value, the higher is the statistical significance of a module in a pathway. Protein sets with a p-value lower than 0.01 were regarded as significantly enriched.

The results of the over-representation analysis are partly in agreement with the observations on the composition of the modules. As expected, the three subnetworks around TPM1_HUMAN, SPTB2_HUMAN and ABLM1_HUMAN have four common significantly enriched pathways: *Pathogenic Escherichia coli infection, Shigellosis, Bacterial invasion of epithelial cells* and *induction of apoptosis through dr3 and dr4/5 death receptors*. While these pathways seem somewhat unlikely related to the EGFR1-SHC complex discussed above, they do support the notion that the three extracted modules have common biological functions. Only the proteins of the subnetwork around ABLM1_HUMAN have been shown to be significantly enriched for the *EGFR1 pathway* with the largest overlap of proteins – 5. Another pathway in which the module is over-represented is the *JAK STAT pathway*. The subnetworks around TPM1_HUMAN and SPTB2_HUMAN are both involved in the regulation of the cytoskeleton and insulin production. The module extracted from the protein IKKA_HUMAN shows highly significant enrichment for the TNF-α pathway and other cascades, which activate the NF-κB protein, e.g. via RIP1 and/or TRAF6. The ENPL_HUMAN induced subnetwork is enriched for the signaling cascades downstream of the VEGF and NOD-like receptors. The modules extracted from PAPOA_HUMAN, PYC_HUMAN and ZN281_HUMAN are as expected over-represented in the pathways governing mRNA splicing and processing, the pyruvate metabolism and gene expression respectively. The subnetworks around COPA_HUMAN and DDB2_HUMAN are not significantly enriched for any pathway defined in ConsensusPathDB. For details see Table 6.

Table 6: The results of the pathway over-representation analysis for each module.

Module center	enriched pathways	overlap	p-value
ABLM1_HUMAN	Bacterial invasion of epithelial cells	4/73	0.000977
	EGFR1	5/180	0.00466
	Pathogenic Escherichia coli infection	3/58	0.00478
	JAK STAT pathway and regulation	2/19	0.00506
	Shigellosis	3/64	0.00672
	KitReceptor	3/65	0.00801
	induction of apoptosis through dr3 and dr4/5 death receptors	2/28	0.00919
COPA_HUMAN			
DDB2_HUMAN			
ENPL_HUMAN	VEGF	2/18	0.000201
	NOD-like receptor signaling pathway	2/62	0.00306
IKKA_HUMAN	TNFalpha	s 4/62	1.25e-05
	Viral dsRNA:TLR3:TRIF Complex Activates RIP1	2/12	4.72e-05
	human TAK1 activates NFkB by activation of IKKs complex	2/16	7.51e-05
	toll-like receptor pathway	2/38	0.000422
	TRAF6 Mediated Induction of the antiviral cytokine IFN-α/β cascade	2/54	0.000932
PAPOA_HUMAN	mRNA Splicing	11/121	3.76e-20
	Elongation and Processing of Capped Transcripts	11/149	4.96e-19
	Processing of Capped Intron-Containing Pre-mRNA	11/154	6.94e-19
	Formation and Maturation of mRNA Transcript	11 /167	2.08e-18

(Continues on the next page)

Module center	enriched pathways	overlap	p-value
	Gene Expression	11/445	2.04e-13
	Spliceosome	7/127	1.55e-10
PYC_HUMAN	Propanoate metabolism	2/32	0.00256
	Pyruvate metabolism	2/44	0.0048
SPTB_HUMAN	Insulin Synthesis and Secretion	4/115	0.00273
	Pathogenic Escherichia coli infection	3/58	0.00308
	Shigellosis	3/64	0.00436
	Regulation of actin cytoskeleton	5/216	0.00576
	Bacterial invasion of epithelial cells	3/73	0.0064
	induction of apoptosis through dr3 and dr4/5 death receptors	2/28	0.00681
	Protein export	2/24	0.00741
	IL1	2/27	0.00932
TPM1_HUMAN	Insulin Synthesis and Secretion	4/115	0.00066
	Pathogenic Escherichia coli infection	3/58	0.00104
	Shigellosis	3/64	0.00148
	Bacterial invasion of epithelial cells	3/73	0.0022
	induction of apoptosis through dr3 and dr4/5 death receptors	2/28	0.00327
	Protein export	2/24	0.00355
	Hypertrophic cardiomyopathy (HCM)	3/85	0.00357
	Dilated cardiomyopathy	3/92	0.00449
	Regulation of actin cytoskeleton	4/216	0.0078
	Diabetes pathways	5/349	0.00795
	Apoptotic cleavage of cellular proteins	2/43	0.00832
ZN281_HUMAN	Generic Transcription Pathway	5/35	6.02e-09
	Gene Expression	5/445	0.00129

Obviously, the extracted functional modules participate in a number of cancer related pathways. From the results we cannot conclude, whether the pathways and their target genes are up- or down-regulated. However, because they are over-represented in the functional modules derived from a cancer cell line, one can conclude that they are deregulated in some manner with regard to healthy tissue. The EGFR1 pathway, enriched in the ABLM1 module, stimulates cell growth and proliferation and inhibits apop-

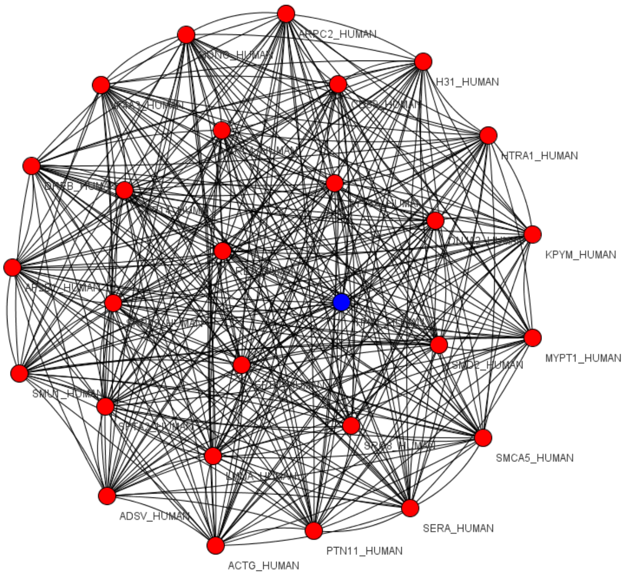

Figure 12: The functional module extracted from TPM1_HUMAN. A dense subnetwork. Participants of the module are components of the complex interaction involving EGFR1 and SHC. Contains the transcription factor c-Jun.

tosis [14, 2, 46]. It can unfold its activity via a number of pathways: *MAPK, STAT, PI3-AKT* and *PLC*γ. The EGFR itself and its pathway are a popular target of anti-cancer therapy [46, 47]. The Jak-STAT signaling pathway is also significantly enriched in the module, suggesting it can be of special importance for the oncogenic properties of the cell line. The VEGF pathway, which is over-represented in the Endoplasmin module, can initiate angiogenesis [2]. It is also an important target of cancer therapy. The activation of NF-κB also causes the upregulation of proliferative genes and may render a cell less sensitive to apoptotic signals [14, 2]. This can be achieved by any of the pathways over-represented in the IKKA_HUMAN module. Moreover, apoptosis related pathways are enriched in the three closely related modules TPM1_HUMAN, SPTB2_HUMAN and ABLM1_HUMAN, which hints towards a general deregulation of the pathway. As previously described, evading apoptosis is an essential characteristic of malignancies. Abnormalities in mRNA splicing and processing can contribute to the differential expression of oncogenes and tumor-suppressor genes. The actin cytoskeleton and its regulation are also crucial for deciding a cell's fate and for acquiring metastatic properties. The altered cellular metabolism is also a key property of malignant cells. It is comprehensible that at least some of the hallmarks of cancer are reflected by the functional modules identified.

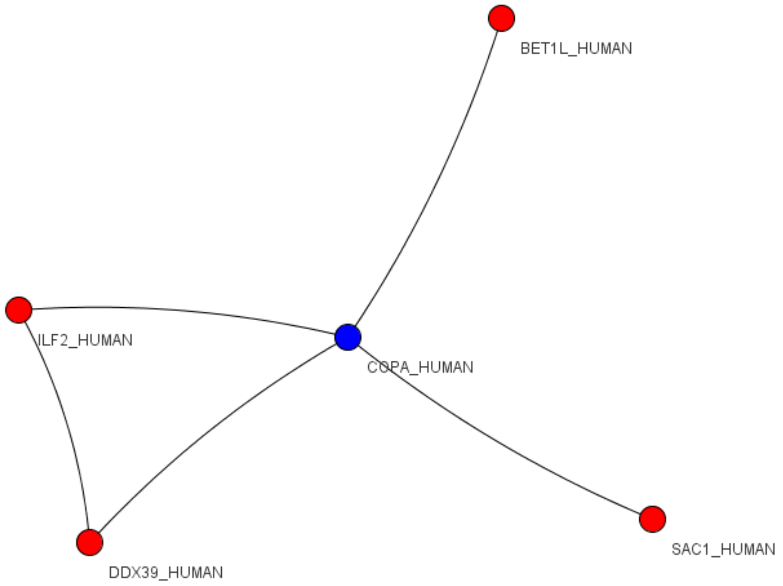

Figure 13: The functional module extracted from COPA_HUMAN. A sparse subnetwork.

3.3 Validation

To test the validity of the functional modules extracted I aimed at determining to what degree the members of each one participated in the same biological process. Other proofs of validity like connectivity density and complex coverage were not relevant to this approach, because the goal was not to locate topologically related proteins, neither to re-identify protein complexes.

The *Gene Ontology* (GO) project is an initiative of the GO Consortium to create a structured, precisely defined, common, controlled vocabulary describing the roles of genes and gene products for different organisms [48]. This is done in order to facilitate the functional annotation of molecular entities by providing an integrated, more complete and unified pool of knowledge. The project supports several organisms: *homo sapiens, mus musculus, drosophila melanogaster, saccharomyces cerevisae* and *caenorhabditis elegans* among others. Three categories describing protein function are defined. *Biological process* describes high-level objectives, which are influenced

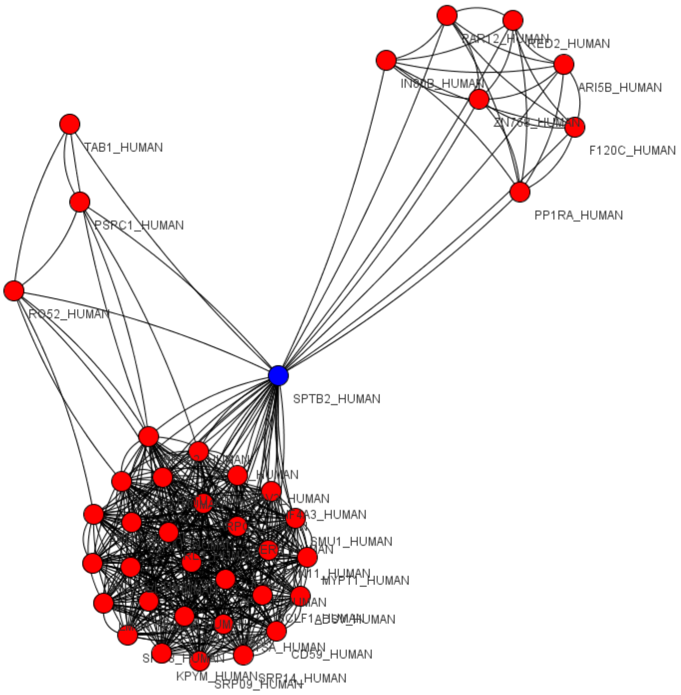

Figure 14: The functional module extracted from SPTB2_HUMAN. A dense subnetwork. Participants of the module are components of two complex interactions. One of them involves EGFR1 and SHC.

or dependent on the gene or gene product, e.g. *cell adhesion* or *pyrimidine metabolism*. *Molecular function* refers to the potential biochemical activity of a gene product, e.g. *adenylate cyclase* or *Toll receptor ligand*. It describes only what is done, but not to what end, where and when. *Cellular component* describes where a gene product unfolds its function with regard to the structure of eukaryotic cells.

In order to validate the functional modules extracted, their biological significance was quantified by determining the GO biological process term enrichment for each one. The ConsensusPathDB over-representation analysis web tool was used for level 2 GO terms. For each set of genes or gene products a p-value is calculated according to a hypergeometric test based on the number of physical entities present in the predefined set, i.e. the set of all participants in a biological process, and the user-specified list of physical entities, i.e. the extracted module. The lower the p-value, the higher is the statistical significance of a module in a GO term. Corrections for multiple testing using the false discovery rate were also calculated.

Interestingly, 9 out of 10 modules are overrepresented in at least one GO biological process with a p-value lower than 0.01 and 6 out of 10 with a q-value

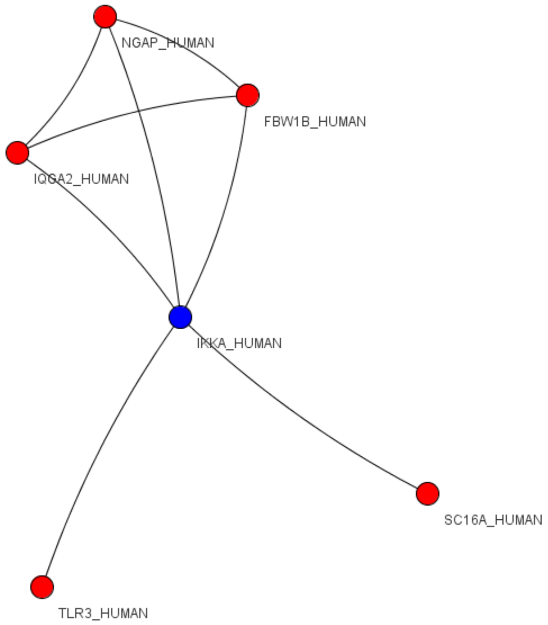

Figure 15: The functional module extracted from IKKA_HUMAN. A sparse subnetwork.

lower than 0.01. The module resulting from the seed protein DDB2_HUMAN is not enriched in any GO biological process with a p-value lower than 0.01. Five of the extracted gene sets are overrepresented in more than one GO term with 50% or more of the proteins being contained in the predefined set. These results obviusly support the notion that the modules, with the exception of one, consist of functionally related proteins, which participate in the same biological processes. Thereby, the confidence level is high, as even after the correction for multiple testing there exist protein sets, which are orders of magnitude lower than the significance threshold of 0.01.

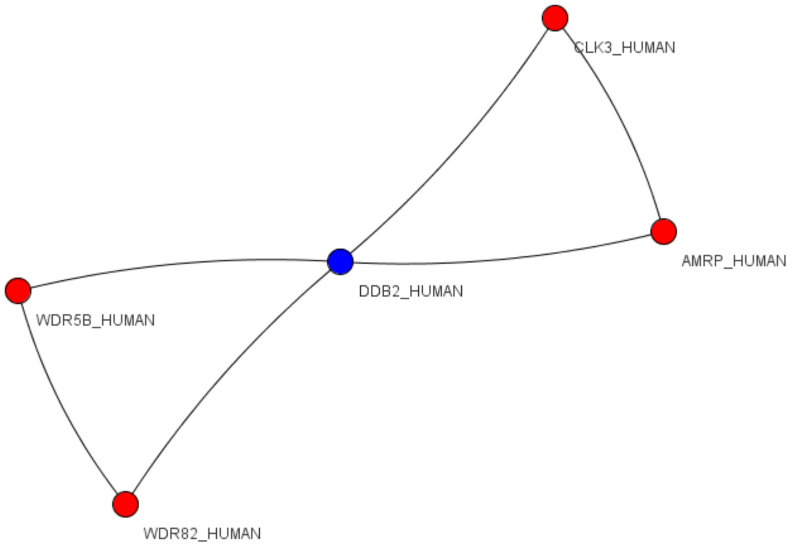

Figure 16: The functional module extracted from DDB2_HUMAN. A sparse subnetwork.

Table 4: The ten functionally specific proteins with lowest entropy. Descriptions taken from UniPROT.

Protein	Score	Description
PAPOA_HUMAN	0.0	Polymerase that creates the 3'-poly(A) tail of mRNA's. Also required for the endoribonucleolytic cleavage reaction at some polyadenylation sites.
ENPL_HUMAN	0.0	Molecular chaperone that functions in the processing and transport of secreted proteins. Functions in endoplasmic reticulum associated degradation (ERAD). Has ATPase activity.
ZN281_HUMAN	0.0	Involved in transcriptional regulation. Represses the transcription of a number of genes including gastrin and ornithine decarboxylase. Binds to the G-rich box in the enhancer region of these genes. Phosphorylated upon DNA damage, probably by ATM or ATR.
PYC_HUMAN	0.0	Catalyzes in a tissue specific manner, the initial reactions of glucose (liver, kidney) and lipid (adipose tissue, liver, brain) synthesis from pyruvate.
TPM1_HUMAN	0.0	Binds to actin filaments in muscle and non-muscle cells. Plays a central role, in association with the troponin complex, in the calcium dependent regulation of vertebrate striated muscle contraction.
COPA_HUMAN	0.102	The coatomer is a cytosolic protein complex that binds to dilysine motifs and reversibly associates with Golgi non-clathrin-coated vesicles, which further mediate biosynthetic protein transport from the ER, via the Golgi up to the trans Golgi network.
SPTB2_HUMAN	0.402	Fodrin, which seems to be involved in secretion, interacts with calmodulin in a calcium-dependent manner and is thus candidate for the calcium-dependent movement of the cytoskeleton at the membrane.
IKKA_HUMAN	0.449	Acts as part of the IKK complex in the conventional pathway of NF-kappa-B activation and phosphorylates inhibitors of NF-kappa-B thus leading to the dissociation of the inhibitor/NF-kappa-B complex and ultimately the degradation of the inhibitor.
DDB2_HUMAN	0.453	Required for DNA repair. Binds to DDB1 to form the UV-damaged DNA-binding protein complex (the UV-DDB complex). The UV-DDB complex may recognize UV-induced DNA damage and recruit proteins of the nucleotide excision repair pathway (the NER pathway) to initiate DNA repair.
ABLM1_HUMAN	0.456	May act as scaffold protein. Has been suggested to play a role in axon guidance.

Table 5: The components of the ten functional modules derived from the corresponding seed protein.

Seed protein	module participants
PAPOA_HUMAN	ABEC1_HUMAN, PRP8_HUMAN, DDX23_HUMAN, TXN4A_HUMAN, NH2L1_HUMAN, CPSF1_HUMAN, MGN_HUMAN, PCF11_HUMAN, U520_HUMAN, CLP1_HUMAN, SF3A1_HUMAN
ENPL_HUMAN	ALDOA_HUMAN, KCRB_HUMAN, SAHH_HUMAN, NUDC_HUMAN, MK13_HUMAN, PDIA6_HUMAN, MARCS_HUMAN, GANAB_HUMAN
ZN281_HUMAN	NUMA1_HUMAN, TCOF_HUMAN, ACINU_HUMAN, CCNC_HUMAN, CHD9_HUMAN, MED25_HUMAN, SR140_HUMAN, NOLC1_HUMAN, MED28_HUMAN, DDX50_HUMAN, ANC2_HUMAN, MED4_HUMAN, MED27_HUMAN, MED16_HUMAN, NOL8_HUMAN, MED9_HUMAN, IDLC_HUMAN
PYC_HUMAN	MCM5_HUMAN, 2AAB_HUMAN, PCCB_HUMAN, ACACA_HUMAN, H14_HUMAN, RS27L_HUMAN, CDCA4_HUMAN, K1C10_HUMAN, DAZP1_HUMAN, TIF1B_HUMAN, IRS4_HUMAN, K1C16_HUMAN, DDX17_HUMAN, PSD4_HUMAN, RM12_HUMAN, LAP2B_HUMAN, K2C6B_HUMAN
TPM1_HUMAN	PTN11_HUMAN, ADSV_HUMAN, NONO_HUMAN, SRP09_HUMAN, RL4_HUMAN, SRC8_HUMAN, SMD2_HUMAN, CYTSA_HUMAN, DREB_HUMAN, LMNA_HUMAN, SMCA5_HUMAN, ACTG_HUMAN, SPTA2_HUMAN, SERA_HUMAN, ARPC2_HUMAN, EFTU_HUMAN, MYPT1_HUMAN, HTRA1_HUMAN, DNJA2_HUMAN, SMU1_HUMAN, IF4A3_HUMAN, AP2B1_HUMAN, SRP14_HUMAN, H31_HUMAN, RL3_HUMAN, KPYM_HUMAN, CD59_HUMAN
COPA_HUMAN	ILF2_HUMAN, DDX39_HUMAN, SAC1_HUMAN, BET1L_HUMAN
SPTB2_HUMAN	PTN11_HUMAN, PAR12_HUMAN, ADSV_HUMAN, NONO_HUMAN, SRP09_HUMAN, PSPC1_HUMAN, RL4_HUMAN, SRC8_HUMAN, SMD2_HUMAN, CYTSA_HUMAN, LMNA_HUMAN, DREB_HUMAN, RO52_HUMAN, F120C_HUMAN, SMCA5_HUMAN, TAB1_HUMAN, ACTG_HUMAN, SPTA2_HUMAN, SERA_HUMAN, ARPC2_HUMAN, EFTU_HUMAN, MYPT1_HUMAN, IN80B_HUMAN, HTRA1_HUMAN, DNJA2_HUMAN, SMU1_HUMAN, AP2B1_HUMAN, IF4A3_HUMAN, BCLF1_HUMAN, SRP14_HUMAN, H31_HUMAN, KPYM_HUMAN, RL3_HUMAN, VAV3_HUMAN, CD59_HUMAN, RED2_HUMAN, PP1RA_HUMAN, ARI5B_HUMAN, ZN768_HUMAN
IKKA_HUMAN	IQGA2_HUMAN, TLR3_HUMAN, SC16A_HUMAN, FBW1B_HUMAN, NGAP_HUMAN
DDB2_HUMAN	AMRP_HUMAN, CLK3_HUMAN, WDR82_HUMAN, WDR5B_HUMAN
ABLM1_HUMAN	PTN11_HUMAN, NONO_HUMAN, SRP09_HUMAN, SRC8_HUMAN, IF4E2_HUMAN, SMD2_HUMAN, ASPP2_HUMAN, CYTSA_HUMAN, LMNA_HUMAN, DREB_HUMAN, SMCA5_HUMAN, ACTG_HUMAN, MAST2_HUMAN, RABE1_HUMAN, ATPA_HUMAN, SPTA2_HUMAN, LC7L2_HUMAN, SERA_HUMAN, ARPC2_HUMAN, VAMP8_HUMAN, EFTU_HUMAN, MYPT1_HUMAN, HTRA1_HUMAN, DNJA2_HUMAN, OSBL3_HUMAN, M3K2_HUMAN, SMU1_HUMAN, AP2B1_HUMAN, IF4A3_HUMAN, LARP1_HUMAN, BCLF1_HUMAN, SRP14_HUMAN, H31_HUMAN, KPYM_HUMAN, RL3_HUMAN, LSR_HUMAN, TRIPB_HUMAN, MAST3_HUMAN, CBLB_HUMAN, TR150_HUMAN, UBP8_HUMAN, CLAP1_HUMAN, FOXO3_HUMAN, CP250_HUMAN

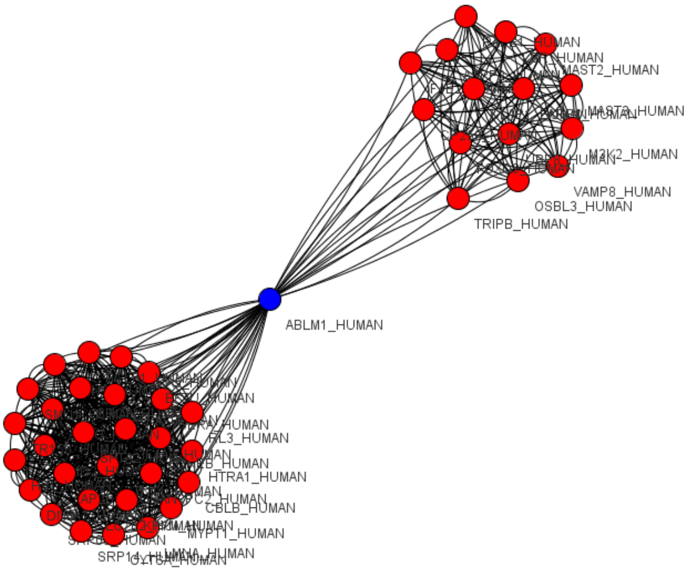

Figure 17: The functional module extracted from ABLM1_HUMAN. A dense subnetwork. Participants of the module are components of two complex interactions. One of these involves the proteins EGFR1 and SHC. Contains MEKK.

4 Discussion

As described above, the majority of the functional modules extracted consist of plausible sets of proteins, which have a common biological function and are enriched for some cancer-related pathways. Indeed, all of them contain proteins and components of protein complexes that have been shown to participate in oncogenesis and breast cancer development in particular. This being the goal set beforehand, the method proposed seems to be delivering reasonable results.

Among the proteins in the functional modules are regulators of the cell cycle (Cyclin C, MCM5), transcription factor associated proteins (TIF1B, c-Jun) and signal transducers from major pathways (SHC, MEKK, MAPK13, IKKA, Protein kinase D etc). Proteins of these groups are often the targets of new cancer therapies. Some of their potential as drug targets is not obvious, e.g. the MED-complex is a rather unlikely candidate with regard to cancer. However, it forms a complex with CCAR1, which upon stimulation from the Estrogen receptor activates the transcription of proliferative genes. This is one of the main pathways leading to cell growth in breast cancers [49]. Other involved proteins are well known key players in breast cancer like the SRC8 protein.

However, there are some subnetworks that could not be validated using the over-representation analysis. No GO biological process is significantly over-represented in the DDB2_HUMAN module and no pathways defined in ConsensusPathDB exist, for which it is enriched (see Table 6). These facts let the module seem incorrectly composed with regard to functional relations between its proteins. However, this can be explained by the small size of the subnetwork (the smallest of the ten) and lack of annotation of the participating proteins in ConsensusPathDB. In general, the validation of the modules is problematic and additional approaches have to be applied in order to verify the functional relationships between the proteins. Hopefully, this will provide better means of assessing the results of the method.

Ideally, the functional subnetworks will be validated in a wet laboratory by means of knock-out experiments, differential expression analysis of malignant versus healthy tissue or others. Comparison to other computational approaches [10, 11, 12] might also provide new insights into the validity and role of the functional modules. To achieve more coherent results, these algorithms can be applied on the same list of seed proteins.

There are several aspects of the proposed approach, which can be improved. The first part concerns the input data used. As previously described, the complex interactions from the ConsensusPathDB were resolved to binary interactions according to the matrix model. The advantage of this model is that no interactions are misrepresented, as all possible interactions are generated. However, this is done at the cost of generating a large number of false interactions. This is also reflected in the very high average degree of the ConsensusPathDB interaction graph. Moreover, matrix topologies are unlikely for larger complexes from a biological point of view, because of steric hindrance. From this point of view the spoke model is more intuitive, but it could misrepresent interactions. In addition, the available protein interaction network provides no information on which is the center/bait protein. As both models are problematic, using the matrix model was reasonable, as it is important for the DMSP algorithm not to miss existing interactions. The bias towards selecting such resolved complex interactions can be somewhat counterbalanced by the co-expression weights. This effect can be observed in the structure of the reduced graphs (see Table 3), which in contrast to the entire network all have average degrees close to the expected 2.5.

The protein interaction network provides another issue worth discussing. Namely, only about 10% of the human interactome have been identified thus far [50]. This fact, combined with the high false positive and false negative rates for Y2H, 50% and 90% respectively [51, 52], have led some scientists to the conclusion that only global network features can be reliably analyzed [22]. Therefore, it is important to keep in mind that any study of local network structures has limited credibility and is relevant only with regard to the contemporary knowledge.

Another critique of the input data are the expression profiles. While these provide a good estimate on the amount of proteins in the cell, they can also mislead, as there are several control steps on later stages on the path from DNA to protein, e.g. mRNA localization, translational regulation etc. Such control effects cannot be detected by microarrays, which only measure mRNA levels. If large-scale protein-array data were available, much more credible results could be achieved.

The second part of possible improvements concerns the method itself and more precisely the DMSP algorithm. While it does perform reasonably well, it is unclear to what extend the parameters p_1 and p_2 affect the results. As previously mentioned, the best-practice values published in the original paper[13] were used for the two parameters. However, it is unclear how these values were selected. Therefore, it might be of interest to test the outcome with variations of the two parameters. Assessing the results however is problematic, because of the validation issue discussed above.

Finally, with regard to the specific configuration of the *in silico* experiment, it might be of interest to analyze the modules around the functionally unspecific proteins. Because the four phenotypes used are all cancer cell lines, they certainly share some malignant characteristics and it can be presumed that they also have a few common molecular aberrations. For example, it is known that the tumor suppressor *p53* is deregulated in a high percentage of all known cancer types. Therefore, identifying functional modules common to different malignant phenotypes is reasonable. However, one must keep in mind that a great part of the modules will not be cancer related at all due to the configuration of the microarray experiment, which does not compare malignant and healthy tissue.

Conclusion

The novel method proposed in this thesis is able to identify functional modules making use of large-scale interaction and expression data. While the verification of these subnetworks remains difficult, they identify sets of proteins, which may play a key role in the development of malignancies. The modules provide an attractive basis for further research and may provide new targets for drug development.

References

[1] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, and M. J. Thun. Cancer statistics, 2009. *CA Cancer J Clin*, 59:225–249, 2009.

[2] D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *Cell*, 100: 57–70, Jan 2000.

[3] T. Ideker and R. Sharan. Protein networks in disease. *Genome Res.*, 18:644–652, Apr 2008.

[4] M. G. Kann. Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Brief. Bioinformatics*, 8: 333–346, Sep 2007.

[5] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, Oct 1999.

[6] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, Feb 2000.

[7] S. Ramaswamy, K. N. Ross, E. S. Lander, and T. R. Golub. A molecular signature of metastasis in primary solid tumors. *Nat. Genet.*, 33:49–54, Jan 2003.

[8] L. Ein-Dor, O. Zuk, and E. Domany. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl. Acad. Sci. U.S.A.*, 103:5923–5928, Apr 2006.

[9] S. A. Tomlins, D. R. Rhodes, S. Perner, S. M. Dhanasekaran, R. Mehra, X. W. Sun, S. Varambally, X. Cao, J. Tchinda, R. Kuefer, C. Lee, J. E. Montie, R. B. Shah, K. J. Pienta, M. A. Rubin, and A. M. Chinnaiyan. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, 310:644–648, Oct 2005.

[10] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18 Suppl 1:S233–240, 2002.

[11] J. Chen and B. Yuan. Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics*, 22:2283–2290, Sep 2006.

[12] H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee, and T. Ideker. Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, 3:140, 2007.

[13] I. A. Maraziotis, K. Dimitrakopoulou, and A. Bezerianos. Growing functional modules from a seed protein via integration of protein interaction and gene expression data. *BMC Bioinformatics*, 8:408, 2007.

[14] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell*. Garland Science, 4 edition, 2004. ISBN 0-8153-4072-9.

[15] S. Fields. High-throughput two-hybrid analysis. The promise and the peril. *FEBS J.*, 272:5391–5399, Nov 2005.

[16] L. Giot, J. S. Bader, C. Brouwer, A. Chaudhuri, B. Kuang, Y. Li, Y. L. Hao, C. E. Ooi, B. Godwin, E. Vitols, G. Vijayadamodar, P. Pochart, H. Machineni, M. Welsh, Y. Kong, B. Zerhusen, R. Malcolm, Z. Varrone, A. Collis, M. Minto, S. Burgess, L. McDaniel, E. Stimpson, F. Spriggs, J. Williams, K. Neurath, N. Ioime, M. Agee, E. Voss, K. Furtak, R. Renzulli, N. Aanensen, S. Carrolla, E. Bickelhaupt, Y. Lazovatsky, A. DaSilva, J. Zhong, C. A. Stanyon, R. L. Finley, K. P. White, M. Braverman, T. Jarvie, S. Gold, M. Leach, J. Knight, R. A. Shimkets, M. P. McKenna, J. Chant, and J. M. Rothberg. A protein interaction map of Drosophila melanogaster. *Science*, 302:1727–1736, Dec 2003.

[17] A. C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck, B. Dümpelfeld, A. Edelmann, M. A. Heurtier, V. Hoffman, C. Hoefert, K. Klein, M. Hudak, A. M. Michon, M. Schelder, M. Schirle, M. Remor, T. Rudi, S. Hooper, A. Bauer, T. Bouwmeester, G. Casari, G. Drewes, G. Neubauer, J. M. Rick, B. Kuster, P. Bork, R. B. Russell, and G. Superti-Furga. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 440:631–636, Mar 2006.

[18] A. C. Gavin, M. Bösche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, C. M. Cruciat, M. Remor,

C. Höfert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415:141–147, Jan 2002.

[19] U. Stelzl, U. Worm, M. Lalowski, C. Haenig, F. H. Brembeck, H. Goehler, M. Stroedicke, M. Zenkner, A. Schoenherr, S. Koeppen, J. Timm, S. Mintzlaff, C. Abraham, N. Bock, S. Kietzmann, A. Goedde, E. Toksöz, A. Droege, S. Krobitsch, B. Korn, W. Birchmeier, H. Lehrach, and E. E. Wanker. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122:957–968, Sep 2005.

[20] J. F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437:1173–1178, Oct 2005.

[21] R. Albert and AL. Barabasi. Statistical mechanics of complex networks. *Review of Modern Physics*, 74:47–97, 2002.

[22] J. Berg, M. Lassig, and A. Wagner. Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. *BMC Evol. Biol.*, 4:51, Nov 2004.

[23] M. Tyers and M. Mann. From genomics to proteomics. *Nature*, 422: 193–197, Mar 2003.

[24] H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41–42, May 2001.

[25] R. J. Britten and E. H. Davidson. Gene regulation for higher cells: a theory. *Science*, 165:349–357, Jul 1969.

[26] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470, Oct 1995.

[27] R. J. Lipshutz, S. P. Fodor, T. R. Gingeras, and D. J. Lockhart. High density synthetic oligonucleotide arrays. *Nat. Genet.*, 21:20–24, Jan 1999.

[28] 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*, 4:249–264, Apr 2003.

[29] W. Huber, A. von Heydebreck, H. Sueltmann, A. Poustka, and M. Vingron. Parameter estimation for the calibration and variance stabilization of microarray data. *Stat Appl Genet Mol Biol*, 2:Article3, 2003.

[30] A. Kamburov, C. Wierling, H. Lehrach, and R. Herwig. ConsensusPathDB–a database for integrating human functional interaction networks. *Nucleic Acids Res.*, 37:D623–628, Jan 2009.

[31] Z. Wu, R. A. Irizarry, R. Gentleman, F. Martinez-Murillo, and F. Spencer. A model-based background adjustment for oligonucleotide expression arrays. *J. of the American Statistical Association*, 99:907–917, Dec 2004.

[32] M. Dai, P. Wang, A. D. Boyd, G. Kostov, B. Athey, E. G. Jones, W. E. Bunney, R. M. Myers, T. P. Speed, H. Akil, S. J. Watson, and F. Meng. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.*, 33:e175, 2005.

[33] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.*, 95:14863–14868, Dec 1998.

[34] X. Zhou, M. C. Kao, and W. H. Wong. Transitive functional annotation by shortest-path analysis of gene expression data. *Proc. Natl. Acad. Sci. U.S.A.*, 99:12783–12788, Oct 2002.

[35] P. Cowin, T. M. Rowlands, and S. J. Hatsell. Cadherins and catenins in breast cancer. *Curr. Opin. Cell Biol.*, 17:499–508, Oct 2005.

[36] S. Bagchi and P. Raychaudhuri. Damaged-DNA Binding Protein-2 Drives Apoptosis Following DNA Damage. *Cell Div*, 5:3, 2010.

[37] T. Stoyanova, N. Roy, D. Kopanja, S. Bagchi, and P. Raychaudhuri. DDB2 decides cell fate following DNA damage. *Proc. Natl. Acad. Sci. U.S.A.*, 106:10690–10695, Jun 2009.

[38] A. L. Richardson, Z. C. Wang, A. De Nicolo, X. Lu, M. Brown, A. Miron, X. Liao, J. D. Iglehart, D. M. Livingston, and S. Ganesan. X chromosomal abnormalities in basal-like human breast cancer. *Cancer Cell*, 9: 121–132, Feb 2006.

[39] E. Huang, S. H. Cheng, H. Dressman, J. Pittman, M. H. Tsou, C. F. Horng, A. Bild, E. S. Iversen, M. Liao, C. M. Chen, M. West, J. R. Nevins, and A. T. Huang. Gene expression predictors of breast cancer outcomes. *Lancet*, 361:1590–1596, May 2003.

[40] K. Melendez, E. S. Wallen, B. S. Edwards, C. D. Mobarak, D. G. Bear, and P. L. Moseley. Heat shock protein 70 and glycoprotein 96 are differentially expressed on the surface of malignant and nonmalignant breast cells. *Cell Stress Chaperones*, 11:334–342, 2006.

[41] G. Gazit, J. Lu, and A. S. Lee. De-regulation of GRP stress protein expression in human breast cancer cell lines. *Breast Cancer Res. Treat.*, 54:135–146, Mar 1999.

[42] S. Bharadwaj, R. Thanawala, G. Bon, R. Falcioni, and G. L. Prasad. Resensitization of breast cancer cells to anoikis by tropomyosin-1: role of Rho kinase-dependent cytoskeleton and adhesion. *Oncogene*, 24:8291–8303, Dec 2005.

[43] R. T. Sormunen, A. S. Leong, J. P. Vaaraniemi, S. S. Fernando, and S. M. Eskelinen. Immunolocalization of the fodrin, E-cadherin, and beta-catenin adhesion complex in infiltrating ductal carcinoma of the breast-comparison with an in vitro model. *J. Pathol.*, 187:416–423, Mar 1999.

[44] A. C. Kim, L. L. Peters, J. H. Knoll, C. Van Huffel, S. L. Ciciotte, P. W. Kleyn, and A. H. Chishti. Limatin (LIMAB1), an actin-binding LIM protein, maps to mouse chromosome 19 and human chromosome 10q25, a region frequently deleted in human cancers. *Genomics*, 46: 291–293, Dec 1997.

[45] J. Hess, P. Angel, and M. Schorpp-Kistner. AP-1 subunits: quarrel and harmony among siblings. *J. Cell. Sci.*, 117:5965–5973, Dec 2004.

[46] L. Ferrer-Soler, A. Vazquez-Martin, J. Brunet, J. A. Menendez, R. De Llorens, and R. Colomer. An update of the mechanisms of resistance to EGFR-tyrosine kinase inhibitors in breast cancer: Gefitinib (Iressa) -induced changes in the expression and nucleo-cytoplasmic trafficking of HER-ligands (Review). *Int. J. Mol. Med.*, 20:3–10, Jul 2007.

[47] C. T. Kuan, C. J. Wikstrand, and D. D. Bigner. EGF mutant receptor vIII as a molecular target in cancer therapy. *Endocr. Relat. Cancer*, 8: 83–96, Jun 2001.

[48] 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.*, 25:25–29, May 2000.

[49] J. H. Kim, C. K. Yang, K. Heo, R. G. Roeder, W. An, and M. R. Stallcup. CCAR1, a key regulator of mediator complex recruitment to nuclear receptor transcription complexes. *Mol. Cell*, 31:510–519, Aug 2008.

[50] G. T. Hart, A. K. Ramani, and E. M. Marcotte. How complete are current yeast and human protein-interaction networks? *Genome Biol.*, 7:120, 2006.

[51] P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, and J. M. Rothberg. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 403:623–627, Feb 2000.

[52] P. Legrain, J. Wojcik, and J. M. Gauthier. Protein–protein interaction maps: a lead towards cellular functions. *Trends Genet.*, 17:346–352, Jun 2001.

Acknowledgments

Here I want to express my gratitude to all those, who have helped me complete this Master Thesis.

I want to especially thank Atanas Kamburov for his numerous corrections, advises and support, both during the actual work process and when writing the thesis.

I am also grateful to Dr. Ralf Herwig for his council regarding the topic of the thesis and the interpretation of the results.

Last but not least, I want to thank all the people at the Bioinformatics group of the Max Planck Institute for Molecular Genetics, with whom I have worked.

Declaration

I confirm that all this work is my own except where indicated, and that I have:

1. Clearly referenced/listed all sources as appropriate
2. Given the sources of all pictures, data etc. that are not my own
3. Not made any use of the report(s) or essay(s) of any other student(s) either past or present
4. Not sought or used the help of any external professional agencies for the work

Date: . . . .   Signature: . . . .