Dissertation

submitted to the

Combined Faculties for the Natural Sciences and for Mathematics

of the Ruperto-Carola University of Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

Presented by:

Kannabiran Nandakumar M.Sc.

Birth place : Chennai, India

February, 2010

Pattern recognition of gene expression data on signalling networks of cancer

Supervisor : Dr. Rainer König

Referees : Prof. Dr. Roland Eils

Prof. Dr. Manfred Schwab

Abstract

Cancer is a result of aberrant cellular signalling. Understanding the properties of these complex networks will enable us to design effective therapeutic strategies against cancer. Often, singular pathways are analyzed to study cancer signalling. This kind of analysis eludes the idea of orchestrated roles of signalling proteins in a network. In the analysis presented in this thesis, a network approach is used to obtain an understanding of the intricate cellular signalling.

In this thesis a sophisticated embedding of human cancer gene expression data onto the human protein-protein interaction network has been performed and pathways were predicted using a graph theoretic approach. Several network properties of normal and cancer signalling were derived from these predicted pathways using 10 cancer datasets. It is shown that the predicted cancer pathways used shorter cascades and more differentiated signalling routes when compared to predicted normal pathways. The cancer signalling network is more differentiated and much more interconnected when compared to the normal cells. Also, the cancer signalling network is less dependent on hubs compared to the normal network.

A network based analysis has been done to compare the different network properties between the normal and cancer cells using several cancer gene expression datasets. All the findings well approve a model of less ordered signalling in cancer leading to more robustness. Finally, from the insights obtained by this study novel signalling motifs have been proposed which were found with high abundance in the analysed data.

Zusammenfassung

Krebs ist ein Ergebnis abweichender zellulärer Signalübertragungen. Das Verständnis der Eigenschaften dieser komplexen Netzwerke wird es ermöglichen, effiziente therapeutische Strategien zu entwickeln. Oft werden bei der Analyse von Tumoren nur einzelne Signalpfade berücksichtigt. Diese Art Analyse vernachlässigt das Prinzip zusammenhängender Signalproteine in einem Netzwerk. Die Analyse, die in dieser Dissertation beschrieben wird, verwendet einen auf Netzwerken basierenden Ansatz, um ein Verständnis der komplexen zellulären Signaltransduktionspfade (sog. *Signalwege*) zu ermöglichen.

In dieser Dissertation wurden menschliche Tumor-Genexpressionsdaten in das menschliche Protein-Protein-Interaktionsnetzwerk eingebettet und *Signalwege* mittels eines auf der Graphentheorie basierenden Ansatzes vorausberechnet. Mehrere Eigenschaften von normalen und Tumorsignalnetzwerken wurden aus diesen berechneten *Signalwege* unter Verwendung von 10 Tumordatensätzen abgeleitet. Es wird gezeigt, dass die *Signalwege* der betrachteten Tumore verglichen mit denen in normalen Gewebe kürzere Kaskaden und stärker differenzierte Signalwege verwenden. Das Signalnetzwerk im Tumor ist allgemein differenzierter und stärker vernetzt als in normalen Zellen.

Eine netzwerkbasierende Analyse wurde ausgeführt, um die verschiedenen Netzwerkeigenschaften zwischen normalen und Tumorzellen mittels mehrerer Tumorgenexpressions-Datensätzen zu vergleichen. Die Ergebnisse bestätigen ein Model weniger geordneter *Signalwege* in Tumoren, was in einer größeren Robustheit der *Signalwege* des Tumors resultiert. Mit den Erkenntnissen dieser Studie wird ein neues Signalübertragungsmotiv vorgeschlagen, das sich in hoher Anzahl in den analysierten Datensätzen findet.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Scope

In the cell, the transfer of a signal from the receptor to a transcription factor involves many intermediary players. The entire system of signalling involving signalling proteins can be modelled as a network providing a systems approach to model cells (Xia, et al., 2004). The different patterns of signalling in cells lead to the different hall marks of cancer progression which include self-sufficiency in growth signals, insensitivity to growth-inhibitory (anti-growth) signals, evasion of programmed cell death (apoptosis), limitless replicative potential, sustained angiogenesis, and tissue invasion and metastasis (Hanahan and Weinberg, 2000). Understanding the signalling networks in cancer can provide means to novel therapeutic strategies. Microarray gene expression data provide a wealth of information about gene expression profiles in cells. Comparative analysis of gene expression profiles of normal and cancer conditions could provide differences in the expression patterns between them.

Several methods have been developed to combine the information from gene expression data and the underlying network topology for the analysis of gene expression data (Chuang, et al., 2007; Ergun, et al., 2007; Konig, et al., 2006; Schramm, et al., 2007). We mapped the gene expression data of human normal and cancer cells to a human protein-protein interaction network. The mapping was done using the Pearson correlation values calculated from the gene expression data of neighbouring proteins. This network with edge weights was later used to calculate highly correlated paths from receptor to transcription factor proteins in the network using a graph theoretic method.

The analysis was performed on 10 datasets comprising of 9 different cancer types. Several graph attributes were compared and results showed significant differences in the signalling pattern between normal and cancer cells. These findings were embedded into signalling motifs which are one of the major findings of the study. One signalling motif showed less ordered signalling in cancer due to usage of dense interconnectivity of the signalling network.

The first section gives an introduction in to basic topics covered in the thesis. The second section describes the method in detail. The results are presented in the third section and discussed in the fourth section.

1.1 Network Properties

1.1.1 Network descriptors

In mathematical terms a network is called a *graph*. There are two kinds of graphs directed and undirected. A *directed graph* G is a pair G = (V,E) where V is a finite set and E is a binary relation on V. The graph G comprises of a set V of *vertices* or *nodes* and a set E of *edges*. In a directed graph G=(V,E), the edge set E consists of ordered pairs of vertices. In the Figure 1 the directed graph has a vertex set {a,b,c,d,e,f} where vertices are represented by circles and edges by arrows.

In an *undirected graph* G = (V,E), the edge set E consists of unordered pairs of vertices. That is, the edge is a set (u,v) where $u,v \in V$ and $u \neq v$. Figure 2 represents an undirected graph with vertex set {a,b,c,d,e,f}. The *degree* k_i of a vertex is the number of edges that connect to it or the number of vertices in its neighbourhood N_i . In a directed graph, the *out-degree* of a vertex is the number of edges leaving it and the *in-degree* is the number of edges coming into it (Cormen, et al., 1990). The *distance* between two nodes in a graph is the number of nodes connecting them in a shortest path.

Figure 1. Directed network

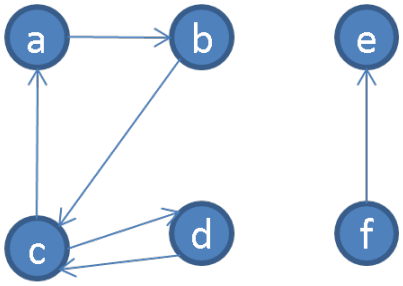
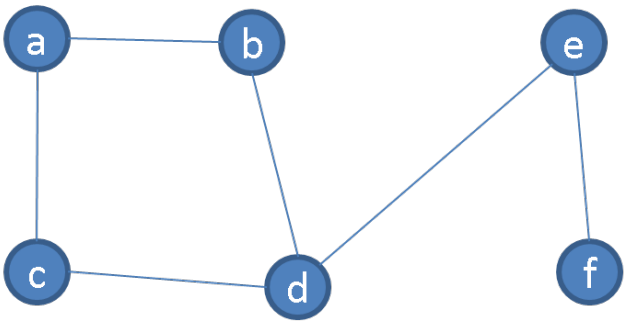


Figure 2. Undirected network



The *clustering coefficient* of a graph is a measure which evaluates the degree to which nodes cluster in a graph . The clustering coefficient for node i was given by

$$C_i = \frac{n_{link}}{k(k-1)} \quad (1)$$

in which n_{link} is the number of links connecting the neighbours of node i. k is the number of neighbours. This feature describes how good the neighbours are connected within each other. If they are fully connected, the clustering coefficient is one, if they were not connected at all, the clustering coefficient is zero.

One of the problems in graph theory is finding the shortest paths between two vertices in a graph. In shortest-path problem we are given a weighted, directed graph G = (V,E) with weight function *w* : E -> R mapping edges to real valued weights. Dijkstra's algorithm solves the single source shortest path problem for a graph with non-negative edge weights. The algorithm finds the shortest path from a source vertex to every other vertex in the graph. A simple description of the steps involved in the algorithm are given below:

1. Every node is assigned a distance value. The source node is set to zero and all other nodes set to infinity.
2. All nodes are marked as unvisited and the source node as current node.
3. The unvisited neighbours of the current node are considered and the distance from the current node calculated. For instance, if current node X has a distance of 0 (starting node)

and the edge connecting to Y has a weight of 2 then the distance of Y will be 0+2=2. If this distance were less than the previously recorded distance (infinity in the beginning, zero for the initial node), it is overwritten.

4. If all the neighbouring nodes of the current node are considered then it is marked as visited and never considered again. It has the minimal distance from the source node.
5. The unvisited node with the smallest distance from the source node is set as the next "current node" and the continued from step3.

1.1.2 Cellular networks

Recent advances in the theory of complex networks has lead to the uncovering the organizing principles that govern the formation and evolution of various complex technological and social networks (Barabasi and Oltvai, 2004; Strogatz, 2001). The impact of these studies on cell biology have contributed to a better understanding of cellular networks. The same architectural features of molecular interaction networks within a cell can be observed in other complex networks like the internet and social networks. Since the same governing principles are observed across complex networks, available knowledge of other well studied complex networks can be used to study cellular function by applying it to cellular networks.

Cellular networks can be studied under three main network categories – the transcriptional regulatory network, cell signalling network and the metabolic network.

Transcriptional regulatory network: Transcription factors regulate other genes through transcription. These relationships can be modelled as a network. The principles governing the topological organization of the regulatory network of yeast was elucidated (Guelzim, et al., 2002). A map of regulator-gene of yeast interactions was used to describe potential pathways used by yeast cells to regulate global gene expression programs and network motifs were identified from this information (Lee, et al., 2002).

Cell signalling network: The cell signalling network comprises the network of interactions between signalling proteins in a cell to carry out various cellular functions. The reconstruction of signalling networks can have different approaches like reconstruction of highly connected nodes in networks, reconstruction of linear pathways

with signalling input and outputs and reconstruction of functional signalling modules(Papin, et al., 2005).

Metabolic network: Metabolic network comprises the connections between different metabolic reactions in a cell. Metabolic network constituents though are different for different organisms, show the same topological scaling properties. Also they are robust and error-tolerant scale-free networks (Jeong, et al., 2000).

1.1.3 Networks are scale-free

Many networks have node connectivities that follow a power-law distribution and such networks are called scale-free networks (Barabasi and Albert, 1999). The degree of a vertex quantifies individual nodes, the diversity of the entire network can be quantified using a *degree distribution*. The degree distribution *P(k)* of a network gives the fraction of nodes that have degree *k* and is obtained by counting the number of nodes *N(k)* that have *k* = 1, 2, 3... edges and dividing it by the total number of nodes *N*. The degree distributions of numerous networks, such as the Internet, human collaboration networks and metabolic networks, follow a well-defined functional form $P(k) = Ak^{-\gamma}$ called a power law. Here, A is a constant that ensures that the *P(k)* values add up to 1, and the degree exponent γ is usually in the range $2 < \gamma < 3$. The cellular networks also share organizational features with many non-biological networks and have a scale-free topology (Albert, 2005).

The highly heterogeneous scale-free topology of complex networks makes it tolerant to random errors. If random nodes were removed in a scale-free network it did not affect the diameter of the network much. Even when up to 5% of the nodes fail, the communication between the remaining nodes are unaffected. However, when the highly connected nodes are targeted the diameter increases in size and they form isolated clusters. This nature of robustness of complex networks is due to the in-homogenous connectivity distribution, where the probability of choosing a highly connected node is less because they are less in number following a power law distribution (Albert, et al., 2000). Similar properties have been observed in protein networks make certain nodes central when compared to the other (Jeong, et al., 2001).

1.2 DNA Microarrays

DNA microarray technology is used for simultaneous measurement of expression levels of thousands of genes. DNA microarrays are chips with arrayed short DNA sequences called “probes” onto which the sample containing cDNA called “target” are allowed to hybridize. Probe-target hybridization is detected by labelling with a fluorophore, silver or chemiluminiscence and is quantified by measuring the flouroscence intensities by a scanner. The fluorescence intensities correspond to abundances of gene expression.

High density arrays containing tens of thousands of synthetic oligonucleotides. They are synthesized *in situ* using a combination of photolithography and oligonucleotide chemistry. Such high density oligonucleotide arrays are called “GeneChip” provided by the company Affymetrix (http://www.affymetrix.com). Using this technology, mRNAs present at a frequency of 1:300,000 can be unambiguously detected. In a typical experiment where two conditions are compared, for instance normal and cancer condition, the samples along with the dye are hybridized to two different GeneChip arrays. Since a single dye is used for quantification the results collected represent absolute gene expression values. This provides the advantage of comparing the absolute gene expression values between different experiments done months or years apart.

1.2.1 Experimental design

One of the major considerations in the experimental design is the use of replicates. A single measurement could give rise to many false positives that pass the filter. However, the use of duplicates further reduce false positives by using the same filter for two measurements. The second consideration is the use of “spiked” control mRNAs in the samples to be analyzed. For Affymetrix GeneChips, kits of premade sets of control RNAs are available. Using such controls could improve the quality of the data (Ness, 2006).

Microarrays detect systematic problems in the analysis or preparation of samples called the “day effect”. This means samples analyzed on the same day correlate with each other more than those analyzed on different days. This can be solved by dividing the sample into groups, with each group having control and experimental samples. Then each group can be analyzed on different days (Ness, 2006).

1.2.2 Data Standardization

The Microarray data is generated from different platforms, assay protocols and analysis methods. This warrants the need for standardization of Microarray data to solve the problem of data interoperability. Several community efforts have offered standardization solutions to this problem.

One of them is MIAME (Minimum Information About a Microarray Experiment) that defines a check list of minimum information required to ensure that microarray data can be easily interpreted and that results derived from its analysis can be independently verified (Brazma, et al., 2001). This is being adopted by many journals as a requirement for the submission of papers incorporating microarray results. But MIAME does not describe the format for the information. The minimum information about a published microarray-based gene expression experiment includes a description of the following six sections (Brazma, et al., 2001):

1. Experimental design: the set of hybridization experiments as a whole needs tp be given.
2. Array design: each array used and each element (spot, feature) on the array needs to be described.
3. Samples: samples used, extract preparation and labelling
4. Hybridizations: procedures and parameters
5. Measurements: images, quantification and specifications
6. Normalization controls: types, values and specifications

Further, the MGED Society has developed standards for the representation of gene expression experiment results and relevant annotations.

1.2.3 Normalization and statistical analysis

The complexity of the Microarray data poses major statistical challenges including the data normalization. There are a number of reasons why data must be normalized, including unequal quantities of starting RNA, differences in labelling or detection efficiencies between the fluorescent dyes used, and systematic biases in the measured expression levels

(Quackenbush, 2002). Several normalization methods have been developed for analysis of microarray data (Huber, et al., 2003; Huber, et al., 2002; Park, et al., 2003).

In the testing for significance with gene expression data from DNA microarray experiments the multiple comparison problem needs to be addressed. This stems from comparison of hundreds or thousands of genes in a typical Microarray experiment. Even though the P-value indicates statistical significance of the differentially expressed genes, the very high number of genes in the array could represent as differentially expressed genes that are false positives. Statistical methods to assess statistical power based on the variation present in the data and the number of experimental replicates, and can help minimize type I and type II errors in the analyses are available (Wei, et al., 2004).

The major statistical hurdle in analyzing the microarray data is the dimensionality of data. A typical microarray study will obtain thousands of numbers per sample for many samples, perhaps a hundred samples.

1.2.4 Data sources

The comparative studies of several Microarray datasets could provide useful information on the underlying biology of disease conditions. This requires storage of Microarray data in specified formats such as MIAME in specialized datasets enabling intuitive querying features. Some of the databases which house such data are Gene Expression Omnibus (GEO) (Barrett, et al., 2007; Edgar, et al., 2002), Array Express (Parkinson, et al., 2009) and Oncomine (Rhodes, et al., 2004).

1.3 Network based analyses

Endogenous signal transduction in cancer cells is systematically disturbed to redirect the cellular decisions from differentiation and apoptosis to proliferation and, later, invasion (Vogelstein and Kinzler, 2004). Cancer cells acquire their malignancy through accumulation of advantageous gene mutations by which the necessary steps to malignancy are obtained (Hanahan and Weinberg, 2000). These selfish adaptations to independence can be described as a result from an evolutionary process of diversity and selection (Goymer, 2008). I was interested to observe these processes on a global view of cellular

signal transduction. Experimental high throughput methods such as gene expression profiling with microarrays enable investigating the pathogenic function of tumours on a mesoscopic level. Large-scale gene expression profiles were successfully used to predict clinical outcome (Fan, et al., 2006; van 't Veer, et al., 2002) and improved risk estimation (Oberthuer, et al., 2006). However these studies didn't relate genes and their expression to a functional context. To gain an understanding on a systems view, gene expression can be mapped onto cellular networks. Several studies have been reported that used gene expression data from microarrays to describe specific characteristics of signalling networks in cancer.

Discriminative components of a protein-protein interaction network were identified by comparing gene expression patterns of metastatic and non-metastatic tumours in breast cancer and suited as risk markers for breast cancer metastasis (Chuang, et al., 2007). The study used a protein-network based approach that used sub-networks as markers rather than individual genes obtained from protein-protein interaction databases. Chuang *et al*, analyzed the expression profiles of the two cohorts of breast cancer patients from a different study (van de Vijver, et al., 2002; Wang, et al., 2005). They restricted the analysis to the 8141 genes present in both data sets. For 78 patients in van de Vijver et al., 2002 and 106 in Wang et al., 2005, metastasis had been detected during follow-up visits within 5 years of surgery. Profiles for these patients were assigned to the class 'metastatic,' whereas profiles for the remaining 217 and 180 patients were labelled 'non-metastatic.' To obtain a corresponding human protein–protein interaction network, they assembled a pooled data set comprising 57,235 interactions from various sources like data integrated from yeast two-hybrid experiments, predicted interactions via orthology and co-citation, and scanning of the literature.

The expression and network data sets were integrated, by overlaying the expression values of each gene on its corresponding protein in the network and then were searched for sub-networks whose activities across the patients were highly discriminative of metastasis. Briefly, a candidate sub-network was first scored to assess its activity in each patient, defined by averaging its normalized gene expression values. This step yielded 295 and 286 activity scores per sub-network, corresponding to the number of breast cancer patients in the two data sets, respectively. Second, the discriminative potential of a candidate sub-network was computed based on the mutual information between its activity score and the metastatic/non-metastatic disease status over all patients. Significantly discriminative sub-

networks were identified by comparing their discriminative potentials to those of random networks.

The use of sub-networks as markers by Chuang et al., 2007 had the following advantages: First, the resulting sub-networks provided models of the molecular mechanisms underlying metastasis. Second, although genes with known breast cancer mutations were typically not detected through analysis of differential expression, they played a central role in the protein network by interconnecting many expression-responsive genes. Third, the identified sub-networks were significantly more reproducible between different breast cancer cohorts than individual marker genes selected without network information. Finally, a higher accuracy in prediction was obtained by network-based classification, as confirmed by selecting markers from one data set and applying them to a second independent validation data set.

Reverse-engineered gene networks were combined with expression profiles to identify the genetic mediators and mediating pathways associated with prostate cancer (Ergun, et al., 2007). They used an approach called mode-of-action by network identification (MNI), which has previously been validated as a means to identify the targets and associated pathways of compounds (di Bernardo, et al., 2005). The MNI algorithm operates in two phases, in phase one, a network model of regulatory interactions is reverse engineered with a diverse training set of whole-genome expression profiles. In phase two, the network is used as a filter to determine the genes affected by a condition of interest. The highest ranked mediator genes, ranked by a *Z*-statistic, are those whose expression was most inconsistent with the model, and this inconsistency is attributed to the external influence of the condition on those genes. Genes implicated in the advancement as well as suppression of a disease are equally likely to be identified as significant genetic mediators by the MNI algorithm. The MNI algorithm identified the AR gene among the top genetic mediators in the metastatic prostate cancer group but not in the non-recurrent primary prostate group. The 100 highest ranked genes in non-recurrent primary and metastatic prostate cancer groups were subjected to enrichment analysis for the AR signalling pathway. The list of the top 100 genes for the metastatic prostate cancer were significantly enriched with genes of the AR signalling pathway, in contrast to the non-recurrent primary prostate cancer, which was not enriched. These results, supported their hypotheses, that the AR gene and the AR pathway are mediators of prostate cancer progression and metastasis.

Ergun et al., 2007 further applied the MNI algorithm to nine recurrent primary prostate cancer samples. Consistent with their hypothesis, MNI ranked the AR gene 970, 155 and 9 for the non-recurrent primary, recurrent primary and metastatic prostate cancer groups, respectively. Thus, these findings suggests that the AR gene, in the context of the reverse-engineered network, can be used as a marker for detecting the aggressiveness of primary prostate cancers. Interestingly, expression change alone ranked the AR gene 641, 668 and 207 in the respective groups, indicating that expression change alone is incapable of capturing the differential involvement of the AR gene in recurrent and non-recurrent primary prostate cancers.

A novel method, to analyze gene expression data on the basis of interaction data, using a metabolic network of enzymes was developed (Konig, et al., 2006). The method was applied to *E. coli* under oxygen deprived conditions and physiologically relevant patterns that represent an adaptation of the cells to changing environmental conditions were extracted.

König et al., extracted the metabolic reactions from the EcoCyc database (Version 9, (Keseler, et al., 2005)). A graph was established by defining neighbours of metabolites in which, two metabolites were neighbours if and only if an enzymatic reaction existed that needed one of the metabolites as input (needed substrate) and produced the other as output (product). In this representation, enzymes were edges and metabolites the nodes. This network was clustered to group enzymes into parts of the network which were highly connected. The clustering algorithm produced a symmetrical sub-matrix of the cluster matrix for each cluster, whose rows and columns were the metabolites and its entries the connecting reactions. The normalized gene expression data of each data-set was mapped onto the corresponding reactions of the transcribed proteins. Mean values were taken if a reaction was catalysed by a complex of proteins. As a case study they analyzed gene expression data for 22 and 21 samples of *E. coli* under normal and oxygen depleted conditions respectively. The expression data of all samples was mapped onto each cluster-matrix, yielding 43 different patterns for each cluster. They calculated a value for every possible expression pattern of neighbouring genes and groups of genes within a cluster that may show substantial differences between samples of different conditions. Hence, they performed a Haar-wavelet transform for each cluster-matrix. The wavelet transformed

expression values served as features for a classifier. Using a step-wise feature extraction , the features were ranked according to their discriminative behaviour. This allowed the identification of regions with a varying pattern between aerobic and anaerobic conditions.

König et al., found a strong differential expression pattern of the transcripts coding for formic acid processing enzymes at the interface of the aerobic and anaerobic glucose catabolism. The aerobic catabolism processes pyruvate further on the respiratory glyoxylate cycle, whereas an anaerobic processing uses pyruvate formate lyase to produce formic acid as a fermentative product to be further degraded or excreted. It was shown that Pyruvate formate lyase may serve as a single switch. Their study also highlighted a concerted regulation reaction on oxygen deprivation. The bacteria adapted to this environmental change not only by degrading pyruvate into formate, but also by formate removal which was enhanced by up-regulated genes for formate exocytosis and formate degradation. They revealed an adapted regulation for aspartate processing enzymes. Interestingly, coming from aspartate, the starting point for purine biosynthesis was up-regulated, whereas that of pyrimidine biosynthesis was not.

In another study, an image processing technique was used for network analysis where the response of the hetero-fermentative bacterium to oxygen deprivation was investigated (Schramm, et al., 2007). This time, feature modules were then generated using the Haar wavelet transformations and significant modules extracted by statistical testing methods. Then the significant reaction pairs were clustered and found clusters were analyzed. Finally, the results were merged to obtain an overall map of metabolic changes under oxygen deprivation in *E. coli.* Schramm et al., detected, as expected, an up-regulation in the pathways of hexose nutrients up-take and metabolism and formate fermentation. Furthermore, their approach revealed a down-regulation in iron processing as well as the up-regulation of the histidine biosynthesis pathway. The latter reflects an adaptive response of *E. coli* against an increasingly acidic environment due to the excretion of acidic products during anaerobic growth in a batch culture.

A systematic graph theory-based analysis of a yeast protein-protein interaction network was performed to construct computational models for describing and predicting the properties of lethal mutations and proteins participating in genetic interactions, functional groups, protein complexes and signalling pathways (Przulj, et al., 2004). Przulj and co-workers showed that, proteins participating in genetic interactions in the graph appeared to

have a degree closer to that of viable proteins. Interestingly, lethal mutations are not only often at highly connected nodes within the network (called hubs), but are at nodes whose removal causes a disruption of the network structure (called articulation points). They also suggested the existence of alternate paths that bypass viable nodes in protein-protein interaction networks, offering an explanation why null mutations of these proteins are not lethal. They hypothesized that highly connected sub-graphs or 'clusters' within a protein-protein interaction network could indicate protein complexes. They defined a highly connected sub-graph as a graph, in which the minimum number of edges whose removal disconnects the graph is greater than 2. They analyzed protein-protein interaction graphs of different sizes to determine the relationship between the size of a graph and the number and complexity of identified clusters, which are feasible candidates for protein complexes. Most of the clusters overlapped with the MIPS database complexes.

Przulj et al., also sought to determine if known signalling pathways had a characteristic structure within the network. The MAPK signalling pathway is a prototypical pathway that exhibits linearity in structure and was used for a linear pathway model. There were 31 MAPK pathway proteins in the full protein-protein graph comprising of 78,390 interactions: four of them were starting points (sources), eight were ending points (sinks) and the rest were internal proteins. They constructed a conservative predictive model that considers sources and sinks with a degree of at most 4 and intermediate nodes of degree of at least 8. From the predicted pathways, they showed that articulation points on linear pathways are much more likely to be lethal mutations or to participate in genetic interactions.

A comprehensive analysis of a manually curated human signalling network was performed by Cui and co-workers (Cui, et al., 2007). The signalling network analyzed was composed of 1634 nodes and 5089 edges. To integrate mutated and methylated genes onto the network, they first collected the cancer mutated genes from the database Catalogue Of Somatic Mutations In Cancer (COSMIC) database, which collects the cancer mutated genes through literature curation and large-scale sequencing of tumour samples in the CGP. This data was then combined with the cancer mutated genes derived from other genome-wide and high-throughput sequencing of tumour samples. The cancer-associated methylated genes were taken from the genome-wide identification of the DNA methylated genes in cancer stem cells. Finally, 227 cancer mutated genes and 93 DNA methylated

genes were mapped onto the network. Among the 227 cancer mutated genes, 218 (96%) and 55 (24%) genes were derived from large-scale gene sequencing of tumours and literature curation, respectively. After mapping of cancer mutated genes onto the network, they found that cancer mutations occured most likely in signalling proteins that were acting as signalling hubs (i.e., RAS) actively sending or receiving signals rather than in nodes that were simply involved in passive physical interactions with other proteins. Alterations of these nodes, or signalling hubs, were predicted to affect more signalling events, resulting in cancer or other diseases. They also showed that cancer mutated genes and methylation-silenced genes have different regulatory mechanisms in oncogene signalling.

Cui et al., also showed that the oncogene-signalling event triggered by mutations is preferentially associated with activating downstream signalling paths or conduits and were less likely to be associated with downstream inhibitory signalling paths. They found that the cancer mutated genes were enriched in positive signalling regulatory loops, whereas the cancer-associated methylated genes were enriched in negative signalling regulatory loops. They further characterized an overall picture of the cancer-signalling architectural and functional organization by constructing an oncogene-signalling map, containing 326 nodes, 892 links and the interconnections of mutated and methylated genes. From this map, they suggested that the crucial players of oncogene signalling tend to be closely clustered and regionalized. This map also uncovered the architectural structure of the basic oncogene signalling and highlights the signalling events that are highly conserved in generating tumour phenotypes. The map could be decomposed into 12 topological regions or oncogene-signalling blocks, including a few 'oncogene-signalling-dependent blocks' in which frequently used oncogene-signalling events were enriched. One such block, in which the genes were highly mutated and methylated, appeared in most tumours and thus played a central role in cancer signalling. Functional collaborations between two oncogene-signalling-dependent blocks occur in most tumours, although breast and lung tumours exhibited more complex collaborative patterns between multiple blocks than other cancer types. Benchmarking two data sets derived from systematic screening of mutations in tumours further reinforced their findings that, although the mutations were tremendously diverse and complex at the gene level, clear patterns of oncogene-signalling collaborations emerge recurrently at the network level. Finally, the mutated genes in the network could be used to discover novel cancer-associated genes and biomarkers.

The global and local regulation of gene expression in the metabolic network of *Saccharomyces cerevisiae* was investigated (Kharchenko, et al., 2005). Kharchenko et al., represented metabolism as a graphical model, with nodes of the graph corresponding to genes encoding metabolic enzymes, and edges to metabolic connections between corresponding enzymes. They investigated how positive and negative correlation of mRNA expression profiles depends on the metabolic network distance, and determined the maximum distance at which genes display statistically significant co-expression. They showed that regulation of metabolic genes was local and extends, generally, to distances smaller than the mean network distance. Such regulation implies that genes close in the metabolic network were usually co-expressed together, possibly to optimize local metabolic fluxes. Positive co-expression was strongest among adjacent genes and decreases monotonically with network distance. In contrast, negative co-expression was most prominent at intermediate distances.

Kharchenko et al., also suggest that regulation of the metabolic network established a number of local, positively co-expressed regions that may exhibit some degree of negative co-expression between each other. Furthermore, they found that positive co-expression and functional associations were strongest in the linear parts of metabolism, while negative co-expression was more pronounced in highly branched regions. Their analysis of the elementary topological motifs showed that co-expression in divergent branches was significantly stronger than that observed in convergent branches. This pattern showed an emphasis on co-regulation of biomass synthesis or degradation from common metabolic precursors. They observed an agreement between the mRNA co-expression and genome context associations suggesting that the observed patterns of metabolic regulation was reflected in genome evolution and affected the location of genes on the chromosomes.

Principles of transcriptional control in the metabolic network of *Saccharomyces cerevisiae* were studied (Ihmels, et al., 2004). By integrating large-scale expression data with the structural description of the metabolic network, they systematically characterized the transcriptional regulation of metabolic pathways. Their analysis revealed recurrent patterns, which may represent design principles of metabolic gene regulation. The three major findings of the study were: First, they showed that transcription regulation biased metabolic flow towards linearity by co-expressing only distinct branches at metabolic branch points. The co-expression pattern of enzymes participating at such branch points,

suggested that many possible branches were in fact suppressed in the actual context-dependent map. This suppression reduced metabolite dissipation and ensured a more efficient metabolic flow. Secondly, they showed that individual isozymes were often separately co-regulated with distinct processes, providing a means of reducing crosstalk between pathways using a common reaction. Finally, they suggested that transcriptional regulation defined a hierarchical organization of metabolic pathways into groups of varying expression coherence. In conclusion, they propose that transcription regulation is prominently involved in shaping the metabolic network of *S. cerevisiae* in response to changing conditions.

The dynamics of a biological network on a genomic scale was studied, by integrating transcriptional regulatory information and gene-expression data for multiple conditions in *Saccharomyces cerevisiae* (Luscombe, et al., 2004). Initially they assembled a static representation of known regulatory interactions from the results of genetic, biochemical and ChIP (chromatin immunoprecipitation)–chip experiments. They obtained a complex network consisting of 7,074 regulatory interactions between 142 transcription factors and 3,420 target genes. To get a dynamic perspective, they then integrated gene-expression data for the following five conditions: cell cycle, sporulation, diauxic shift, DNA damage and stress response. From these data, they traced paths in the regulatory network that were active in each condition using a trace-back algorithm.

They showed that the topological measures changed considerably between the endogenous and exogenous sub-networks. In biological terms, the small in-degrees for target genes in exogenous conditions indicated that transcription factors were regulating in simpler combinations, and the large out-degrees showed that each transcription factor had greater regulatory influence by targeting more genes simultaneously. Short paths implied faster propagation of the regulatory signals which was needed for exogenic perturbations. Conversely, long paths in multi-stage, endogenous conditions suggested slower action arising from the formation of regulatory chains to control intermediate phases. Finally, high clustering coefficients in endogenous conditions signified larger inter-regulation between transcription factors. In summary, sub-networks have evolved to produce rapid, large-scale responses in exogenous states, and carefully coordinated processes in endogenous conditions. They showed that, in response to diverse stimuli, transcription factors alter their interactions to varying degrees, thereby rewiring the network. A few

transcription factors serve as permanent hubs, but most act transiently only during certain conditions.

Yeger-Lotem and co-workers identified characteristic network patterns consisting of both transcription–regulation and protein–protein interactions that recur significantly more often than in random networks were identified (Yeger-Lotem, et al., 2004). They developed algorithms for detecting motifs in networks with two or more types of interactions and applied them to an integrated data set of protein–protein interactions and transcription regulation in *Saccharomyces cerevisiae*. They found a two-protein mixed-feedback loop motif, five types of three-protein motifs exhibiting co-regulation and complex formation, and many motifs involving four proteins. Virtually all four-protein motifs consisted of combinations of smaller motifs. Their study presents a basic framework for detecting the building blocks of networks with multiple types of interactions.

The figure 3 depicts the four-protein motifs discovered by Yeger-Lotem, et al., 2004. In the figure (*a*) represents motifs that can be represented as combinations of three-protein network motifs. When there is more than one possible way to generate a four-protein motif, the combination involving the more abundant three-protein motifs is presented. In the figure (*b*) represented motifs that cannot be constructed from three-protein motifs. *i*, the bi-fan motif; *ii*, a motif containing a feed-forward loop; *iii–vi*, motifs that appear as extensions of smaller network motifs, for which one of the protein-protein interactions in each smaller motif (*Left*) was extended to a series of protein-protein interactions by means of an intermediate protein (*Right*). A node represents a gene and its protein product; a red, directed edge represents a transcription-regulation interaction; and a black, bi-directed edge represents a protein-protein interaction.

An important family of motifs in cellular signalling is the feed-forward loop (FFL) (Alon, 2007). This motif consists of three genes: a regulator, X, which regulates Y, and gene Z, which is regulated by both X and Y. Because each of the three regulatory interactions in the FFL can be either activation or repression, there are eight possible structural types of FFL (figure 4a). In FFLs, X and Y are integrated to regulate the Z promoter. Two common 'input functions' are an 'AND gate', in which both X and Y are needed to activate Z, and an 'OR gate', in which binding of either regulator is sufficient. These feed forward loops can be coherent, that is inputs to Z are not contradictory or in-coherent. In-coherent motifs are

needed for short activations. Other input functions are possible, such as the additive input function and the hybrid of AND and OR logic. However, much of the essential behaviour of FFLs can be understood by focusing on the stereotypical AND and OR gates. Each of the eight FFL types can thus appear with at least two input functions. In the best studied transcriptional networks (*E. coli* and yeast), two of the eight FFL types occur much more frequently than the other six types. These common types are the coherent type-1 FFL (C1-FFL) and the incoherent type-1 FFL (I1-FFL) (figure 4b and 4c respectively).

Figure 3. Four-protein network motifs discovered in the stringent network identified by Yeger-Lotem, et al., 2004.

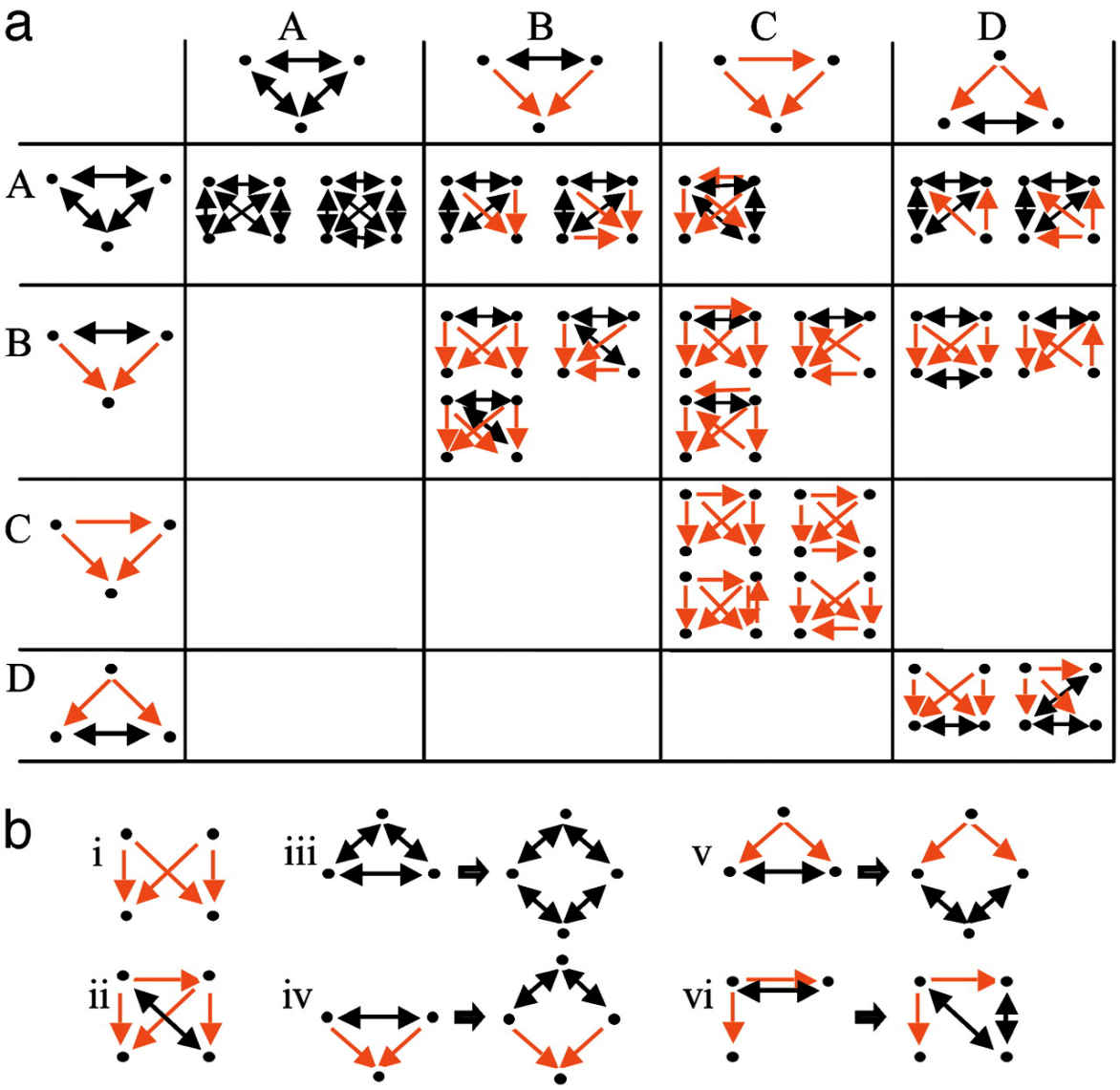
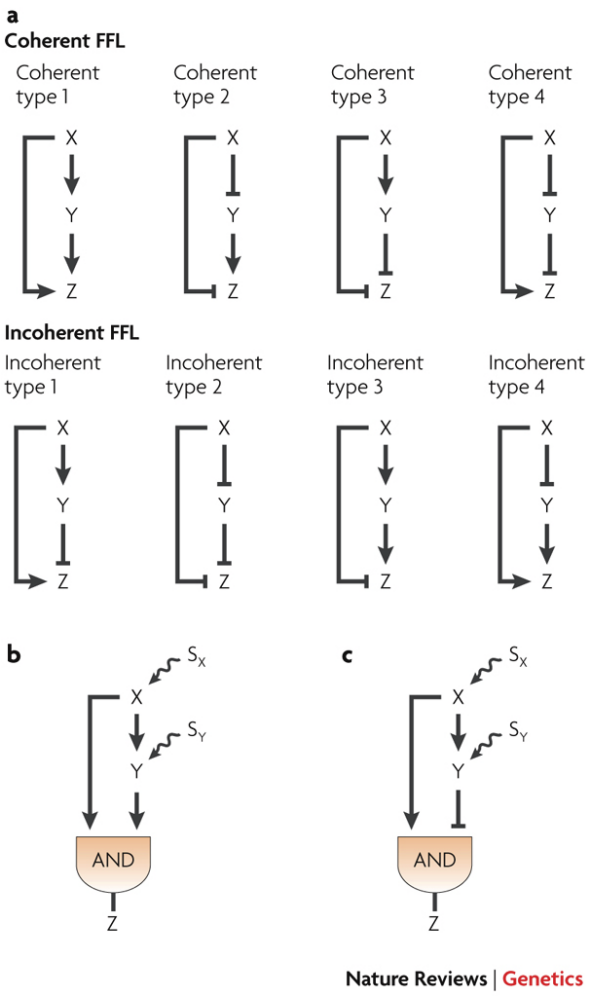


Figure 4. Feed forward loops. The figure shows different types of feed forward loops in literature (Alon, 2007).



1.4 Biological background

Cancer is a disease of uncontrolled cell proliferation. This aberrant behaviour of the cell is a result of the cumulative effects of the signalling circuitry. The cells express around 20,000 or more distinct proteins which are involved in many regulatory circuits. These proteins talk to each other and hence participate in signalling cascades where the signal from the receptor is transmitted down to the transcription factor. These signalling circuits determine the transformation from a normal cell to a malignant form. Hence, cancer can be

regarded as a disease of aberrant signal processing. In this section some of the major cancer signalling pathways are explained.

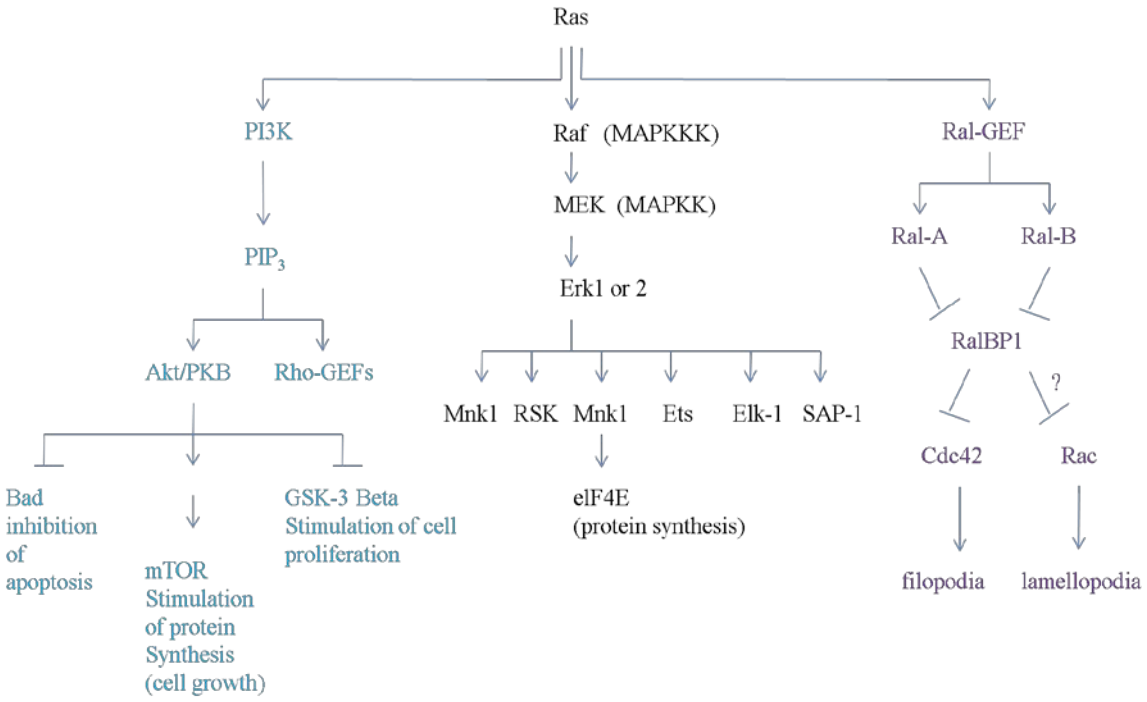
A variety of biochemical signals radiate from ligand-activated growth factor receptors and some of them are directed directly to the nucleus where they alter gene expression programs, and other have cytoplasmic targets. Many of the traits of cancer can be traced back to the effects evoked growth factors. ***Ras*** is a family of genes encoding small GTPases that are involved in cellular signal transduction. The activation of Ras signalling leads to cell growth, differentiation and survival. Since Ras communicates signals from outside the cell to the nucleus, mutations in *ras* genes can lead to permanent activation resulting in aberrant signalling. Since these signals result in cell growth and division, inappropriate Ras signalling can ultimately lead to oncogenesis and cancer. A common signalling channel can be observed across phyla leading from receptor to Ras. This involves,

Tyrosine kinase receptor ⟶ Shc ⟶ Grb ⟶ Sos ⟶ Ras

The Ras protein stands in the middle of a complex signalling cascade and many signalling cascades emerge downstream of Ras evoking a number of distinct changes in the cell.

There are three signalling cascades that operate downstream of Ras. One of them is **MAPK (Mitogen-activated protein kinase) pathway**. The MAPK pathway is illustrated in blue color in Figure 5.The GTP bound Ras binds to several downstream signalling partners known as Ras effectors. Raf kinase is a Ras effector which interacts with GTP bound Ras. Raf is now activated and phosphorylates a second kinase known as MEK (MAPKK) thereby activating it. The MEK inturn phosphorylates two other kinases , the extracellular signal-regulated kinases 1 and 2, also called Erk1 and Erk2. The Erks phosphorylate transcription factors (Ets, Elk-1, SAP-1) and in addition phosphorylate other kinases which activate yet other transcription factors. The MAPK pathway contributes to several Ras induced phenotypes in the cancer cell. This pathway activates several growth-promoting genes, confers anchorage independence and loss of contact inhibition and also contributes to change in cell shape which has been associated with the transformation of the oncogene *ras* (Weinberg, 2007).

Figure 5. Pathways downstream of Ras



The second pathway downstream of Ras is the **PI3 kinase (Phosphatidylinositol 3-kinase or PI3K) pathway**. It is illustrated in black color in Figure 5. The PI3K attaches a phosphate group to 3' hydroxyl of the inositol moiety of $PI(4,5)P_2$ (also called PIP_2) converting it to Phosphatidylinositol (3',4',5')-triphosphate (PIP_3). Once PIP_3 is formed by PI3K, a serine-threonine kinase known as Akt, also called protein kinase B (PKB) can become tethered via its PH domain to the inositol head group of PIP_3. This association activates Akt/PKB, which in turn phosphorylates several protein substrates that have multiple effects on the cell. The three major effects of Akt/PKB on the cell are:

1. It reduces the possibility of activation of the apoptotic program and aids cell survival.
2. It stimulates cell proliferation.
3. It stimulates cell growth.

Independent of theses proliferative functions Akt/PKB also control the rate protein synthesis in the cell. Akt/PKB phosphorylates and inactivates a protein called TSC2, which otherwise triggers the inactivation of mTOR kinase which regulates the rate of translation.

A family of guanine nucleotide exchange factors (GEFs) called Rho-GEFs use theier PH domains to associate with PIP_3. These include Rho proper and its two cousins, Rac and Cdc42. Once activated the Rho proteins participate in reconfiguring the structure of the cytoskeleton and the cell's attachments to its physical surroundings. Thereby they control cell shape, cell motility and in case of cancer cells invasiveness. For example, Cdc42 is involved in skeletal reorganization and controlling **filopodia**, small finger-like extensions used by the cell to explore its surroundings; while Rac is involved in the formation of **lamellopodia**, broad ruffles extending from the plasma membrane which are found at the leading edges of motile cells (Weinberg, 2007).

The PI3K pathway is deregulated in a number of human cancer types. The low levels of PIP_3 is maintained by phosphatases that inactivate PIP_3. One of the phosphatases is PTEN which removes the 3'phosphate group from PIP_3. Hence hyperactivity of PI3K or inactivity of PTEN can deregulate the pathway. One form of PI3K is over expressed and active in certain ovarian carcinomas. In lymphomas, head and neck tumours and colon carcinomas the Akt/PKB is over expressed and hyperactivated. In tumour types such as breast, prostate and glioblastoma the PTEN activity is lost due to mutation or methylation events that suppress PTEN gene expression. Such loss of PTEN activity is found in 30-40% of all human cancers (Weinberg, 2007).

The third major pathway downstream of Ras involves Ras-like proteins termed Ral-A and Ral-B. It is illustrated in pink color in Figure 5. The communication between Ras and Ral is carried out by Ral guanine nucleotide exchange factors (Ral-GEFs). The Ral-GEFs causes a Ral protein to shed GDP and bind GTP. The activated Ral-A and Ral-B can further inactivate Rac and Cdc42. Ral proteins are considered to play roles in cell motility enabling invasion and metastasis of cancer cells. In conclusion, by activating multiple pathways simultaneously Ras brings about several phenotypic changes observed during neoplastic transformation (Weinberg, 2007).

The **Jak-STAT pathway** is also responsible for transformation in different cancer types. The Jak enzyme (Janus kinase enzyme) on binding to the receptor causes receptor dimerization and phosphorylation of tyrosine residues in the cytoplasmic tail of the receptor. These phosphotyrosine are bound by STATs (signal transducers and activators of transcription) and are phosphorylated. This activates STAT forms STAT-STAT dimers which are translocated to nucleus to function as transcription factors. STATs activates target genes which are involved in cell proliferation and cell survival like *myc*, cyclins D2

and D3 and genes encoding anti-apoptotic protein Bcl-X_L. There is growing evidence that connects STATs to cancer pathogenesis. For example, Stat3 is constitutively activated in a number of human cancers like melanoma and breast cancers showing that they act as important mediators of transformation (Weinberg, 2007).

The **Wnt-β-catenin pathway** enables cells to remain in a undifferentiated state which is typical of many cancer cells. The Wnt factors act through Frizzled receptors and suppress the activity of glycogen synthase kinase-3β (GSK-3β). The GSK-3β are active in the absence of Wnts and phosphorylate many protein substrates thereby marking them for degradation. One of the most important substrates is β-catenin, they are either bound to the cytoplasmic domain of cell-cell adhesion receptors (eg. E-cadherin) or operate in the nucleus as a vital component of a transcription factor. When Wnt pathway is activated, the activity of GSK-3β is suppressed leading increased concentrations of β-catenin. Many of the β-catenin molecules move into the nucleus and activate transcription by binding to Tcf/Lef proteins. This transcription factor complex activate expression of target genes involved in cell growth and proliferation like *myc* and cyclin D1. GSK-3β can also phosphorylate cyclin D1 marking it for degradation. Hence Wnt pathway also modulates cyclin D1 expression both at transcriptional and post-translational levels. In many human breast cancers Wnt expression is increased 4-10 fold and there is evidence of nuclear translocation of β-catenin in approximately 20% of advanced prostate carcinomas. β-catenin mutation which eludes from GSK-3β phosphorylation have been observed in have been observed in carcinomas of prostate liver, colon, endometrium, ovary and melanomas (Weinberg, 2007).

The **Nuclear factor-κB** (NF- κB) pathway is a important pathway implicated in cancer. NF- κB is commonly found as heterodimer composed of p65 and p50 subunit. In the cytoplasm, it is sequestered by a third polypeptide IκB (inhibitor of NF- κB) showing suppressed activity in this state. Signals from diverse sources phsophorylate IκB, marking it for degradation. This liberates NF- κB which migrates to the nucleus and activates the expression of over 150 target genes. NF- κBs have effects on cell survival and proliferation in cancer. In the nucleus, they can induce the expression of key anti-apoptotic proteins Bcl-2 and IAP-1 and -2. They also induce expression of *myc* and cyclin D1 genes involved in cell growth and proliferation. Thus NF- κB can protect the cell from apoptosis and simultaneously drive their proliferation. In human cancers, NF- κB was found to be

constitutively activated. It is highly activated in breast cancers and also plays a role in malignancies of lymphocyte lineages (Weinberg, 2007).

The **Notch pathway** is controlled by the Notch protein which is a transmembrane protein and four different forms of Notch are expressed in mammalian cells. Notch binds its ligand (NotchL) and undergoes proteolytic cleavage, liberating a cytoplasmic fragment which migrates into the nucleus and along with other proteins acts as a transcription factor. It was seen that altered forms of Notch contribute to cancer pathogenesis. Overexpression of one of the forms of Notch was seen in a majority of cervical carcinomas, a subset of colon carcinomas and in lung squamous carcinomas. Increased expression of Notch ligands, Jagged and Delta was found in cervical and prostate carcinomas (Weinberg, 2007).

The **Hedgehog-patched pathway** involves the binding of patched receptor by its ligand Hedgehog causing it to release the Smoothened protein from inhibition, which later emits downstream signals. The Smoothened protein prevents the cleavage of cytoplasmic Gli protein. If cleaved in the absence of Smoothened protein, one of the fragments moves into the nucleus to act as a transcriptional repressor. In the presence of Smoothened the intact Gli proteins migrate into the nucleus and act as transcriptional activators. About 40% of sporadic basal cell carcinomas of skin carry mutant PTCH (human patched) or SMO (Smoothened) alleles. Somatically mutated alleles of PTCH were also found in medulloblastomas, meningiomas, breast and esophageal carcinomas (Weinberg, 2007).

The **TGF-β pathway** plays a role in pathogenesis of many of the carcinomas, acting in the early stage by arresting the growth of many cell types and later stages leading to cancer progression by contributing to the phenotype of tumour invasiveness. On binding its ligand the TGF-β receptor phosphorylate Smad2 (or Smad3) protein molecules which then binds to Smad4 proteins and the resulting heterodimeric protein complex migrates to the nucleus functioning as a transcription factor expressing a large constituency of genes. In the absence of Smads the epithelial cancer cells can escape from the growth inhibitory signals of TGF-β and thrive. This state was observed in precursors of invasive pancreatic carcinomas (Weinberg, 2007).

Some of the major signalling pathways that are deregulated in cancers have been briefly explained above. As it is evident from the central role of Ras, the signalling cascades are not linear pathways but a network of signalling interactions that consist of a complex signalling circuitry. In this thesis, a method has been developed to understand this complex

circuitry using information from gene expression. The technical background for the study are elaborated in the subsequent sections.

Chapter 2

Methods

2.1 Different cancer types analyzed

The method developed in this thesis is used to analyze a broad range of cancer types. These included:

- Lung adenocarcinoma: Lung cancer are tumours arising from the cells lining the airways of the respiratory system. Lung adenocarcinoma is one of major types of lung cancer, it arises from the secretory (glandular) cells located in the epithelium lining the bronchi.
- Breast cancer: This cancer that starts in the breast, usually in the inner lining of the milk ducts or lobules. There are different types of breast cancer, with different stages (spread), aggressiveness, and genetic makeup.
- Prostate cancer: This form of cancer develops in the prostate, a gland in the male reproductive system. The cancer cells may metastasize (spread) from the prostate to other parts of the body, particularly the bones and lymph nodes.
- Head and neck squamous carcinoma: Head and neck cancer includes the squamous cell carcinomas of the oral cavity, pharynx and larynx.
- Oral tongue cancer: There are two parts of the tongue, the oral tongue and the base of the tongue. The cancer can develop in either part. The Oral tongue cancer involves cancer in the front two-thirds region of the tongue.
- Acute myeloid leukemia: It is a cancer of the myeloid line of blood cells, characterized by the rapid growth of abnormal white blood cells that accumulate in the bone marrow and interfere with the production of normal blood cells.
- Renal cancer: This is the cancer of the kidneys where it arises in the lining of very small tubes in kidneys that filter blood and remove waste products.
- Vulvar interstitial neoplasia: This form of cancer arises in the Vulva which is the external female genital organ including the clitoris, vaginal lips and the opening to the vagina

- Cervical cancer: It is a cancer that forms in the tissues of the cervix (the organ connecting the uterus and vagina).

2.2 Datasets

2.2.1 Gene expression datasets

The gene expression datasets were used either from collaborators or were downloaded from the NCBI GEO (gene expression omnibus) database (Barrett, et al., 2007; Edgar, et al., 2002). We used the neuroblastoma gene expression dataset which was measured by the Agilent array platform (Oberthuer, et al., 2006). The study analyzed tumour samples from 251 neuroblastoma patients across different disease stages. The stage 1 and stage 4 MYCN amplified classes of neuroblastoma were used to the comparative analysis. Two lung cancer datasets, one containing 17 normal patients and 12 patients with highly aggressive adenocarcinoma belonging to the cluster C2 for analysis. C2 is a cluster of patients with highly aggressive adenocarcinoma derived from the clustering of patients in the original study based on gene expression data (Bhattacharjee, et al., 2001). A second lung cancer dataset containing gene expression data of 27 normal and adenocarcinoma patients (Su, et al., 2007). Breast cancer datasets of 43 normal and cancer patients each were used (NCBI GEO: GSE15852). Prostate cancer datasets with 50 and 52 normal and cancer patients respectively were used (Singh, et al., 2002). Data from a study of head and neck squamous carcinoma consisting of 22 normal and cancer patients each were used in the analysis (Kuriakose, et al., 2004). Oral tongue cancer gene expression profiles of 26 and 31 normal and cancer patients respectively (Estilo, et al., 2009) were also used. AML (acute myeloid leukemia) datasets containing 18 and 25 (Stirewalt, et al., 2008), renal cancer datasets with 23 and 69 (Jones, et al., 2005), vulvar interstitial neoplasia with 10 and 9 (Santegoets, et al., 2007) and Cervical cancers dataset with 8 and 19 cervical cancer (Pyeon, et al., 2007) normal and cancer samples were used respectively. Except for the neuroblastoma dataset the rest of the analyzed datasets were from the Affymetrix platform.

2.2.2 Protein interaction dataset

In order to perform a network-based analysis of various cancer gene expression datasets a protein interaction network was assembled from the Human Protein Reference Database (HPRD)(Mishra, et al., 2006; Peri, et al., 2003). The protein-protein interaction dataset contained around 36000 protein-protein interactions. Depending on the gene to probe mapping, the largest connected component of the network was used for the analysis. The dataset broadly describes interaction relationships between proteins and captures signalling events like phosphorylation and binding reactions which form a major part of the signalling network.

2.3 Network reconstruction and analysis

The protein-protein interaction network was reconstructed from the HPRD database. Initially the probe ids of the gene expression datasets were mapped to corresponding NCBI Entrez Gene IDs using either a BLAST analysis (Altschul, et al., 1990) or the BioMart package in R. The BLAST analysis was used to map Agilent probe IDs to Entrez Gene IDs, where all the probe ids were aligned to human nucleotide sequences and alignments with up to 2 mismatches were taken. For probe ids from Affymetrix existing R functions were used to map to Entrez Gene IDs to the probes.

From the available Entrez Gene ID mappings a network was reconstructed. This number of mapped Entrez Gene IDs changed depending on the different platforms and platform formats of the gene expression data. Finally the gene expression data was mapped onto the protein-protein interaction network. This was done by calculating the Pearson correlation coefficients of the gene expression data of two interacting proteins and assigning them as edge weights. The Pearson's correlation coefficient r is given by,

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2)}{N}\right)\left(\sum Y^2 - \frac{(\sum Y)^2)}{N}\right)}} \quad (2)$$

Where X and Y are variables and N is the number of samples in X and Y.

To equally handle induction and inhibition events, absolute values of all correlation coefficients were used. For each cancer dataset two networks were compared. A network was constructed for normal samples and for cancer samples respectively. For Neuroblastoma “Stage1” samples and “Stage 4 MYCN amplified” samples were compared.

Pathways were then predicted from different pairs of receptors and transcription factors in the network. Genes with the molecular function term "receptor activity" from the definitions of Gene Ontology (www.geneontology.org) were used as receptors in the network. Genes listed as a transcription factor in the TRANSFAC database were used as transcription factors (Matys, et al., 2003). Dijkstra algorithm which is already explained in the introductory section was used for calculating the shortest paths for every pair of receptors and transcription factors in all the networks. The correlation values were subtracted from one before using the Dijkstra's algorithm to obtain paths with high correlation. The receptor-transcription factor pairs ranged from 21,353 to 72,124 (these are also the number of predicted pathways for each dataset) depending on the dataset platform formats used. Several properties of these predicted pathways were later compared between normal and cancer samples. Table 1 gives the network statistics showing the number of edges, nodes and receptor- transcription factor pairs.

Table 1. Network Statistics

	Edges	Nodes	Percentage used nodes in normal	Percentage used edges in normal	Percentage used nodes in tumour	Percentage used edges in tumour	Number of receptors	Number of transcription factors
Lung1	17529	4692	30.64	28.00	30.64	30.00	210	203
Lung2	27511	7092	26.38	24.00	27.72	27.00	292	247
Breast	27511	7092	28.52	28.00	31.11	31.00	292	247
Prostate	17529	4692	23.74	19.00	25.40	22.00	210	203
HN1	17529	4692	26.30	22.00	26.38	23.00	210	203
OT	17529	4692	27.11	23.87	30.29	30.67	210	203
AML	27511	7092	24.22	20.95	31.07	33.55	292	247
Renal	27511	7092	26.64	26.54	28.85	28.72	292	247
Vulva	12490	4671	23.12	21.55	25.78	25.49	175	148
Cervical	12490	4671	20.57	16.53	24.64	22.21	175	148

2.4 Defining the network features

Path length, link and node frequency, and the signalling motif are explained in the results. The (average) network diameter has been used as a measure for error tolerance of a network against removals of nodes in scale free networks (Albert, et al., 2000). The network diameters for the networks were obtained by the average of the shortest paths of each pair of nodes in the network. The network diameter was calculated for undisturbed (whole) networks and networks in which the top 20% of the hubs were removed. The ratio of these values were calculated to yield the increase of the average network diameter after hub removal. The calculation of the information content based on the assumption that signals enter the network at any receptor with equal probability within a certain time interval. These signals are passed by the links of the network to the transcription factors via the defined pathways from the receptors, again with equal probability. It was assumed that the signals vanish from the signalling network after having entered the corresponding transcription factor at the end of the path. Signals enter the receptors with a certain frequency, resulting in an equal distribution and therefore uniform density of the signals in each pathway. The probability of a signal to pass through the link of node i and j is then proportional to the number of pathways passing through this link. With this, we calculated the information content by Shannon's definition (Shannon, 1948)

$$I = \sum_{i=1}^{n} p_i \ log_2(p_i) \qquad (3)$$

in which n denotes the number of links and p_i the probability of a signal to be passed through link i. The clustering coefficient for node i was given by

$$C_i = \frac{n_{link}}{k(k-1)} \qquad (4)$$

in which n_{link} is the number of links connecting the neighbours of node i. k is the number of neighbours. This feature described how good the neighbours were connected within each

other. If they were fully connected, the clustering coefficient was one, if they were not connected at all, the clustering coefficient was zero.

2.5 Combined linear model for link frequency distributions

For each tumour and normal sample, the link frequency distributions were plotted on a log-log scale (basis = 10) using hist() (breaks = 10) of the package R (www.r-project.org). The combined linear models for these logarithmic distributions were calculated assuming same slopes with different intercepts for normal tissue and tumour. First, a linear regression was performed for each dataset yielding regression coefficients $\beta_{0,normal,i}$ and $\beta_{1,normal,i}$ of normal sample i for the intercept and the slope, respectively, and $\beta_{0,tumor,i}$ and $\beta_{1,tumor,i}$ for tumour sample i, i ∈ {AML, breast, cervical, head-and-neck, lung-1, lung-2, oral-tongue, prostate, renal, vulva}. Then combined linear models were obtained by calculating combined regression coefficients $\gamma_{0,normal,i}$, $\gamma_{1,normal,i}$, $\gamma_{0,tumor,i}$, $\gamma_{1,tumor,i}$ for intercept-normal, slope-normal, intercept-tumour and slope-tumour, respectively, i.e.

$$slope_{m,i} := \frac{\beta_{1,normal,i}+\beta_{1,tumor,i}}{2} \quad (5)$$

$$\gamma_{0,normal,i} = \overline{y_{normal,i}} - slope_{m,i} \cdot \overline{x_{normal,i}} \quad (6)$$

$$\gamma_{0,tumor,i} = \overline{y_{tumor,i}} - slope_{m,i} \cdot \overline{x_{tumor,i}} \quad (7)$$

$$\gamma_{1,normal,i} = slope_{m,i} \quad (8)$$

$$\gamma_{1,tumor,i} = slope_{m,i} \quad (9)$$

in which $\overline{x_{normal,i}}, \overline{x_{tumor,i}}, \overline{y_{normal,i}}, \overline{y_{tumor,i}}$ are the mean values of the logarithms of the breaks and corresponding frequencies of the distributions of normal and tumour sample i, respectively. With this mean slopes of the distributions for normal and cancer sample i was obtained, but distinct intercepts for them which well fitted the data (Figure 8 and Figure 12 (Supplement)).

2.6 Defining and counting the integration and the maintenance motif

Hubs of cancer mutated genes were defined by intersecting the top 20 most frequently involved nodes and the list of cancer genes from Cui and co-workers (Cui, et al., 2007). This was done for normal and tumour of every sample i, i ∈ {AML, breast, cervical, head-and-neck, lung-1, lung-2, oral-tongue, prostate, renal, vulva}. For each datasets, all triangles were collected in which at least one node was such a cancer mutated hub. Out of these triangles, triangles having the motifs for integration (motif A in Figure 9) and maintenance (motif B in Figure 9) were selected. For motif A, triangles were selected where distances between all pairs of nodes (hub-n_1, hub-n_2, n_1-n_2, n_1 and n_2 are the two other nodes in the triangle) were equal or below the medians $m_{normal,i}$, $m_{tumor,i}$ of all distances of the datasets normal and tumour for sample i, respectively. For motif B, the triangles were selected in which hub-n_1 and hub-n_2 were larger or equal *and* n_1-n_2 were lower or equal than $m_{normal,i}$ and $m_{tumor,i}$ for normal and tumour sample i, respectively.

2.7 Identification of high node frequency genes

While tracking the shortest paths from the receptor nodes to the transcription factor nodes in the network, certain nodes may be used more frequently when compared to other nodes. The node frequency is simply the number of times a node was used for every pair of and transcription factor. The non-aggressive “stage 1” condition was compared to the aggressive “Stage 4 MYCN amplified”(stage4A) tumour condition for the neuroblastomas. Identifying those nodes which are used with a high frequency in aggressive tumour when compared to the non-aggressive tumour may be useful for potential drug targets. When such genes are knocked-down or silenced the signalling of the aggressive tumour can be

targeted leading to breakdown of signalling machinery of the cancer cell. Therefore the frequencies for tumour cell and normal cell were compared.

The nodes with differential node frequency, that is, those with high node frequency in aggressive and less node frequency in non-aggressive were calculated. The stage1 condition contains 65 patients and stage4A condition contains 17 patients. Due to unequal number of patient samples the datasets were stratified into 3 sets of stage1 patients with 17 samples each. The shortest paths were calculated for 4 sets – three sets of stage1 and one set of stage4A. The node frequencies of 3 sets of stage1 were subtracted from stage4A node frequencies ending up in three lists of difference of node frequencies. Finally, a rank-product(RP) test was used to calculate the top ranking nodes from the three lists. The Rank product tests have previously been used to calculate differentially expressed genes in replicated microarray experiments (Breitling, et al., 2004). One of the advantage of this test is that it can be used for experiments even with a very less number of replicates. The significance of genes in a rank product test can be calculated as follows:

1. Calculate the ranks of the x genes in m replicates. Here m=3 for three lists of node frequency difference.
2. Calculate the product of ranks(RP) divided by x^m. That is, if ranks for the three lists are r1,r2 and r3 then RP value is $(r1*r2*r3)/ x^m$.
3. Sort all the data by increasing RP value, the most significantly used genes will be at the top of the list.

Chapter 3

Results

3.1 Properties of the cancer signalling network

Gene expression data of 10 cancer datasets comprising of one dataset each of breast cancer, prostate cancer, oral tongue carcinoma, acute myeloid leukemia, renal cell carcinoma, vulvar interstitial neoplasia, cervical cancer and two datasets each of lung adenocarcinoma and were used for detecting properties of the signalling networks of cancer. A paired wilcoxon test was used to calculate the statistical significance of the obtained results.

3.1.1 Cancer showed shorter signalling pathways

For each dataset shortest paths or the predicted pathways were calculated for normal and cancer samples. The path length for each of these pathways is the number of proteins in the pathway starting from the receptor to the transcription factor. The average path length for all the paths from different receptor-transcription factor pairs for the normal and cancer was calculated. Path lengths for the normal samples were significantly longer (mean for cancer 5.73:, mean for normal: 6.04, P= 0.009) longer. The results are given in Table 2.

3.1.2 Tumours use more edges and less hubs

There was an interest to know how often the same edges (interactions) were used for different signalling pathways. For this, the mean frequency of every edge to be involved in a receptor-transcription-factor pathway was calculated. This frequency was obtained by the number of edges used in each single pathway divided by the number of all used edges. The edge frequency was higher in normal cells (mean for cancer: 40.58, mean for normal: 51.59, P = 0.001). Similarly, the node frequency was calculated and showed the same tendency (mean for cancer: 154.96, mean for normal: 178.79, P = 0.001). Hence normal

cells used more often the same central nodes and interactions for different signalling tasks. The results are given in Table 2.

Table 2. Features of signalling network

	Lung-1	Lung-2	Breast	Prostate	HN	OT	AML	Renal	Vulva	Cervical	P-value	Tendency for cancer
Signalling motifs												
Normal	351764	647730	371984	83356	136676	256492	140030	67706	115240	167852	0.001	Down
Cancer	295884	397290	278772	63722	134056	147170	26294	60346	82326	89800		
Path length												
Normal	5.4	5.66	5.24	6.2	5.76	5.66	6.65	5.93	6.36	7.55	0.009	Down
Cancer	5.36	5.4	5.17	5.84	5.68	5.13	5.56	5.92	6.37	6.89		
Link frequency												
Normal	37.43	49.87	39.21	63.52	51.11	47.50	55.41	38.15	51.55	82.21	0.001	Down
Cancer	34.65	42.26	34.26	51.86	48.31	32.72	27.91	35.19	43.70	55.00		
Node frequency												
Normal	160.2	218.0	186.8	237.3	199.1	189.8	151.0	89.9	152.4	203.5	0.001	Down
Cancer	158.8	197.9	168.8	208.8	195.7	153.8	98.4	75.3	137.1	155.1		
Size of the network												
Normal	5014	6734	7799	3490	3974	4185	2176	2758	2691	2064	0.001	Up
Cancer	5362	7501	8770	3977	4132	5376	3485	2984	3184	2774		
Clustering coefficient												
Normal	0.049	0.048	0.053	0.054	0.052	0.049	0.030	0.038	0.032	0.024	0.09	Up
Cancer	0.052	0.053	0.055	0.058	0.052	0.064	0.041	0.033	0.034	0.024		
Clustering coefficient >0												
Normal	521	729	829	412	445	459	217	279	261	182	0.005	Up
Cancer	599	807	935	457	473	617	382	308	338	235		
Integration motifs												
Normal	230	442	188	565	229	840	397	223	80	116	0.01	Down
Cancer	91	471	152	482	179	634	312	220	74	83		
Maintenance motifs												
Normal	14	177	8	59	227	65	131	2	64	56	0.02	Up
Tumour	51	211	13	95	291	119	123	2	104	57		

Increase of average path length after hub removal												
Normal	1.65	1.64	1.65	1.71	1.79	1.64	1.03	1.70	1.79	1.64	0.01	Down
Cancer	1.56	1.63	1.56	1.66	1.63	1.52	1.03	1.68	1.64	1.67		
Information entropy												
Normal	10.74	10.88	11.26	10.26	10.33	10.59	9.05	10.04	9.67	9.48	0.001	Up
Cancer	10.87	11.18	11.53	10.43	10.35	10.92	10.48	10.07	9.93	9.98		

3.1.3 The used signalling network is less centralized

Barabasi and co-workers used the clustering coefficient for a measure of inter-connectivity of networks (Barabasi and Oltvai, 2004). We calculated the clustering coefficient and obtained higher values in cancer supporting our findings that cancer showed a more connected, less centralized structure (mean of cancer: 0.046, mean of normal: 0.042, P = 0.09). Also the number of nodes with a clustering coefficient greater zero was higher in cancer cells (mean for cancer: 515.10, mean for normal: 433.40, P = 0.005). The results are given in Table 2.

3.1.4 Tumour networks are more robust against directed attacks

The robustness of the used network was checked by removal of the top 50% of the hubs. When a certain number of hubs were removed, the network becomes disconnected forming one large component and several small clusters. After removal of the top 50% of the highly connected nodes were removed and the number of clusters in the normal and cancer used network were counted. Except for two cancers the rest showed either equal or higher number of clusters in normal when compared to cancer. The results are shown in Table 3 . This shows a slightly more robust cancer signalling network.

Table 3. Number of clusters after hub removal

Cancer type	Normal	Cancer
Lung1	4	1
Lung2	5	3
Breast	2	1
Prostate	2	2
Head and Neck	4	2
Oral tongue	5	2
AML	7	3
Renal	4	3
Vulva	4	3
Cervical	7	5

Albert and co-workers showed that the average path length of a network increases as nodes are removed randomly (Albert, et al., 2000). In order to understand the effect of hub removal on robustness of the network, up to top 20% of the hubs were removed systematically. The ratio of average path length on hub removal to the average path length without hub removal were obtained for 1-20% of hub removal (Figure 6). The plots show that as a larger percentage of hubs are removed, the average path length increases more for normal samples when compared to cancer. This again confirms a higher dependency on hubs in networks of normal tissue. Also, the cancer network is comparatively more robust to removal of hubs. The area under the curve (AUC) for hub removal is significantly higher in normal (mean for cancer: 32.32, mean for normal: 34.98, $P = 0.001$) as shown in Table 2. A plot for the area under the curve for normal and cancer for different datasets is given in Figure 7.

3.1.5 Frequently involved genes are enriched with cancer mutated genes

Cui and co-workers compiled a list of 284 cancer mutated genes which were derived from large scale sequencing studies and other literature (Supplementary table S10 in (Cui, et al., 2007)). This list was compared with the 20 most frequently involved nodes (hubs) of each network and significant enrichment was found for 9 out of 10 normal and tumour samples (Table 4). Then gene-lists of cancer mutated hubs for every cancer were defined by intersecting the hubs of the network with the list of cancer mutated genes of Cui *et al.* (Table 5). Interestingly, most of the genes which showed up in the tumour networks were also present in the normal networks. This may indicate that normal cells intrinsically pave the way for their specific evolvement into malignancy.

Figure 6. This figure shows the effect of hub removal on average path length of the network in different cancer datasets. (Black represents normal and red represents cancer).

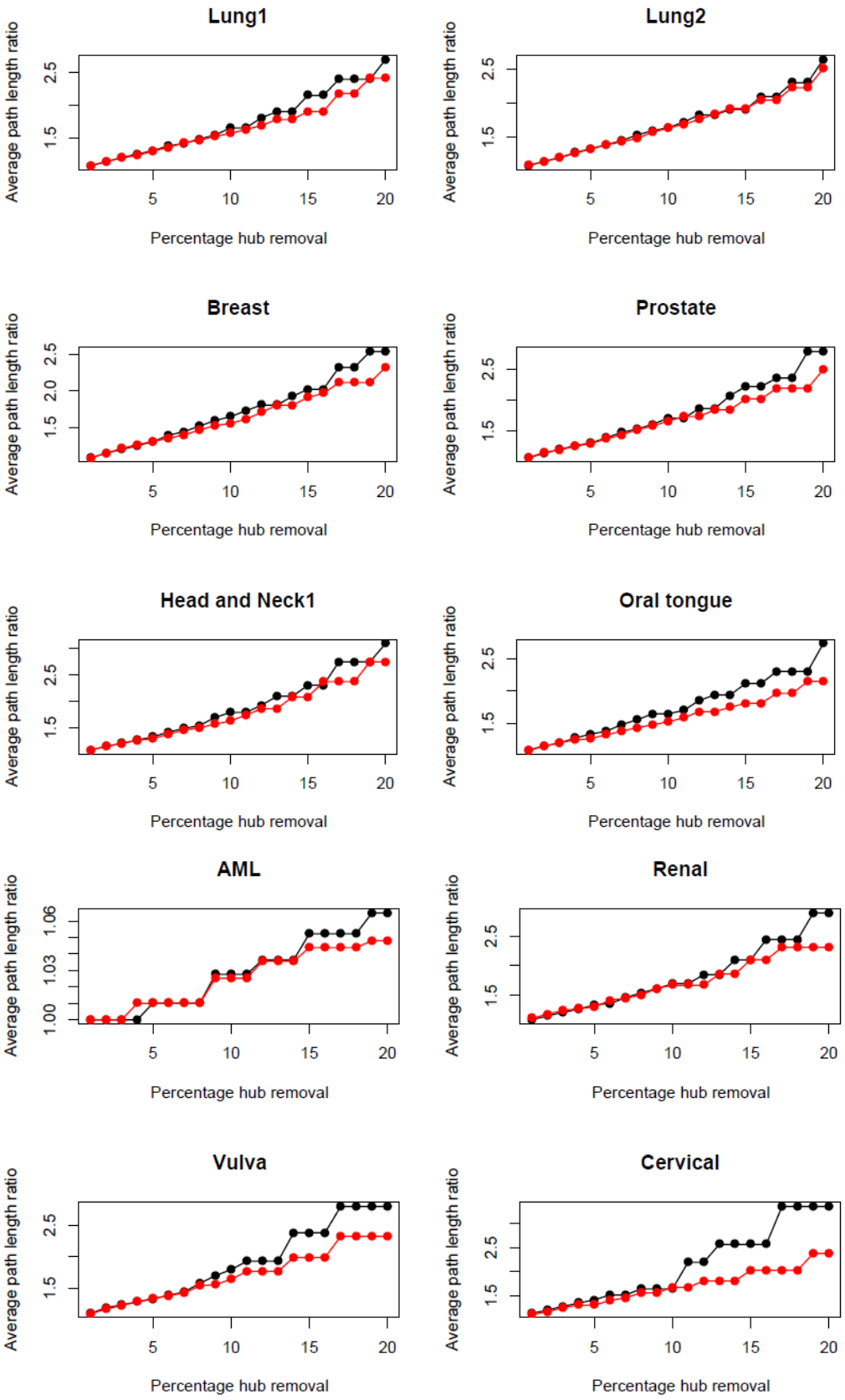
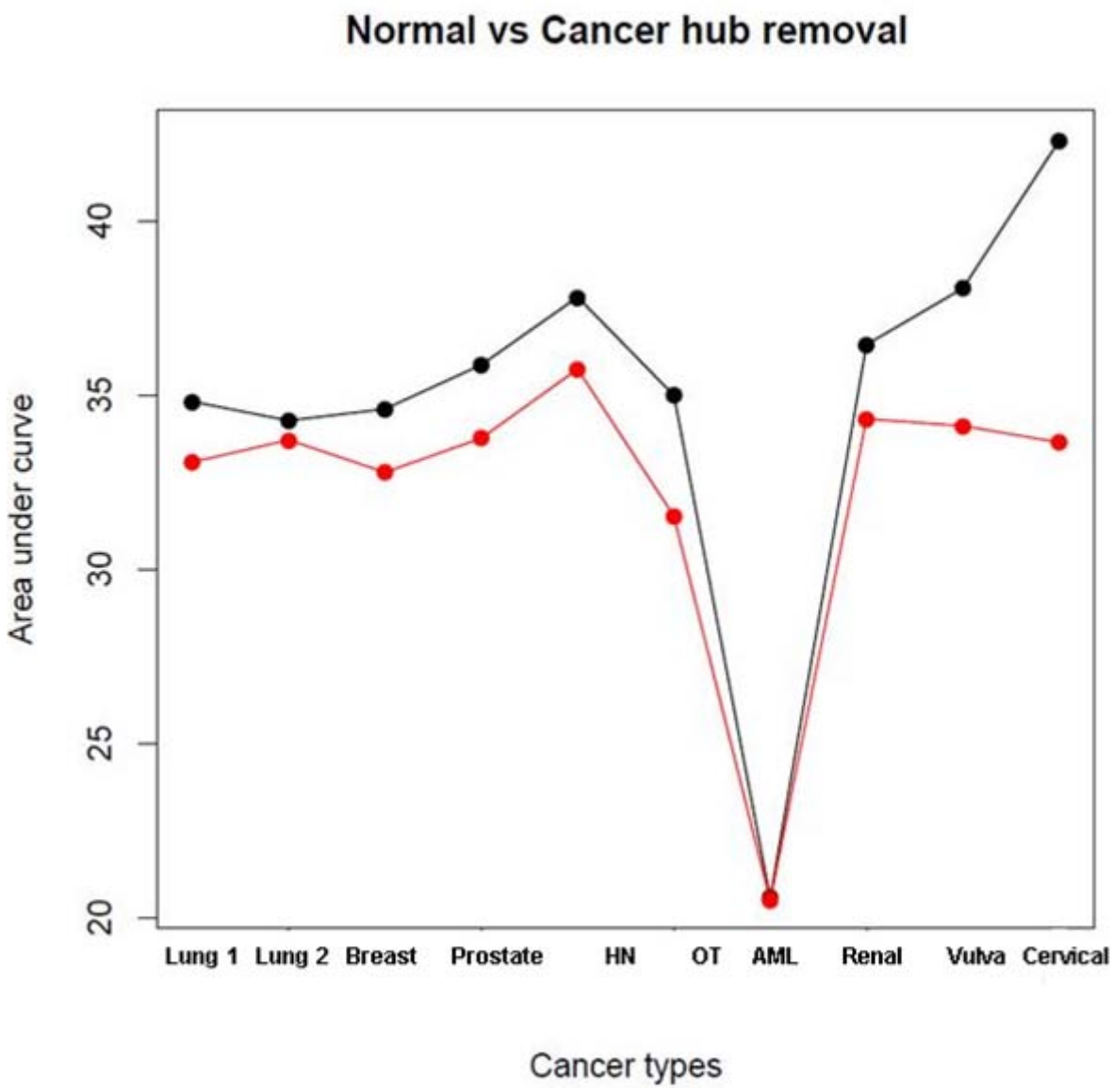


Figure 7. This figure shows the area under the curve for the previous graph for different cancer types



3.1.6 Signalling-regulation in cancer is detached at cancer mutated hubs but maintained in their vicinity

Uri Alon and his co-workers have studied occurrences of direction-motifs in triangles and revealed a large variety of substantial characteristics in signalling networks characterized by consistent and non-consistent feed-forward and feedback loops (Alon, 2007). Local regulation patterns of the networks at cancer mutated hubs were studied. For this, regulation motifs of every triangle consisting of at least one hub and two of its neighbours

which on their part also interact were analyzed. Two regulation motifs were defined. The first motif reflected the degree of regulatory integration of a hub and its network-vicinity and was defined by a high correlation of all pairs of nodes in the triangle motif (integrated motif, motif A in Figure 9). This motif was found significantly more often in normal cells (P = 0.01, Table 2). The second motif (maintenance motif, motif B in Figure 9) described a hub which regulation is independent from its vicinity (low correlation between the hub and its two neighbours in the motif), but which vicinity is co-regulated (high correlation of the two neighbours). Such a scenario is reasonable for a mutated cancer protein with loss of function but which neighbours maintain signalling propagation. Indeed, this motif occurred more often in the cancer networks (P = 0.02, Table 2).

Figure 8. Frequency distribution for breast cancer (red, circles) and the corresponding normal sample (blue, crosses). Both networks showed the typical scale-free distribution for the frequency of proteins being involved in our defined signalling pathways. Proteins in the cancer network exhibited a distinct shift to the left indicating less frequency not only for the hubs but for all proteins in the network. Both distributions were fitted by a combined linear model of same slopes but different intercepts for normal and cancer cells.

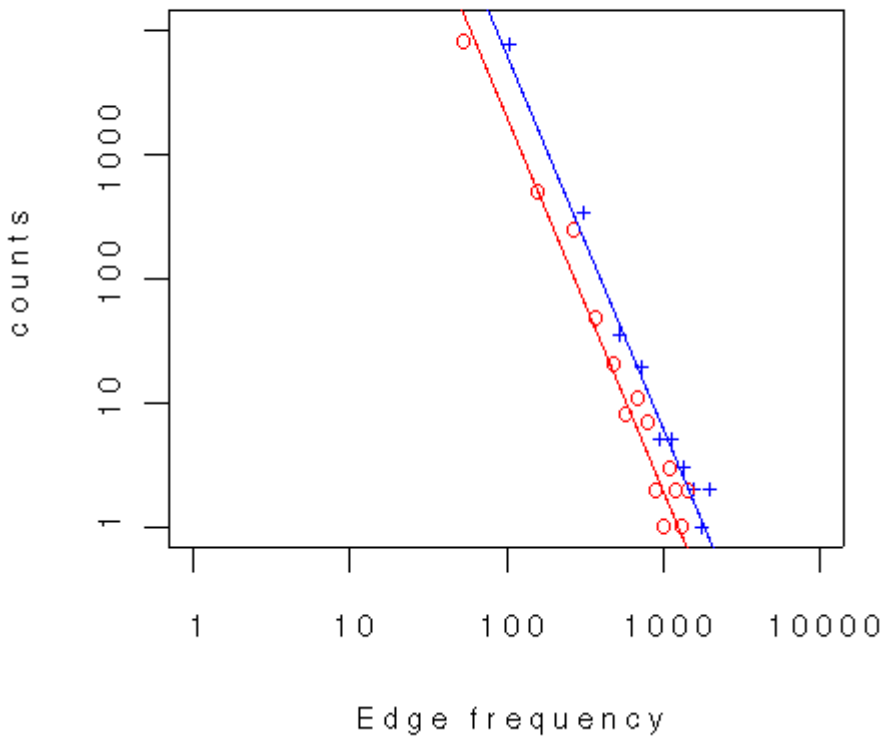
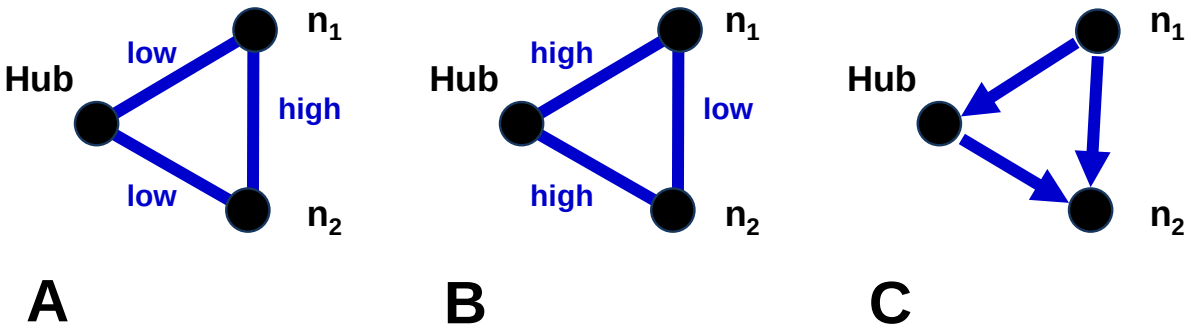


Figure 9. Triangle motifs. The motifs were derived for each triple of nodes consisting of a hub and two of its network-neighbours (n_1, n_2) which on their part were also connected. In the integration motif (motif A) all nodes are pair-wise co-regulated. Accordingly, the motif is defined by low distances for links hub-n_1, hub-n_2 and n_1-n_2. In the maintenance motif (motif B) only n_1 and n_2 are co-regulated. It is defined by a low link-distance for n_1-n_2 and high link-distances for hub-n_1 and hub-n_2. Motif C is a consistent feed-forward loop, taken from the literature (Alon, 2007).



3.1.7 A novel motif for degenerate signalling

From the obtained results a new signalling motif was designed which is illustrated in Figure 10. Within this model, cancers use differentiated pathways whereas normal cells utilize the same signalling interactions for different tasks. Cancer utilizes pathways with different interactions by different operator-receiver pairs (R_1 - TF and R_2 – TF in Figure 10) whereas normal cells utilize common interactions for this task. We compared the abundance of this motif with the abundance of its counterpart in which cancer pathways were different and common pathways were used in the normal samples. The significantly higher number of the motif in normal (mean for cancer: 157566, mean for normal: 233883, P = 0.001) supported the idea of "decentralized signalling" in the cancer samples under study (Table 2).

Figure 10. Comparative cancer motif. Two different signals are transmitted from two receptors (R1 and R2) to a transcription factor (TF). Green and grey arrows indicate the pathways for normal and cancer cells, respectively. The motif was defined for each pair of pathways (R1,TF) and (R2,TF) such that the pathways of normal cells share at least one common link whereas the pathways for cancer cells didn't share any link.

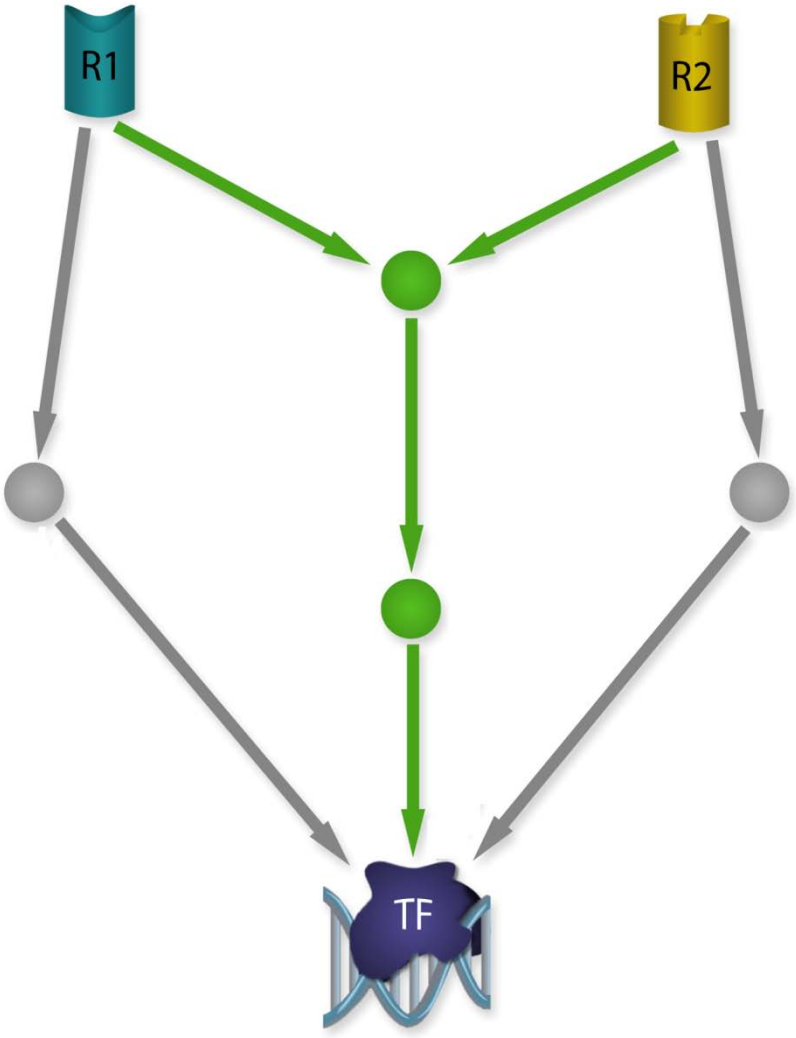


Table 4. Cancer mutated genes are significantly enriched in the most frequently involved nodes (hubs)

Dataset	Network size	Cancer genes in our hubs	Cancer genes in the whole network	P-value
AML normal	920	8	80	1.57e-03
AML tumour	1186	11	96	1.34e-05
Breast normal	2003	9	165	2.49e-04
Breast tumour	2187	7	177	3.37e-03
Cervical normal	941	6	88	2.26e-02
Cervical tumour	1131	7	94	4.17e-03
Head and neck normal	1214	6	113	2.17e-02
Head and neck tumour	1218	6	115	2.29e-02
Lung1 normal	1418	10	136	1.95e-04
Lung1 tumour	1418	8	134	2.30e-03
Lung2 normal	1851	9	154	2.70e-04
Lung2 tumour	1946	12	149	1.41e-06
Oral tongue normal	1252	7	114	6.40e-03
Oral tongue tumour	1401	8	131	2.17e-03
Prostate normal	1094	2	100	5.63e-01
Prostate tumour	1172	3	111	3.25e-01
Renal normal	707	9	59	3.63e-04

Renal tumour	792	11	59	8.54e-06
Vulva normal	1060	9	91	3.77e-04
Vulva tumour	1184	11	99	1.76e-05

Table 5.Intersection of the hubs and cancer mutated genes

Dataset	**Gene symbols (EntrezGene ids in brackets)** of the intersection of the top 20 most frequently involved genes (hubs) and cancer mutated genes from table S10 of Cui and co-workers (Cui, et al., 2007)
AML normal	CBP (1387) bARK (156) EGFR (1956) p300 (2033) Hsp90 (3320) JAK1 (3716) c-Myc (4609) nPKC (5579)
AML tumour	CBP (1387) bARK (156) EGFR (1956) p300 (2033) ABL1 (25) INSR (3643) JAK1 (3716) SMAD2 (4087) PKCA (5578) PKCz (5590) SHP2 (5781)
Breast normal	CBP (1387) bARK (156) Hsp90 (3320) JAK1 (3716) SMAD2 (4087) SMAD3 (4088) PKCA (5578) RB (5925) SRC (6714)
Breast tumour	CBP (1387) bARK (156) p300 (2033) SMAD2 (4087) SMAD3 (4088) PKCA (5578) SRC (6714)
Cervical normal	CBP (1387) EGFR (1956) p300 (2033) PKCz (5590) RB (5925) TYK2 (7297)
Cervical tumour	CBP (1387) p38 (1432) p300 (2033) ABL1 (25) Hsp90 (3320) SMAD2 (4087) PKCA (5578)
Head and neck normal	CBP (1387) p300 (2033) FYN (2534) JAK1 (3716) SMAD2 (4087) SMAD4 (4089)
Head and neck tumour	CBP (1387) p300 (2033) FYN (2534) LCK (3932) SMAD2 (4087) SMAD4 (4089)
Lung 1 normal	CBP (1387) p300 (2033) SMAD4 (4089) PKCA (5578) nPKC (5579) SHP2 (5781) RAF1 (5894) RB (5925) SRC (6714) p53 (7157)
Lung 1 tumour	CBP (1387) p300 (2033) FYN (2534) SMAD2 (4087) SMAD3 (4088) SMAD4 (4089) PKCA (5578) SRC (6714)
Lung 2 normal	CBP (1387) p300 (2033) FYN (2534) Hsp90 (3320) JAK1 (3716) VEGFR (3791) SMAD2 (4087) RAF1 (5894) RB (5925)
Lung 2 tumour	CBP (1387) CTNNB1 (1499) p300 (2033) GAQ (2776) Hsp90 (3320) INSR (3643) JAK1 (3716) VEGFR (3791) SMAD2 (4087) PKCA (5578) RAF1 (5894) RB (5925)
Oral tongue normal	CBP (1387) SMAD2 (4087) SMAD4 (4089) PKCA (5578) RAF1 (5894) SRC (6714) p53 (7157)
Oral tongue tumour	CBP (1387) p300 (2033) Hsp90 (3320) LCK (3932) SMAD2 (4087) SMAD4 (4089) PKCA (5578) SRC (6714)
Prostate normal	SMAD2 (4087) SMAD4 (4089)
Prostate tumour	p300 (2033) SMAD2 (4087) SMAD4 (4089)
Renal normal	CBP (1387) EGFR (1956) ABL1 (25) Hsp90 (3320) JAK1 (3716) SMAD2 (4087) PKCA (5578) RAF1 (5894) RB (5925)
Renal tumour	CBP (1387) p38 (1432) EGFR (1956) ABL1 (25) Hsp90 (3320) INSR (3643) JAK1 (3716) SMAD2 (4087) PKCA (5578) RAF1 (5894) RB (5925)
Vulva normal	CBP (1387) bARK (156) EGFR (1956) p300 (2033) ABL1 (25) JAK1 (3716) SMAD2 (4087) PKCA (5578) RB (5925)
Vulva tumour	IKKA (1147) CBP (1387) bARK (156) p300 (2033) INSR (3643) JAK1 (3716) LYN (4067) nPKC (5579) RAF1 (5894) RB (5925) ZAP70 (7535)

3.1.8 Neuroblastoma – properties of its cancer signalling network

Neuroblastoma is a malignant tumour consisting of undifferentiated neuroectodermal cells derived from the neural crest. As is characteristic case of embryonic tumours, neuroblasts are histologically indistinguishable from developing neuroblastic cells in the embryo. Neuroblastoma is the most common malignant disease in children with 7.5 cases for every 100,000 infants. There are 1.3 new cases per 100,000 children under the age of 15 years every year, which accounts for 9% of all childhood cancers. Almost, 90% of children with the disease are diagnosed in their first 5 years (Schwab, et al., 2003). Neuroblastoma is often unpredictable, because it is associated with contrasting patterns of clinical behaviour, ranging from life-threatening progression, maturation to ganglio-neuroblastoma or ganglioneuroma, and spontaneous regression. The “age” and “stage” of the patients enable physicians to predict, to some extent, the clinical course of the disease which is supported with histologic information.

A significant proportion of tumours (>10%) undergo complete spontaneous regression in the absence of or with minimal therapeutic intervention. The spontaneous regression is evident as primary neuroblastoma and metastatic disease disappear without any treatment. This situation is generally associated with a clinically recognisable syndrome called 4s, defined as a small primary tumour in the abdomen or thoracic cavity accompanied with metastasis in the liver or bone marrow and skin (or both) but not in the cortical bone. Although spontaneous regression is most commonly observed in patients with stage 4s, it is also well described in stage 1–3 neuroblastoma in children and older patients. Spontaneous maturation to benign ganglioneuroma is much less frequent than spontaneous regression. Additionally, several biological markers have been found to describe therapeutic risk groups (Schwab, et al., 2003).

From the clinical perspective a better prediction of tumour behaviour at diagnosis will help to avoid overtreatment of spontaneously regressing tumours and treatment failure in high-risk patients. The current prognostic evaluation is based primarily on the extent of tumour spread at diagnosis and age of the patient. The problem is that this classification overestimates the number of patients who need chemotherapy, because it cannot define patients with stage 1–3 tumours that are regressing. Recently, several biological markers have been incorporated to describe therapeutic risk groups. Such an approach involving

molecular classification by detection of amplified *MYCN* and deletion of 1p chromosomal material underestimates the proportion of high-risk patients because only about 30% of the children with high-risk stage 4 disease have amplified *MYCN* and 47% have 1p alteration (Schwab, et al., 2003).

Oncogene *MYCN* is a member of the MYC family of oncogenes that encode nuclear proteins serving as transcription factors. In neuroblastoma, amplified MYCN is a strong prognostic indicator of poor prognosis, particularly in localized tumours where patients with normal *MYCN* gene dosage fare quite well. The active MYC family of genes usually due to genetic damage result in enhanced expression of wild-type proteins has important indications in human and animal cancers. Mostly, the activation mechanism involves the increase of the *MYCN* gene dosage, either by amplification resulting in up to several hundred gene copies or by more subtle mechanisms, like duplication or polyploidization. Nowadays, the status of *MYCN* is used widely as a standard marker for neuroblastoma stratification (Schwab, 2004). Apart from *MYCN* amplification, cytogenetic and molecular level analysis of tumours identified non-random genetic changes, including ploidy changes, deletions of chromosome 1p, gains of chromosome arm 17q, and deletions of 11q as well as deletions of other genomic regions that allow tumours to be classified into subsets with distinct biological features and clinical behaviour (Westermann and Schwab, 2002).

In this study gene expression datasets of 65 non-aggressive Stage1 and 17 aggressive Stage4A (Stage4 *MYCN* amplified) classes were used for a comparative analysis. Stage 1 tumours will be denoted as “Normal” in the following, in contrast to the denotation of “Cancer” for stage 4A tumours. Due to an unequal number of samples in the two classes, the gene expression data was stratified to obtain correlation values and three groups of stage 1 samples were assembled by random selection, yielding the groups “Normal 1”, “Normal 2” and “Normal 3”. Normal 1 group gives consistent results consistent with the previous analysis for 10 cancer datasets. However, the other two groups – Normal 2 and Normal 3 showed opposite trends. This could be because of the heterogeneity within the normal samples that lead to fluctuations in the resulting correlation values. The results are shown in Table 6.

Table 6. Features of signalling network - Neuroblastoma

	Normal1	Normal2	Normal3	Cancer1	Cancer2	Cancer3
Signalling motifs	944072	1023018	1013944	847492	994278	1036862
Path length	5.51	5.35	5.45	5.38		
Edge re-usage per edge	86.50	80.66	86.47	84.89		
Average node frequency	432.54	423.75	440.53	428.31		
Size of used network	6996	7232	6907	6922		
Average clustering coefficient of used network	0.069	0.079	0.072	0.071		
Clustering coefficient >0	810	876	815	828		
AUC for hub removal	121.72	119.58	119.13	122.09		
Information entropy - edge	10.96	11.03	11.06	11.01		

To obtrain proteins which were highly involved in tumour signalling and comparatively few involved in normal signalling, proteins with differential node frequency between stage 1 and stage 4A in neuroblastoma were identified using the method described in the "Methods" section. Table 7 below shows a list of the top few genes obtained from the analysis. The table contains the genes identified in the analysis, the significance value from the rank-product test, the gene regulation (up or down regulated), the P-value obtained by wilcoxon test and the ranking based on differential gene expression. The ranking based on differential gene expression shows that some proteins which have a lower ranking in the conventional analysis are in the top of the list in the analysis used in this thesis showing that it is a new approach to detect highly used signalling nodes in a network. The up-regulated genes in the list could be candidates where a knock-down assay could better elucidate the involvement of these proteins in aggressive neuroblastoma cells.

Table 7. Significant genes with differential node frequency

Gene symbol	Rank product test	Gene regulation	Significance testing by Wilcoxon test	Ranking based on differential expression
MAPK1	2.46e-09	Down	0.024	490
PRKCA	1.47e-08	Up	0.16	593
PRKACA	2.36e-07	Up	0.16	593
PRKCD	9.45e-07	Down	0.0001	278
CDC25B	9.97e-07	Up	0.045	522
LYN	1.03e-06	Down	0.038	513
HTT	2.17e-06	Up	0.037	512
CAV1	2.64e-06	Down	0.009	451
CALR	3.49e-06	Up	0.45	652
MAPK3	4.91e-06	Down	0.0002	300
PAK1	6.30e-06	Up	0.17	594
AKT1	6.80e-06	Up	0.24	611
CSNK2A1	7.29e-06	Up	5.13e-05	256
LCK	7.75e-06	Down	0.0001	285
GNAQ	8.74e-06	Down	0.25	613
GNA15	8.80e-06	Down	0.96	713
SYK	8.84e-06	Down	1.70e-05	221
PIK3R1	1.08e-05	Down	9.15e-05	273
LCP2	1.25e-05	Down	1.61e-05	220
VIM	1.38e-05	Down	0.36	635
EGFR	1.39e-05	Down	0.43	648
RGS14	2.36e-05	Down	0.84	700
TUBB	2.88e-05	Down	0.004	408
PSMC5	3.61e-05	Up	0.50	659
GNAO1	3.66e-05	Down	0.48	656
PLCB1	3.76e-05	Down	0.008	443
ARRB2	4.94e-05	Down	0.001	367
SRC	4.95e-05	Down	0.88	704
PTPN6	5.24e-05	Down	0.005	419
BTK	5.50e-05	Down	0.014	467
YWHAG	6.30e-05	Down	1.64e-06	162
CASP7	6.89e-05	Down	0.005	415

The oncogenic relevance of some of these genes in neuroblastoma or cancer in general have been assembled by intensive literature search. A short summary of the results are given below:

1. MAPK1 (mitogen-activated protein kinase 1): MAP kinases, also known as extracellular signal-regulated kinases (ERKs), act as an integration point for multiple biochemical signals, and are involved in a wide variety of cellular processes such as proliferation, differentiation, transcription regulation and development. In SK-N-MC neuroblastoma cells, the basic fibroblast growth factor (FGF2) induces apoptosis by directing the formation of the ERK (or MAPK1)-GSK3beta (Glycogen synthase kinase-3beta) complex. This complex formation results in retaining ERK in the cytoplasm, otherwise it leads to ERK nuclear translocation (Ma, et al., 2008). MAPK1 was down-regulated in stage 4A thereby contributing to survival rather than apoptosis.

2. PRKCA (Protein kinase C, alpha): Protein kinase C (PKC) is a family of serine- and threonine-specific protein kinases that can be activated by calcium and the second messenger diacylglycerol. This kinase has been reported to be involved in many different cellular processes, such as cell adhesion, cell transformation, cell cycle checkpoint, and cell volume control. GSK 3-beta is implicated in regulation of apoptosis. In human SH-SY5Y neuroblastoma cells, PKC-alpha and PKB proteins were activated by apoE4 leading to GSK-3beta inhibition (Cedazo-Minguez, et al., 2003). Inhibition of GSK-3beta thus leads to escape from apoptosis. This relates well with the gene expression data where PRKCA is up-regulated in stage 4A.

3. PRKACA (protein kinase, cAMP-dependent, catalytic, alpha): The second messenger cAMP exerts its effects by activating the cAMP-dependent protein kinase, which transduces the signal through phosphorylation of different target proteins. In the SH-SY5Y human neuroblastoma cell line the regulatory mechanism of Bcl-2 proteins was investigated. Bcl-2 proteins are involved in serum depletion-induced apoptosis. In SH-SY5Y cell lines, it was shown that Bcl-2 was negatively regulated by PKA (Itano, et al., 1996). They are also up-regulated in stage4A thereby leading to cell survival. Also in SH-SY5Y cells, in presence of cAMP PKA was activated leading to neurite outgrowth (Sanchez, et al., 2004). In stage 4A cells the PRKACA or PKA was up-regulated.

4. PRKCD: Protein kinase C (PKC) is a family of serine- and threonine-specific protein kinases that can be activated by calcium and the second messenger diacylglycerol. The

protein encoded by this gene is one of the PKC family members. Studies in human and mice demonstrated that this kinase is involved in B cell signalling and in the regulation of growth, apoptosis, and differentiation of a variety of cell types. In PC12 rat pheochromocytoma cells, selective activation of PKC delta may play a role in neuritogenic signals in PC12 cells (O'Driscoll, et al., 1995).

5. CDC25B: It is a member of the CDC25 family of phosphatases. CDC25B activates the cyclin dependent kinase CDC2 by removing two phosphate groups and it is required for entry into mitosis. CDC25B shuttles between the nucleus and the cytoplasm due to nuclear localization and nuclear export signals. The protein is nuclear in the M and G1 phases of the cell cycle and moves to the cytoplasm during S and G2. In neuroblastoma, overexpression of the proto-oncogene N-MYC is correlated with malignancy. It was found that CDC25B expression levels were significantly correlated with N-MYC m-RNA levels suggesting that CDC25B may play an active role as a target of N-MYC (Sato, et al., 2001).

6. LYN: It is an src-related intracellular protein tyrosine kinase. It acts as a signal transducing molecule for surface immunoglobulin M and is expressed predominantly in hemopoietic cells. Expression of LYN has been reported in neuroblastoma. It was reported that in surgical tumour samples LYN transcripts were found preferentially at early stages whereas they were barely detectable in highly malignant tumours. I was also proposed that LYN may be involved in a signalling pathway of neuroblasts committed to neuronal differentiation (Bielke, et al., 1992).

7. HTT: Huntingtin is a disease gene linked to Huntington's disease, a neurodegenerative disorder characterized by loss of striatal neurons. This is thought to be caused by an expanded, unstable trinucleotide repeat in the huntingtin gene, which translates as a polyglutamine repeat in the protein product. It is predicted that mutated htt acquires toxic properties in specific brain regions. It was reported that transfection of mutant HTT in SK-N-MC neuroblastoma cells that endogenously express D1 receptors was associated with a minor increase in cell death (Robinson, et al., 2008). This correlates well with the gene expression data where HTT is down-regulated in stage 4A.

8. CAV1: The scaffolding protein Caveolin-1 encoded by this gene is the main component of the caveolae plasma membranes found in most cell types. In studies on CAV1 was suggested that Caveolin-1 inhibits neurite growth by blocking Rac1/Cdc42 and p21-

activated kinase 1 interactions in the basic fibroblast growth factor receptor (bFGF) pathway (Kang, et al., 2006).

9. CALR: Calreticulin is a multifunctional protein that acts as a major Ca(2+)-binding (storage) protein in the lumen of the endoplasmic reticulum. It is also found in the nucleus, suggesting that it may be involved in transcription regulation. It was reported that Calreticulin can inhibit the binding of the androgen receptor to its hormone-responsive DNA element and inhibit androgen receptor and retinoic acid receptor transcriptional activities *in vivo*, as well as retinoic acid-induced neuronal differentiation. This shows that Calreticulin can act as an important modulator of the regulation of gene transcription by nuclear hormone receptors.

10. PAK1: As already discussed above, it is suggested that the up-regulated caveolin-1 in neuronal cells can inhibit neurite outgrowth by interfering with the bFGF signalling pathway from small GTPases to PAK1 by directly binding to PAK1 (Kang, et al., 2006). In our data it is known that PAK1 was up-regulated in stage 4A. It was also shown that PAK1 was involved in neuronal migration where adhesion molecule L1 stimulates neuronal migration through Vav2-Pak1 signalling (Schmid, et al., 2004).

Chapter 4

Discussion

In the above analysis, cancers exhibited very distinct mechanisms in signal transduction when compared to the normal samples. The cancers showed shorter signalling paths and more differentiated pathways. Luscombe and co-workers analyzed the dynamics of regulatory networks in yeast (Luscombe, et al., 2004). In comparison to endogenously caused changes, they discovered large differences in topological changes when yeast responded to environmental changes. For having quick responses, yeast reacted to environmental changes (nutrition depletion, stress response) by short regulatory cascades. Interestingly, these findings can be compared to the regulation of the signalling network for the human cancers studied here. The cancer cell is similar to the yeast cell under stress forcing it to organize short regulatory cascades.

A higher average clustering coefficient was observed in the used signalling network in cancer. The used network in cancer was larger with less edge and node frequency in cancer all pointing to high inter-connectedness of the cancer signalling using different routes for signal transfer. There was also a lower entropy in the cancer signalling network signifying less order also showed by other network parameters. The cancers showed a tendency for disparate signalling in contrast to the normal cells. Cancers utilized different signalling pathways for same tasks, modelled by pathways between pairs of receptors and transcription factors, this result correlates well with the result that the cancer signalling network is a more inter-connected network. The used network for cancer is much more diverse with more connections and links. These results hint that there is a less dependency on hubs in the cancer signalling network.

In order to determine if the cancer signalling network was less dependent on hubs, systematic removal of hubs was performed. Previous studies have shown that even after random removal of up to 5% of the nodes from a scale-free network still doesn't harm it. This is because nodes with less connectivity would be selected with greater probability when compared to the hubs which are less in number. On the other hand, removal of hubs cause the network to break down into smaller clusters indicating vulnerability to these attacks (Albert, et al., 2000). We detected that the different network toplogy in the

network of cancers made the networks more robust against these attacks. When the top 50% of highly connected nodes were removed, the number of clusters in the network were counted. The networks of normal samples had more clusters (only one cancer showed same number of clusters). These results show a clear tendency towards a higher robustness in cancer signalling due to a different wiring when compared to the normal signalling network. Furthermore, the increase in average path length of all possible pairs of paths in the network to hub removal was seen. The increase in path length was less for the cancers when compared to the normal samples. From these results, it can be concluded that the cancer signalling networks were more robust to hub removal when compared normal networks. Hence, the higher inter-connectedness (or more links) contributed by disparate signalling in cancer could provide additional robustness or tolerance to attacks making targeted drug design very challenging.

An analysis was done confirming the findings of Cui and co-workers (Cui, et al., 2007), that cancer specific mutations occur distinctively more often at hubs for signal transduction. Such a mutation can cause a loss of function. This is beneficial for the cancer if the protein gets insensitive to regulation-upstream-signals and fires constitutively an oncogenic signal as e.g. the ABL-BCR fusion protein in chronic myelogenous leukemia (Druker, 2008). If the protein acts as a tumor suppressor, a complete loss of function is beneficial for oncogenesis. In both scenarios, the regulation for signaling homeostasis of the local network environment is detached from this mal-functional protein and a coordinated regulation between the environment and this protein is not necessary any more. This was observed by counting distinctively less integration-motifs in tumors (motif A in Figure 9). Interestingly, tumors still sustain the original signals between the environment. This was observed by higher counts of the maintenance motif in tumors which reflects low co-regulation between hubs and their neighbors, but high co-regulation between the neighbors of the hubs (motif B in Figure 2). Even though tumors may exhibit de-regulation of mal-functional hubs with their neighbors, such a maintained co-regulation of their neighbors gives evidence that bypass regulations are still necessary. Ma'ayan and co-workers observed an accumulation of feedback and feed-forward loops at such hubs (Ma'ayan, et al., 2005) which supports this idea. Tumors need to maintain the direct signal of e.g. a feed-forward loop which is necessary for the effect of the constitutive signal of an oncogenic hub. Such oncogenic signaling motifs may have implications to drug therapy. If an oncogenic hub is treated (as e.g. ABL-BCR with imatinib (Druker, 2008)) resistance can occur by mutations of the target protein which reduce the affinity of the drug to the

target. A combined therapy may avoid this evolvement by additionally blocking the neighboring signaling-maintenance. In addition, the observed cancer networks showed higher error tolerance against directed attacks of hub removals. Hence, some maintenance signals may not only support cancer mutated hubs but also pave the way for the signaling network to get independent of them, specifically for proteins of cancer mutated genes with a complete loss of function. It is challenging but highly relevant to shed light into these effects experimentally with cell lines exhibiting drug resistances at such hubs. A novel comparative signaling-motif for malignant signaling-regulation is also proposed, which sums up these findings. There have been elaborated studies on network motifs (Alon, 2007). The comparative cancer motif is different from these motifs in that it shows signaling-regulation in cancer reflecting less centralized formation. It is to note that the comparative cancer motif agrees with the findings of non-integration (motif A, Figure 9) but signal-maintenance (motif B, Figure 9) of proteins with higher involvement in signal propagation.

In conclusion, a method that based on the correlation between interacting genes was used, which is simple and enabled tracking basic principles of signaling by its regulation. The malignant signaling networks showed more diverse signaling pathways which were shorter and used less hubs. They indicated signaling maintenance and increased error tolerance to punctual attacks even at hubs which makes cancer treatment at specific targets challenging.

To detect common signalling patterns 10 cancer datasets were chosen from the same Affymetrix platform to avoid cross platform differences which could affect the analysis. In all the datasets chosen there were sufficient number of normal and cancer patients enabling performance of a comparative study. In Table 2, a common pattern can be seen across datasets for the different measures. Some of the datasets comprise of normal and cancer tissue samples derived from different groups of patients. The method is unaffected by this, however in some cases these distant expression patterns between the groups may give inconsistent results. Hence homogenous sets of samples were selected in these cases by a cluster analysis.

In the second part, the Neuroblastoma gene expression dataset was analyzed in detail. The analysis determined nodes which are used in high frequency in the predicted signalling pathways, predicted by combining the gene expression data and protein-protein interaction network information. Many of the genes listed in top of the list have been reported to be involved in neuroblastoma progression and in cancer in general. Proteins with high

frequency in stage 4A and low frequency with stage 1 were indentified. Those up-regulated genes can be knocked down in *MYCN* amplified cell lines to determine their effect upon cell survival.

Most of the top genes obtained in the analysis include kinases which are active players in the signalling machinery in cancer. It is clear from the list that some of the top genes MAPK1, PRKCA and PRKACA have a role in regulating apoptosis whereas CAV1 and PAK1 are involved in neurite outgrowth. The respective up-regulation or down-regulation of the genes mostly correlated well with the reported findings of their role in neuroblastoma cell lines. For instance, MAPK1 was down-regulated in stage4A, but it's role is to form a complex with GSK3-Beta and induce apoptosis. On the other hand, PRKCA which was up-regulated in stage 4A inhibits GSK3-Beta thereby inhibiting apoptosis. PRKACA inhibits Bcl-2 thereby preventing apoptosis and it was up-regulated in stage 4A. These genes in the top of the list seem to play an important role in apoptosis. They are all interlinked to form a small network. This could be small sub-network of the larger network of signalling interactions involved in apoptosis. These were obtained by linking the most likely oncogenic candidates found by our method.

Apart from genes implicated in apoptosis, the method is able to identify genes associated with symptoms of aggressive neuroblastomas. PRKACA is invoved in neuronal differentiation and CAV1 and PAK1 are involved in neurite outgrowth. CAV1 is down-regulated in stage4A and inhibits PAK1 which is up-regulated in stage4A and PAK1 induces in neurite outgrowth. It is also reported in literature that both CAV1 and PAK1 form components of the same pathway. In conclusion, the method used in this study enables detection of highly involved signalling proteins and functionally related signalling proteins involved in specific oncogenetic programs.

Outlook

A network based approach was used to dissecting properties of signalling networks and to identifying central signalling genes in cancer. The method used is based on the correlation between interacting genes which is simple and enabled tracking basic principles of signaling by its regulation. The malignant signaling networks showed more diverse signaling pathways which were shorter and used less hubs. They indicated signaling maintenance and increased error tolerance to punctual attacks even at hubs which makes cancer treatment at specific targets challenging.

This study shows the difference in the regulatory dynamics of signalling regulation between the normal and the cancer network. The results across 10 datasets are consistent with the proposed model. This knowledge can be used for designing effective strategies for targeted drug discovery for cancer. However, precautions should be taken to select homogenous gene expression data (grouped by age, sex, cancer stage etc.) in particular cancer types to avoid unnecessary fluctuations that could be causative to these differences.

This method can be further extended to identify important signalling in cancer. In this thesis, this was done by identifying genes with highly differential node frequency. Several active signalling proteins in neuroblastoma were identified. These proteins could be considered as potential drug targets. Further, knock-down studies can be done in the laboratory with cell lines to confirm their importance as drug targets. Further, this method can be used to analyze gene expression datasets pertaining to specific cancer subtypes, to identify subtype specific, as well as common players enabling us to move a step further to understanding the complex cancer signalling network.

In future, the analysis of many large scale gene expression datasets of different cancer types can be used to look for key molecules of signal regulation. Such analysis will provide key signaling molecules which could fill the missing links in the currently available sparse cancer signaling network.

References

Albert, R. (2005) Scale-free networks in cell biology, *J Cell Sci*, **118**, 4947-4957.

Albert, R., Jeong, H. and Barabasi, A.L. (2000) Error and attack tolerance of complex networks, *Nature*, **406**, 378-382.

Alon, U. (2007) Network motifs: theory and experimental approaches, *Nat Rev Genet*, **8**, 450-461.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool, *J Mol Biol*, **215**, 403-410.

Barabasi, A.L. and Albert, R. (1999) Emergence of scaling in random networks, *Science (New York, N.Y*, **286**, 509-512.

Barabasi, A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization, *Nat Rev Genet*, **5**, 101-113.

Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M. and Edgar, R. (2007) NCBI GEO: mining tens of millions of expression profiles--database and tools update, *Nucleic acids research*, **35**, D760-765.

Bhattacharjee, A., Richards, W.G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E.J., Lander, E.S., Wong, W., Johnson, B.E., Golub, T.R., Sugarbaker, D.J. and Meyerson, M. (2001) Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses, *Proc Natl Acad Sci U S A*, **98**, 13790-13795.

Bielke, W., Ziemieki, A., Kappos, L. and Miescher, G.C. (1992) Expression of the B cell-associated tyrosine kinase gene Lyn in primary neuroblastoma tumours and its modulation during the differentiation of neuroblastoma cell lines, *Biochem Biophys Res Commun*, **186**, 1403-1409.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. and Vingron, M. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data, *Nature genetics*, **29**, 365-371.

Breitling, R., Armengaud, P., Amtmann, A. and Herzyk, P. (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments, *FEBS Lett*, **573**, 83-92.

Cedazo-Minguez, A., Popescu, B.O., Blanco-Millan, J.M., Akterin, S., Pei, J.J., Winblad, B. and Cowburn, R.F. (2003) Apolipoprotein E and beta-amyloid (1-42) regulation of glycogen synthase kinase-3beta, *J Neurochem*, **87**, 1152-1164.

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

Cormen, T.H., Leiserson, C.E. and Rivest, R.L. (1990) *Introduction to algorithms*. MIT Press ;McGraw-Hill, Cambridge, Mass.New York.

Cui, Q., Ma, Y., Jaramillo, M., Bari, H., Awan, A., Yang, S., Zhang, S., Liu, L., Lu, M., O'Connor-McCourt, M., Purisima, E.O. and Wang, E. (2007) A map of human cancer signaling, *Mol Syst Biol*, **3**, 152.

di Bernardo, D., Thompson, M.J., Gardner, T.S., Chobot, S.E., Eastwood, E.L., Wojtovich, A.P., Elliott, S.J., Schaus, S.E. and Collins, J.J. (2005) Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks, *Nat Biotechnol*, **23**, 377-383.

Druker, B.J. (2008) Translation of the Philadelphia chromosome into therapy for CML, *Blood*, **112**, 4808-4817.

Edgar, R., Domrachev, M. and Lash, A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, *Nucleic acids research*, **30**, 207-210.

Ergun, A., Lawrence, C.A., Kohanski, M.A., Brennan, T.A. and Collins, J.J. (2007) A network biology approach to prostate cancer, *Mol Syst Biol*, **3**, 82.

Estilo, C.L., P, O.c., Talbot, S., Socci, N.D., Carlson, D.L., Ghossein, R., Williams, T., Yonekawa, Y., Ramanathan, Y., Boyle, J.O., Kraus, D.H., Patel, S., Shaha, A.R., Wong, R.J., Huryn, J.M., Shah, J.P. and Singh, B. (2009) Oral tongue cancer gene expression profiling: Identification of novel potential prognosticators by oligonucleotide microarray analysis, *BMC Cancer*, **9**, 11.

Fan, C., Oh, D.S., Wessels, L., Weigelt, B., Nuyten, D.S., Nobel, A.B., van't Veer, L.J. and Perou, C.M. (2006) Concordance among gene-expression-based predictors for breast cancer, *The New England journal of medicine*, **355**, 560-569.

Goymer, P. (2008) Natural selection: The evolution of cancer, *Nature*, **454**, 1046-1048.

Guelzim, N., Bottani, S., Bourgine, P. and Kepes, F. (2002) Topological and causal structure of the yeast transcriptional regulatory network, *Nature genetics*, **31**, 60-63.

Hanahan, D. and Weinberg, R.A. (2000) The hallmarks of cancer, *Cell*, **100**, 57-70.

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

Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. and Vingron, M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression, *Bioinformatics*, **18 Suppl 1**, S96-104.

Ihmels, J., Levy, R. and Barkai, N. (2004) Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae, *Nat Biotechnol*, **22**, 86-92.

Itano, Y., Ito, A., Uehara, T. and Nomura, Y. (1996) Regulation of Bcl-2 protein expression in human neuroblastoma SH-SY5Y cells: positive and negative effects of protein kinases C and A, respectively, *J Neurochem*, **67**, 131-137.

Jeong, H., Mason, S.P., Barabasi, A.L. and Oltvai, Z.N. (2001) Lethality and centrality in protein networks, *Nature*, **411**, 41-42.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. and Barabasi, A.L. (2000) The large-scale organization of metabolic networks, *Nature*, **407**, 651-654.

Jones, J., Otu, H., Spentzos, D., Kolia, S., Inan, M., Beecken, W.D., Fellbaum, C., Gu, X., Joseph, M., Pantuck, A.J., Jonas, D. and Libermann, T.A. (2005) Gene signatures of progression and metastasis in renal cell cancer, *Clin Cancer Res*, **11**, 5730-5739.

Kang, M.J., Seo, J.S. and Park, W.Y. (2006) Caveolin-1 inhibits neurite growth by blocking Rac1/Cdc42 and p21-activated kinase 1 interactions, *Neuroreport*, **17**, 823-827.

Keseler, I.M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I.T., Peralta-Gil, M. and Karp, P.D. (2005) EcoCyc: a comprehensive database resource for Escherichia coli, *Nucleic acids research*, **33**, D334-337.

Kharchenko, P., Church, G.M. and Vitkup, D. (2005) Expression dynamics of a cellular metabolic network, *Molecular systems biology*, **1**, 2005 0016.

Konig, R., Schramm, G., Oswald, M., Seitz, H., Sager, S., Zapatka, M., Reinelt, G. and Eils, R. (2006) Discovering functional gene expression patterns in the metabolic network of Escherichia coli with wavelets transforms, *BMC Bioinformatics*, **7**, 119.

Kuriakose, M.A., Chen, W.T., He, Z.M., Sikora, A.G., Zhang, P., Zhang, Z.Y., Qiu, W.L., Hsu, D.F., McMunn-Coffran, C., Brown, S.M., Elango, E.M., Delacure, M.D. and Chen, F.A. (2004) Selection and validation of differentially expressed genes in head and neck cancer, *Cell Mol Life Sci*, **61**, 1372-1383.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.B., Volkert, T.L., Fraenkel, E., Gifford, D.K. and Young, R.A. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae, *Science (New York, N.Y.*, **298**, 799-804.

Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., Teichmann, S.A. and Gerstein, M. (2004) Genomic analysis of regulatory network dynamics reveals large topological changes, *Nature*, **431**, 308-312.

Ma'ayan, A., Jenkins, S.L., Neves, S., Hasseldine, A., Grace, E., Dubin-Thaler, B., Eungdamrong, N.J., Weng, G., Ram, P.T., Rice, J.J., Kershenbaum, A., Stolovitzky, G.A., Blitzer, R.D. and Iyengar, R. (2005) Formation of regulatory patterns during signal propagation in a Mammalian cellular network, *Science*, **309**, 1078-1083.

Ma, C., Bower, K.A., Chen, G., Shi, X., Ke, Z.J. and Luo, J. (2008) Interaction between ERK and GSK3beta mediates basic fibroblast growth factor-induced apoptosis in SK-N-MC neuroblastoma cells, *The Journal of biological chemistry*, **283**, 9248-9256.

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., Kloos, D.U., Land, S., Lewicki-Potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S. and Wingender, E. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles, *Nucleic Acids Res*, **31**, 374-378.

Mishra, G.R., Suresh, M., Kumaran, K., Kannabiran, N., Suresh, S., Bala, P., Shivakumar, K., Anuradha, N., Reddy, R., Raghavan, T.M., Menon, S., Hanumanthu, G., Gupta, M., Upendran, S., Gupta, S., Mahesh, M., Jacob, B., Mathew, P., Chatterjee, P., Arun, K.S., Sharma, S., Chandrika, K.N., Deshpande, N., Palvankar, K., Raghavnath, R., Krishnakanth, R., Karathia, H., Rekha, B., Nayak, R., Vishnupriya, G., Kumar, H.G., Nagini, M., Kumar, G.S., Jose, R., Deepthi, P., Mohan, S.S., Gandhi, T.K., Harsha, H.C., Deshpande, K.S., Sarker, M., Prasad, T.S. and Pandey, A. (2006) Human protein reference database--2006 update, *Nucleic Acids Res*, **34**, D411-414.

Ness, S.A. (2006) Basic microarray analysis: strategies for successful experiments, *Methods Mol Biol*, **316**, 13-33.

O'Driscoll, K.R., Teng, K.K., Fabbro, D., Greene, L.A. and Weinstein, I.B. (1995) Selective translocation of protein kinase C-delta in PC12 cells during nerve growth factor-induced neuritogenesis, *Mol Biol Cell*, **6**, 449-458.

Oberthuer, A., Berthold, F., Warnat, P., Hero, B., Kahlert, Y., Spitz, R., Ernestus, K., Konig, R., Haas, S., Eils, R., Schwab, M., Brors, B., Westermann, F. and Fischer, M. (2006) Customized oligonucleotide microarray gene expression-based classification of neuroblastoma patients outperforms current clinical risk stratification, *J Clin Oncol*, **24**, 5070-5078.

Oberthuer, A., Berthold, F., Warnat, P., Hero, B., Kahlert, Y., Spitz, R., Ernestus, K., König, R., Haas, S., Eils, R., Schwab, M., Brors, B., Westermann, F. and Fischer, M. (2006) Customized oligonucleotide microarray gene expression-based classification of neuroblastoma patients outperforms current clinical risk stratification, *J Clin Oncol*, **24**, 5070-5078.

Papin, J.A., Hunter, T., Palsson, B.O. and Subramaniam, S. (2005) Reconstruction of cellular signalling networks and analysis of their properties, *Nat Rev Mol Cell Biol*, **6**, 99-111.

Park, T., Yi, S.G., Kang, S.H., Lee, S., Lee, Y.S. and Simon, R. (2003) Evaluation of normalization methods for microarray data, *BMC bioinformatics*, **4**, 33.

Parkinson, H., Kapushesky, M., Kolesnikov, N., Rustici, G., Shojatalab, M., Abeygunawardena, N., Berube, H., Dylag, M., Emam, I., Farne, A., Holloway, E., Lukk, M., Malone, J., Mani, R., Pilicheva, E., Rayner, T.F., Rezwan, F., Sharma, A., Williams, E., Bradley, X.Z., Adamusiak, T., Brandizi, M., Burdett, T., Coulson, R., Krestyaninova, M., Kurnosov, P., Maguire, E., Neogi, S.G., Rocca-Serra, P., Sansone, S.A., Sklyar, N., Zhao, M., Sarkans, U. and Brazma, A. (2009) ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression, *Nucleic acids research*, **37**, D868-872.

Peri, S., Navarro, J.D., Amanchy, R., Kristiansen, T.Z., Jonnalagadda, C.K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T.K., Gronborg, M., Ibarrola, N., Deshpande, N., Shanker, K., Shivashankar, H.N., Rashmi, B.P., Ramya, M.A., Zhao, Z., Chandrika, K.N., Padma, N., Harsha, H.C., Yatish, A.J., Kavitha, M.P., Menezes, M., Choudhury, D.R., Suresh, S., Ghosh, N., Saravana, R., Chandran, S., Krishna, S., Joy, M., Anand, S.K., Madavan, V., Joseph, A., Wong, G.W., Schiemann, W.P., Constantinescu, S.N., Huang, L., Khosravi-Far, R., Steen, H., Tewari, M., Ghaffari, S., Blobe, G.C., Dang, C.V., Garcia, J.G., Pevsner, J., Jensen, O.N., Roepstorff, P., Deshpande, K.S., Chinnaiyan, A.M., Hamosh, A., Chakravarti, A. and Pandey, A. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans, *Genome Res*, **13**, 2363-2371.

Przulj, N., Wigle, D.A. and Jurisica, I. (2004) Functional topology in a network of protein interactions, *Bioinformatics*, **20**, 340-348.

Pyeon, D., Newton, M.A., Lambert, P.F., den Boon, J.A., Sengupta, S., Marsit, C.J., Woodworth, C.D., Connor, J.P., Haugen, T.H., Smith, E.M., Kelsey, K.T., Turek, L.P. and Ahlquist, P. (2007) Fundamental differences in cell cycle deregulation in human papillomavirus-positive and human papillomavirus-negative head/neck and cervical cancers, *Cancer Res*, **67**, 4605-4619.

Quackenbush, J. (2002) Microarray data normalization and transformation, *Nature genetics*, **32 Suppl**, 496-501.

Rhodes, D.R., Yu, J., Shanker, K., Deshpande, N., Varambally, R., Ghosh, D., Barrette, T., Pandey, A. and Chinnaiyan, A.M. (2004) ONCOMINE: a cancer microarray database and integrated data-mining platform, *Neoplasia*, **6**, 1-6.

Robinson, P., Lebel, M. and Cyr, M. (2008) Dopamine D1 receptor-mediated aggregation of N-terminal fragments of mutant huntingtin and cell death in a neuroblastoma cell line, *Neuroscience*, **153**, 762-772.

Sanchez, S., Jimenez, C., Carrera, A.C., Diaz-Nido, J., Avila, J. and Wandosell, F. (2004) A cAMP-activated pathway, including PKA and PI3K, regulates neuronal differentiation, *Neurochem Int*, **44**, 231-242.

Santegoets, L.A., Seters, M., Helmerhorst, T.J., Heijmans-Antonissen, C., Hanifi-Moghaddam, P., Ewing, P.C., van Ijcken, W.F., van der Spek, P.J., van der Meijden, W.I. and Blok, L.J. (2007) HPV related VIN: highly proliferative and diminished responsiveness to extracellular signals, *Int J Cancer*, **121**, 759-766.

Sato, Y., Sasaki, H., Kondo, S., Fukai, I., Kiriyama, M., Yamakawa, Y. and Fujii, Y. (2001) Expression of the cdc25B mRNA correlated with that of N-myc in neuroblastoma, *Jpn J Clin Oncol*, **31**, 428-431.

Schmid, R.S., Midkiff, B.R., Kedar, V.P. and Maness, P.F. (2004) Adhesion molecule L1 stimulates neuronal migration through Vav2-Pak1 signaling, *Neuroreport*, **15**, 2791-2794.

Schramm, G., Zapatka, M., Eils, R. and Konig, R. (2007) Using gene expression data and network topology to detect substantial pathways, clusters and switches during oxygen deprivation of Escherichia coli, *BMC bioinformatics*, **8**, 149.

Schwab, M. (2004) MYCN in neuronal tumours, *Cancer Lett*, **204**, 179-187.

Schwab, M., Westermann, F., Hero, B. and Berthold, F. (2003) Neuroblastoma: biology and molecular and chromosomal pathology, *Lancet Oncol*, **4**, 472-480.

Shannon, C. (1948) A Mathematical Theory of Communication, *The Bell System Technical Journal*, **27**, 379-423, 623-656.

Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P., Lander, E.S., Loda, M., Kantoff, P.W., Golub, T.R. and Sellers, W.R. (2002) Gene expression correlates of clinical prostate cancer behavior, *Cancer Cell*, **1**, 203-209.

Stirewalt, D.L., Meshinchi, S., Kopecky, K.J., Fan, W., Pogosova-Agadjanyan, E.L., Engel, J.H., Cronk, M.R., Dorcy, K.S., McQuary, A.R., Hockenbery, D., Wood, B., Heimfeld, S. and Radich, J.P. (2008) Identification of genes with abnormal expression changes in acute myeloid leukemia, *Genes, chromosomes & cancer*, **47**, 8-20.

Strogatz, S.H. (2001) Exploring complex networks, *Nature*, **410**, 268-276.

Su, L.J., Chang, C.W., Wu, Y.C., Chen, K.C., Lin, C.J., Liang, S.C., Lin, C.H., Whang-Peng, J., Hsu, S.L., Chen, C.H. and Huang, C.Y. (2007) Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap re-sampling scheme, *BMC Genomics*, **8**, 140.

van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S.,

Bernards, R. and Friend, S.H. (2002) Gene expression profiling predicts clinical outcome of breast cancer, *Nature*, **415**, 530-536.

van de Vijver, M.J., He, Y.D., van't Veer, L.J., Dai, H., Hart, A.A., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., Roberts, C., Marton, M.J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E.T., Friend, S.H. and Bernards, R. (2002) A gene-expression signature as a predictor of survival in breast cancer, *N Engl J Med*, **347**, 1999-2009.

Vogelstein, B. and Kinzler, K.W. (2004) Cancer genes and the pathways they control, *Nat Med*, **10**, 789-799.

Wang, Y., Klijn, J.G., Zhang, Y., Sieuwerts, A.M., Look, M.P., Yang, F., Talantov, D., Timmermans, M., Meijer-van Gelder, M.E., Yu, J., Jatkoe, T., Berns, E.M., Atkins, D. and Foekens, J.A. (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer, *Lancet*, **365**, 671-679.

Wei, C., Li, J. and Bumgarner, R.E. (2004) Sample size for detecting differentially expressed genes in microarray experiments, *BMC genomics*, **5**, 87.

Weinberg, R.A. (2007) *The biology of cancer*. Garland Science, New York.

Westermann, F. and Schwab, M. (2002) Genetic parameters of neuroblastomas, *Cancer Lett*, **184**, 127-147.

Xia, Y., Yu, H., Jansen, R., Seringhaus, M., Baxter, S., Greenbaum, D., Zhao, H. and Gerstein, M. (2004) Analyzing cellular biochemistry in terms of molecular networks, *Annu Rev Biochem*, **73**, 1051-1087.

Yeger-Lotem, E., Sattath, S., Kashtan, N., Itzkovitz, S., Milo, R., Pinter, R.Y., Alon, U. and Margalit, H. (2004) Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction, *Proc Natl Acad Sci U S A*, **101**, 5934-5939.

Supplement

Figure 11. Distribution of the correlation coefficients of the different cancers (black bars: normal, red bars: cancer)

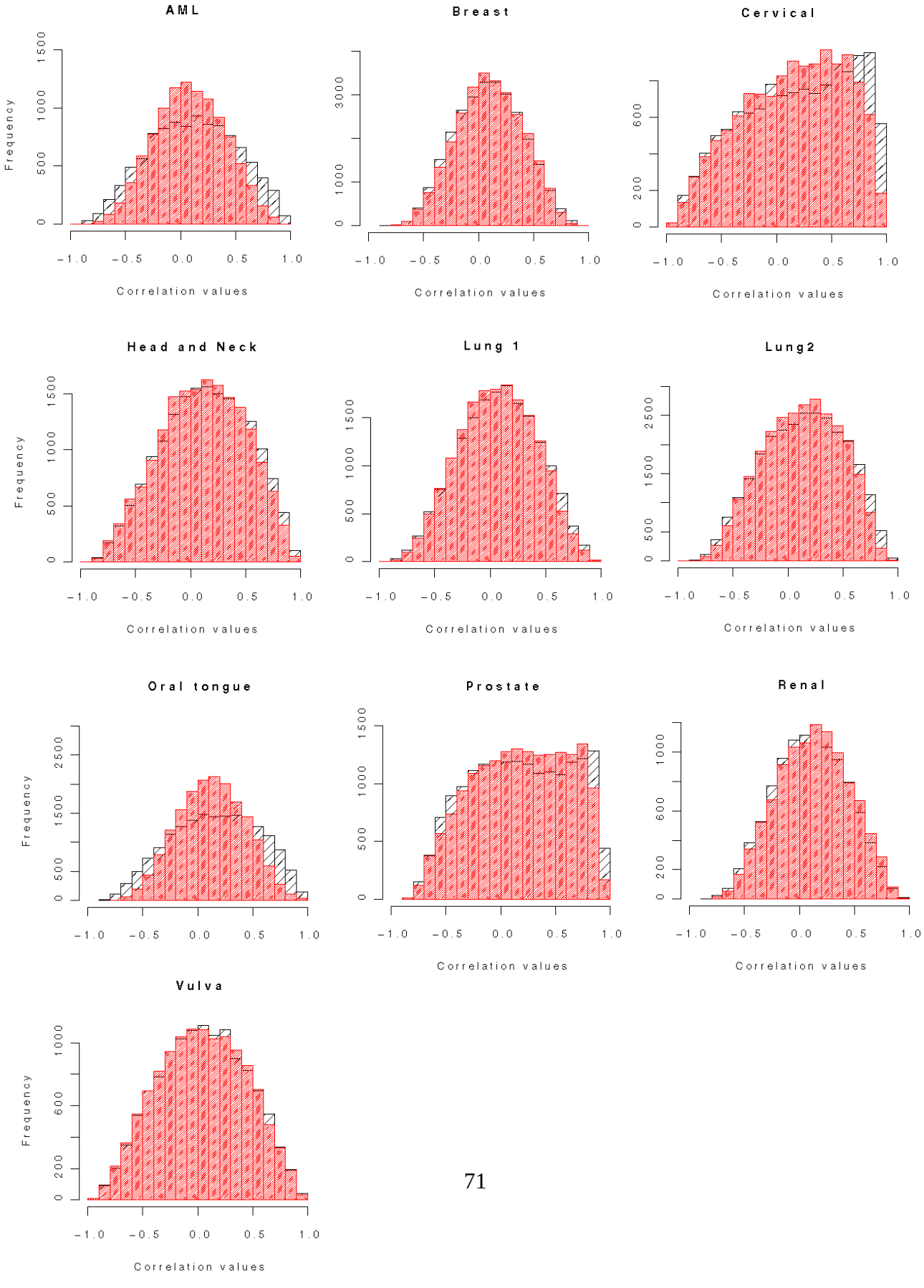


Figure 12. Link frequency distribution for all datasets

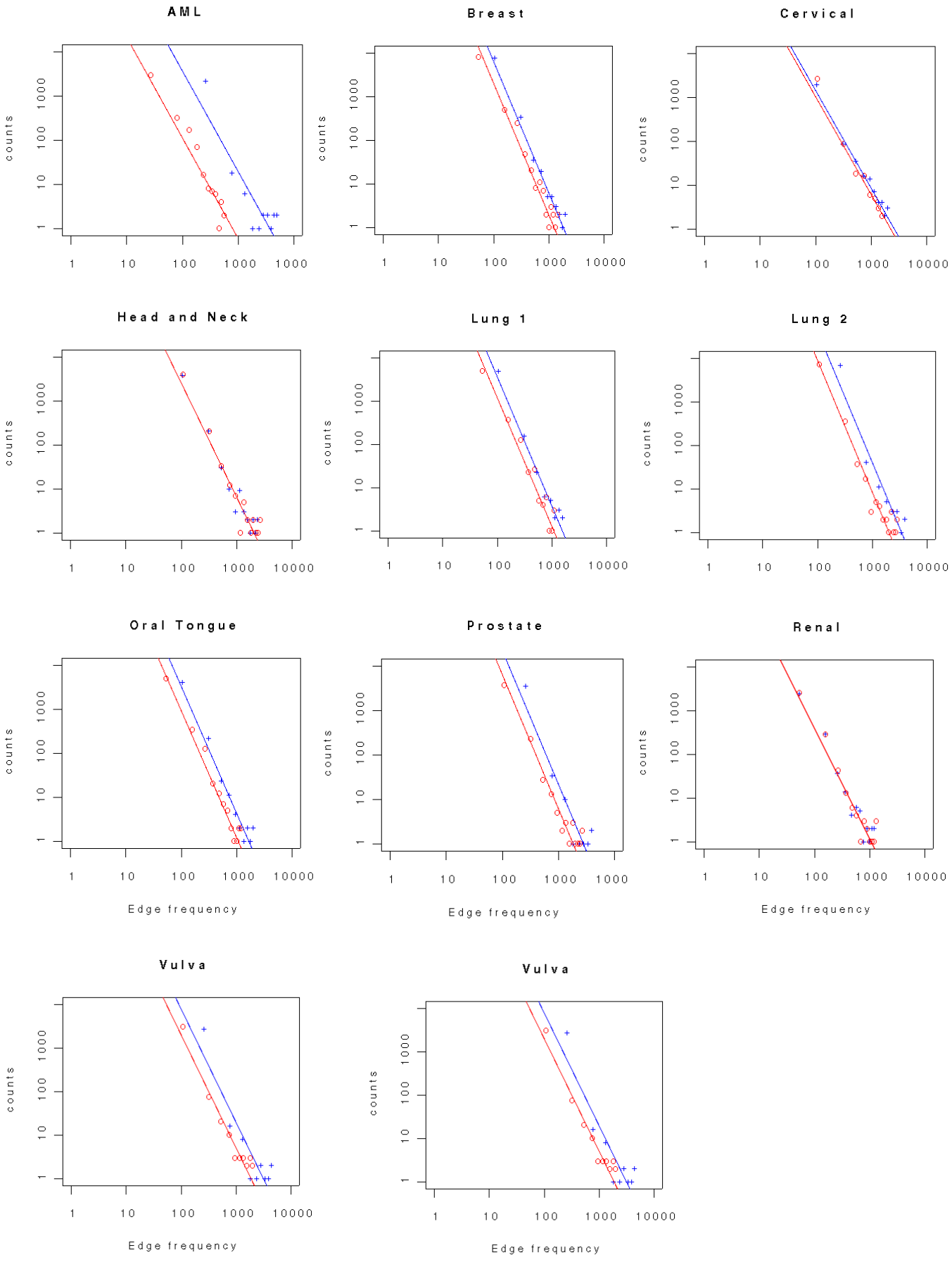


Figure 12. In the above figure, Link frequency distribution for all datasets are shown (normal: blue, crosses, cancer: red, circles). All networks show the typical scale-free distribution for the frequency of the genes to be involved in the defined signalling pathways. Genes in the cancer network exhibit a distinct shift to the left indicating less frequency not only for the hubs but for all genes in the network. All distributions were fitted by a combined linear model (see methods).

y

Acknowledgements

Several people helped me in making this thesis possible. I would especially like to thank Prof. Dr. Roland Eils who provided me with the opportunity to work in this division. He was a constant source of support and encouragement. I am also very grateful to my second supervisor Prof. Dr. Manfred Schwab for his support and guidance. I sincerely thank Dr. Rainer König for fruitful discussions I had with him, encouragement and able guidance in this thesis work. I also thank the Hemlholtz International Graduate school for Cancer Research for providing me a fellowship for my study at DKFZ.

Furthermore, I enjoyed working in the network modeling group with Dr. Gunnar Schramm, Tobias Bauer, Anna-Lena Kranz, Kitiporn Plaimas, Apichat Suratanee and Richa Batra. I experienced a friendly atmosphere in group and a great team spirit.

I would also like to thank Dr. Frank Westermann of the Tumour genetics division of the DKFZ, Dr. Benedikt Brors from my division and Dr. Marc Zapatka for their valuable suggestion during my thesis work.

I would like to thank my parents and my sister who provided me support during my thesis work.

Erklärung

Ich versichere, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Heidelberg, den Januar 14, 2010 ……

(Kannabiran Nandakumar)