1-1-2017

Network-Based Approaches To Identify The Impacted Genes And Active Interactions

Sahar Ansari
Wayne State University,

NETWORK-BASED APPROACHES TO IDENTIFY THE IMPACTED GENES AND ACTIVE INTERACTIONS

by

SAHAR ANSARI

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the

requirements for the degree of

DOCTOR OF PHILOSOPHY

2017

MAJOR: COMPUTER SCIENCE

Approved By:

_____

Advisor Date

_____

_____

_____

_____

DEDICATION

To my parents, Robab and Dariush, for their kindness, love, and prayers. I owe all my achievements to them.

To my life partner, Calin, for his constant care, support, and patience. This wouldn't be possible without him.

ACKNOWLEDGEMENTS

First, I would like to show my gratitude to my advisor, Dr. Sorin Draghici, who inspired me during the last few years to be a perfectionist not only in research, but also in other aspects of life. He taught me how to begin a research project by looking at important problems that has high impact in real life, and find the best possible solution for them. I thank him for his insights, patience, and comments during this Ph.D. journey.

Second, I would like to thank my family, Dariush, Robab, Calin, Elnaz, Armin, and Samin, who helped me through these years with their support, and guidance. This would not be possible without their love, good thoughts and encouragement.

I thank the committee members, Dr. Fotouhi, Dr. Kotov, and Dr. Hao, for their helpful feedbacks. I strongly appreciate their time to review this thesis.

Next, I would like to thank all the past and current members in ISBL team, who made a very friendly environment in the lab. I appreciate their friendships and motivations. Specially, I like to thank Dr. Michele Donato, Cristina Mitrea, Nafiseh Saberian, and Azam Peyvandipour for their great collaborations and helpful comments.

I extend my thanks to all people who were involved in the process.

This research was supported in part by the following grants: NIH R01 DK089167, R42 GM087013 and NSF DBI-0965741, and by the Robert J. Sokol Endowment in Systems Biology in Reproduction. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the funding agencies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

The ultimate goal of any biological experiment is to understand the underlying phenomenon of the condition investigated. Technologies such as microarray [102], and RNA-seq [151] make it possible to capture the expression level of thousands of genes at the same time. Understanding how genes interact with each other is the key to understand how the cell works and as a result, how diseases evolve. At the same time, this knowledge can be used to design better drugs and improve the standard of medical care. Previously, single-gene statistical analyses were used to identify the genes that are responsible for a given condition. This type of analysis is not able to provide a complete understanding of the disease. By understanding how genes interact, we can first construct networks that describe specific mechanisms and later combine them at a system level to provide a global perspective.

A very important step in system biology, is the identification of the networks that are most impacted in the given condition. These networks, called gene regulatory networks or pathways, are modeled as graphs where the nodes represent genes and the edges represent the interactions between them. Such networks explain where the target genes are affected by some other genes, and therefore describe the mechanisms involved in a biological process.

Genes that have similar behavior (e.g., have high correlation) in different conditions are considered to be involved in similar cellular functions. These genes may share a transcription factor. We need to answer the question of which genes cause the impact in a specific condition and how they interact [52].

The networks that explain the interactions between genes can be used to: 1) predict the disease or the responses of the system to a specific impact, 2) find the subset of genes that interact with each other and play an important role in the condition of interest, and 3) understand the mechanisms involved in that condition [7, 29, 83].

We are addressing above mentioned problem by introducing two main strategies. First, we are taking advantage of pre-defined pathways obtained from existing databases to identify the impact of a phenotype studied on such pathways.

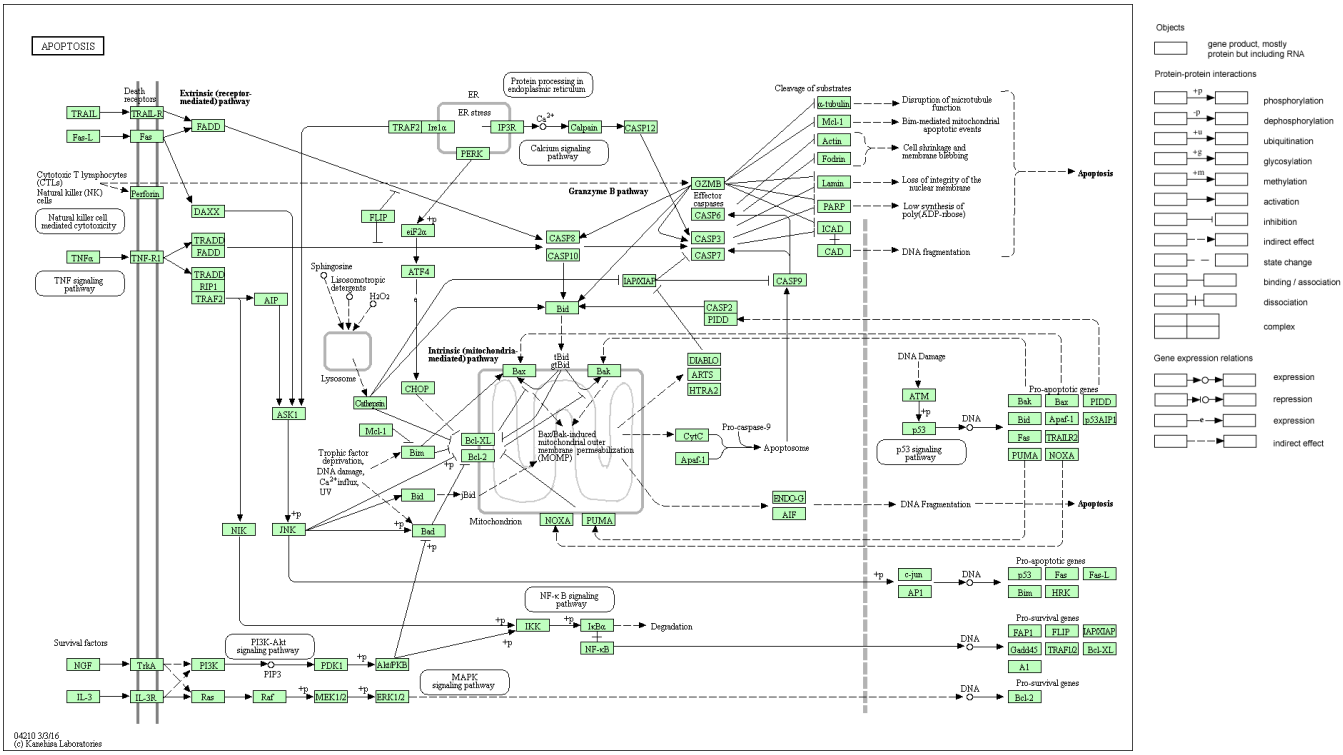


Figure 2.1: KEGG apoptosis signaling pathway (from: `http://www.genome.jp/kegg/pathway/hsa/hsa04210.html`).

A pathway describes all the known phenomena involved in a given biological process, to which it is associated. It is part of a larger system that has a set of components interacting with each other. These components work together to achieve a common goal. The name of a pathway usually represents the biological process, phenomenon, or disease process described by the pathway. Different types of interactions are described by different types of edges, or weights in the structure of a pathway. As an example, Figure 2.1 shows the apoptosis signaling pathway, which includes the genes and interactions involved in the known mechanism for cell death. Different types of interactions are shown by different arrows.

In this research, we are introducing a pathway analysis method that compares two phenotypes and helps identifying the pathways that are significantly impacted between the two phenotypes. The ranks of the pathways represent the significance of their perturbation in one phenotype versus another. Correctly identifying the pathways that are impacted in a disease condition is a crucial step in understanding that disease. This method aims to reduce

the false positives and false negatives in the existing methods by focusing on the primary dis-regulation of the gene itself in comparison with the effects coming from upstream genes.

A common avenue pursued in order to understand the differences between phenotypes involves the selection of a subset of "differentially expressed" (DE) genes. Usually, the tens of thousands of measured genes are reduced to a small set of a few hundred DE genes, essentially discarding about 99% of the measurements. Our hypothesis is that the drastic filtering necessary in order to select this subset of DE genes could discard many genes that play important roles in the given phenotype. The pathway analysis method proposed here takes full advantage of the current technologies and uses the measured data for the entire set of 30,000-100,000 transcripts in the genome, thus allowing the full use of the data provided by current techniques. The proposed method is validated on 24 datasets involving 12 different human disease, as well as 8 yeast knock-out datasets. It yields significant improvements with respect to the state-of-the-art methods.

Second, we are proposing a method that identifies a subset of genes and interactions relevant to the given phenotype. The identified subset known as the active network includes the genes that are strongly involved in the phenotype studied. One of the drawbacks of the existing curated pathways or gene interactions databases is that each interaction is extracted from literature or experimentally validated by independent studies. However, most such interactions were found in specific tissues and/or phenotype, and not all studies employed the same tissue and/or phenotype. Therefore, these independently identified interactions in the databases may not exist in the actual phenotype or tissue studied in a subsequent experiment. Furthermore, new phenomena may be involved in the tissue or phenotype currently being studied. Utilizing only existing pathways from pathways databases or literature, limits one's ability to discover new phenomena and new interactions.

In order to overcome such limitations, some existing methods try to build the regulatory networks based on the correlation or the co-expression existing in the given datasets [55, 71, 107, 157]. The networks resulting from such methods are specific to the condition under

study, but the interactions identified are only based on the genes' expression level. This limitation can produce many false positives as well as false negatives, because an interaction between two genes is not necessarily reflected in the correlation between their expression levels. The interactions between genes can involve an indirect relation between them via their protein products or their transcription factors, and sometimes interactions take place on different time scales. Therefore, there is a need for computational algorithms able to construct network of active interactions by analyzing data in more sophisticated ways, by combining gene expression data with existing pathway information, as well as with data from protein-protein interactions databases. Such methods identify the network of interactions that is most relevant to a given phenotype based on the retrieved prior knowledge, referred to as "active network". This network is also known as "network hotspot" or "responsive subnetwork" [86].

The active network, as part of a global interaction network, explains the sudden changes in the genes activity or the characteristic of the phenotype in a given disease. This network is identified based on the given data and can be considered as the putative mechanism involved in the given phenotype. The advantage of identifying an active network is that it is specific to the condition studied, as opposed to existing curated pathways that can describe more generic knowledge, not necessarily applicable to the given condition.

In this manuscript, we propose a method that uses interactions obtained from different databases to infer the active networks. Also, we propose two validation approaches to evaluate the results. First approach is computing a pathway enrichment score to assess the significance of the overlaps between identified network and known pathways. The second approach is computing the percentage of the genes in the identified network that are known to be relevant to the condition based on an existing database. We validated the method on multiple datasets from experiments studying colorectal cancer, renal cancer and prostate cancer. We compare our result with the results of two widely used methods: NetWalker and HotNet, and the classical approach of considering just the union of differentially expressed

(DE). The results of these comparisons show that the proposed method yields improvements in constructing networks, which are significantly relevant to the given diseases based on the resulting genes and interactions.

CHAPTER 2: PRIMARY DIS-REGULATION

2.1 Background

The goal of pathway analysis methods is to identify the most perturbed pathways in a given condition. Pathways are divided in two main categories: i) signaling pathways, that are defined as graphs in which nodes represent genes/proteins and edges are interactions between them, and ii) metabolic pathways in which the nodes represent biochemical compounds and the edges represent reactions, carried out by enzymes which are coded by genes [87]. Such pathways describe all known phenomena involved in a biological process (e.g. cell cycle), or a disease (e.g. Alzheimer's disease), etc. In this thesis, we focus on signaling pathways to be able to map the measured expression level of the genes to the corresponding nodes in those pathways. Intuitively, the impact of a given phenotype on a given pathway should be determined by the number of differentially expressed (DE) genes on that pathway, the magnitude of the changes in the expression level of the genes, and the type, direction and strength of the interactions between the genes in that pathway.

The simplest pathway analysis approach is the over-representation analysis (ORA) [62]. This approach considers only the number of DE genes that are present in a given pathway. ORA techniques calculate the probability of finding a certain number of DE genes among all the genes in a pathway just by chance. Another approach to pathway analysis is the functional class scoring (FCS) [63, 87]. This approach takes into consideration all measured expression changes, as well as the correlation between the expression change of the genes and the phenotype. The most popular techniques in the FCS category are Gene Set Enrichment Analysis (GSEA) [121] and Gene Set Analysis (GSA) [35]. These two techniques rank the genes based on the correlation between their expression and a given phenotype, and calculate a score that reflects the degree to which a given pathway is represented at extremes of the ranked list. Neither of these two approaches considers the interactions between genes, their direction, type, strength, etc. In essence, all these methods treat the pathways as simple sets of genes.

However, databases such as KEGG [93], BioCarta [16] and Reactome [56] provide pathways that consist of much more than just sets of genes. These databases provide complex graphs for each signaling pathways in which each node is a gene/protein and each edge is an interaction between two such genes or proteins. Ignoring the wealth of knowledge captured in the topology of the pathway is clearly sub-optimal. Even though these databases provide more detailed information about the topology of the pathways, there are thousands of genes that have not been annotated yet. Furthermore, many of the existing annotations may be inaccurate [63]. However, we believe that accuracy and reliability of pathways annotation is growing and using this type of information can only help the interpretation of high-throughput experiments.

More recently, more sophisticated methods have been proposed that are able to fully take into consideration all the interactions between genes in signaling pathways to find which pathway is most impacted by a given phenotype [32]. These are sometimes referred to as "topology-aware" or "third generation" pathway analysis methods [63, 87]. The method, proposed in this manuscript, belongs to this latest generation of pathway analysis methods, inasmuch it considers the topology of the pathways, as well as the changes in expression level of the genes.

However, even the most sophisticated current pathway analysis methods still produce both false positives as well as false negatives in certain circumstances. We hypothesized that such incorrect results are due to the fact that the existing methods fail to distinguish between the primary dis-regulation of a given gene itself and the effects of signals coming from upstream. We hypothesize that better results could be achieved if one distinguishes between genes that are true sources of perturbation, e.g. due to mutations, copy number variations, epigenetic changes, etc. and genes that merely respond to perturbation signals coming from upstream. Intuitively, a pathway should be more significantly impacted if it hosts more genes that are such true sources of perturbation. The method proposed here is

an attempt at capturing these differences by calculating a "primary dis-regulation" for every gene and using them to compute a total pathway perturbation and subsequent significance.

Another issue related to the traditional topological data analysis approaches involves the need for a selection of differentially expressed (DE) genes. Traditionally, the pathway analysis step is performed after a set of DE genes has been selected using some thresholds on some criteria such as fold-change and/or p-values. Typically, a set of a few hundred genes are selected as DE. However, a modern whole-genome experiment performed with a next-generation technology (NGS) provides measurements for the entire set of transcripts in the genome, albeit for a non-trivial cost in computation necessary for the assembly and quantification of millions of short reads. In addition to the high computational cost, other drawbacks are related to the large amount of storage space, and the need to specialized bioinformatics expertise to set-up and run the environment necessary for the analysis. Given that this great deal of effort is spent in order to measure over 30,000 transcripts, it makes little sense to discard approximately 99% of these measurements in order to focus on 300 or so genes that are declared to be differentially expressed. Subsequently, the pathway analysis step aims to identify system-level changes based on only these 1% of the original data collected. More recently, approaches that are able to identify significantly impacted pathways based on the entire set of measurements have been proposed [144]. Henceforth, we will refer to the original approach based on DE genes as the *cut-off*-based approach, and to the threshold-free approach as the *all genes* approach. We assessed the novel method proposed here with both types of input.

In the following section, we describe our new proposed method in details. We evaluate the preliminary results of our method using 24 datasets involving 12 conditions from different experiments comparing disease versus normal tissues. The results of the proposed method using the *cut-off*-based approach are compared with SPIA (cut-off) [127], which also uses a pre-selected list of DE genes as input. The results of the proposed method using the *all genes* approach are compared with GSEA [121], GSA [35] and SPIA (all genes) [144]

which use entire set of genes as input. These existing methods have been selected as references in our comparisons because they are among the most cited and widely used methods in the literature [87]. We also evaluate our method using eight yeast knock-out datasets from different experiments comparing samples with knock-out gene versus normal samples. The comparisons show that the proposed method is able to perform better than the most widely used pathway analysis methods, in identifying the relevant pathways as statistically significant.

2.2 Primary dis-regulation analysis

The measured expression change of a gene in a given phenotype can be seen as the result of influences from upstream genes superposed on the dis-regulation incurred by that particular gene itself. We will refer to this later quantity as the primary dis-regulation (pDis). The diffusion of signals between genes in regulatory networks, called "network propagation", can be used to find the active genes and subnetworks as well as the function of the genes in different conditions [53]. Widely used methods in this field are introduced in [149] and [141]. Here, we are using a similar approach that uses propagation between genes to calculate pDis in order to find the most impacted pathways. We propose a pathway analysis method that focuses on this primary dis-regulation.

The change in the expression level of a gene i, $\Delta E(g_i)$, can be seen as a sum of the primary dis-regulation (pDis) and the secondary dis-regulation (sDis):

$$\Delta E(g_i) = pDis(g_i) + sDis(g_i) \tag{2.1}$$

The secondary dis-regulation of the gene g_i is the term that is meant to capture the perturbation reaching this particular gene from upstream. This can be calculated by adding the expression change of upstream genes normalized by the number of their downstream

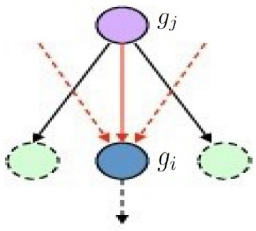


Figure 2.1: An example of one upstream gene and its three downstream genes. $pDis(g_i)$ is calculated using its measured fold change of $\Delta E(g_i)$ and measured fold change of upstream genes (e.g. $\Delta E(g_j)$). In this example, the number of downstream genes for g_j is $N_{ds}(g_j) = 3$.

genes:

$$\Delta E(g_i) = pDis(g_i) + \sum_{j \in U} \frac{\beta_{i,j} \cdot \Delta E(g_j)}{N_{ds}(g_j)} \tag{2.2}$$

In the equation above, $\Delta E(g_j)$ is the measured fold change of the gene g_j that is somewhere directly upstream of g_i, U is the set of all such genes directly upstream of g_i, and $N_{ds}(g_j)$ is the number of genes immediately downstream of g_j (see Figure 2.1). The quantity $\beta_{i,j}$ represents efficiency of the interaction between $gene_i$ and $gene_j$. It captures a specific value if an interaction is available between two genes. We used $+1$ if the interaction type is activation or expression and -1 if it is inhibition or repression as default values. This is the same approach used by the impact analysis [32].

The primary dis-regulation, which gives the change in a gene expression inherent to the gene itself, can then be derived as follows:

$$pDis(g_i) = \Delta E(g_i) - \sum_{j \in U} \frac{\beta_{i,j} \cdot \Delta E(g_j)}{N_{ds}(g_j)} \tag{2.3}$$

The primary dis-regulation is meant to capture information about the genes that are sources of perturbation in a given phenotype, rather than those genes that change as a result of upstream changes. For instance, a mutation that induce expression changes would be captured by the gene's primary dis-regulation, while expression changes due to upstream

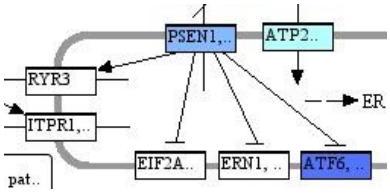


Figure 2.2: A small fragment of a pathway, which includes PSEN1. The gene PSEN1 has 4 downstream genes. There are two types of interactions in this example: the arrows show activation, while the terminal bars show inhibition. The colors represent the expression changes of the genes (blue is down-regulation).

signaling would be captured by the secondary dis-regulation. A mutation is an example that is sufficient but not necessary to create primary dis-regulation. Other potential cause could be copy number variations, epigenetic changes such as methylation, etc. The intuition motivating the computation of the primary dis-regulation is that pathways that have more genes that are sources of perturbation are more likely to be truly involved in the phenotype.

This computation can be illustrated on the small example shown in Figure 2.2. This figure shows a subgraph from the KEGG Alzheimer's disease pathway. The colors represent the log transformed expression change of those genes in the experiment (darker blue represents more down-regulation). The value of β between genes PSEN1 and RYR3 has been assigned +1 because PSEN1 activates RYR3; the values of betas between PSEN1 and there other 3 genes are -1 because PSEN1 inhibits those downstream genes.

We can see in this figure that gene PSEN1 is down-regulated and it interacts with four downstream genes (one activation and three inhibitions). Among the four downstream genes, RYR3, EIF2A and ERN1 are not differentially expressed, thus their measured ΔEs are zero. The effect of the gene PSEN1 on each downstream gene is equal to its expression change divided by the number of its downstream genes, which in this case is four. Since RYR3, EIF2A, ERN1 and ATF6 have only PSEN1 as an upstream gene, we calculate their $pDis$ based on equation 2.3 as follows:

$$pDis(RYR3) = \Delta E(RYR3) - \Delta E(PSEN1)/4 \tag{2.4}$$

$$pDis(EIF2A) = \Delta E(EIF2A) + \Delta E(PSEN1)/4 \tag{2.5}$$

$$pDis(ERN1) = \Delta E(ERN1) + \Delta E(PSEN1)/4 \tag{2.6}$$

$$pDis(ATF6) = \Delta E(ATF6) + \Delta E(PSEN1)/4 \tag{2.7}$$

The process of calculating all values of the primary dis-regulation for all the genes in a given pathway can be summarized using the matrix equation:

$$pDis = \Delta E^T \cdot (I - B) \tag{2.8}$$

In this equation, the matrix B represents the adjacency matrix of each signaling pathway normalized by the number of downstream genes of each gene.

$$B = \begin{pmatrix} \beta_{1,1}/N_{ds(g_1)} & \beta_{1,2}/N_{ds(g_2)} & ... & \beta_{1,n}/N_{ds(g_n)} \\ \beta_{2,1}/N_{ds(g_1)} & \beta_{2,2}/N_{ds(g_2)} & ... & \beta_{2,n}/N_{ds(g_n)} \\ ... & ... & ... & ... \\ \beta_{n,1}/N_{ds(g_1)} & \beta_{n,2}/N_{ds(g_2)} & ... & \beta_{n,n}/N_{ds(g_n)} \end{pmatrix}$$

In equation 2.8, I is an identity matrix with dimensions equal to the number of genes in a pathway, and ΔE is the vector of measured expression changes of the genes in that

pathway:

$$\Delta E = \begin{pmatrix} \Delta E(g_1) \\ \Delta E(g_2) \\ ... \\ \Delta E(g_n) \end{pmatrix}$$

The score for pathway k is calculated as the sum of the absolute values of primary dis-regulation of all the genes in the pathway, $totalpDis$:

$$totalpDis_k = \sum_{i \in pathway_k} |pDis(g_i)| \tag{2.9}$$

The quantity $totalpDis$ of a pathway represents the amount of primary dis-regulation of the whole pathway in the condition under study.

The significance of each pathway is assessed by computing the probability of obtaining just by chance a $totalpDis$ value more extreme than the one observed. This probability is estimated using a bootstrap approach where the null distribution for $totalpDis$ for each pathway is generated by sampling random gene expression changes from the original set of expression changes. The number of bootstraps used was 2,000. This process is repeated for all pathways and yields a p-value for each pathway. Subsequently, the set of p-values for all pathways are corrected for multiple comparisons using the false discovery rate (FDR). The average running time for a dataset is 6.3 minutes on an architecture using a single Intel Xeon core @ 2.66GHz with 1TB of RAM. The complexity of the method depends on the number of pathways investigated multiply by the number of genes in each pathways ($O(N^2)$).

2.3 Cut-off dependent versus cut-off free analysis

Pathway analysis techniques often take a subset of statistically significant genes as input, based on cut-offs for expression change and/or p-value. It has been shown in [95] that small variations of the threshold used to select the subset of differentially expressed

(DE) genes has dramatic effects on the outcome of the methods. Hence, the accuracy of any pathway analysis methods using a subset of DE genes will also be very dependent on the threshold(s) used. Furthermore, when using a cut-off, some genes that play an important biological role may fail to meet the selection criteria and thus, not included in the set of DE genes. This can potentially impede the identification of the biologically meaningful pathways.

Recently, it has also been shown that the accuracy of a pathway analysis method can be improved by using the entire set of measurement from an experiment rather than a subset of DE genes [144]. This means that a selection of a set of DE genes may no longer be needed in many situations.

With respect to the method proposed in this research, the use of a subset of DE genes will affect the values of the pDis of other genes in a pathway. The pDis of a gene is simply equal to the expression change when there are no upstream DE genes. However, when such upstream genes do exist, pDis is calculated using the expression changes of upstream genes as well. The inclusion of all genes in the calculation will have a strong impact on the result, even if the expression changes are small. This allows the analysis to retain all of the information in the data, avoiding arbitrary threshold choices.

We refer to this method as *pDis analysis (all genes)*, as opposed to *pDis analysis (cut-off)* for cut-off based. We show the results from both types of input sets applied to our new method proposed.

2.4 Discussion and results

To date there is no universally accepted technique for the validation of the results of pathway analysis methods. The assessment of the results of different pathway analysis methods usually involves the selection of a few datasets, and then the interpretation of the results either with the help of biologists in the field, or by searching the published literature. This approach is very limited because it can only be applied to a small number of datasets. Furthermore, it is subjective, and may lead to bias in the results since most of the time the

expert who performs the assessment is also a co-author of the paper. Finally, the biological phenomena are so complex that with enough literature search, a large number of pathways can be implicated in almost any condition. In this work, we follow two validation approaches. The first one is the validation approach introduced in [126]. We like this evaluation approach because it is objective, reproducible, based on multiple datasets, and it does not require an unavoidably biased "expert" human evaluation of the results [126]. This approach requires testing on a large number (at least 10 but preferably more) of different datasets coming from a variety of different conditions, tissues, and laboratories. The datasets are selected such that there are specific pathways in the pathway databases that model each of the given diseases. For each dataset, the pathway corresponding to the phenotype is considered to be the target pathway (e.g. the colorectal cancer pathway will be the target pathway in a colorectal cancer dataset). The evaluation focuses on the ability of each method to identify these true positive pathways as significant, and rank them as high as possible. In this thesis, we validated the proposed method using 24 datasets involving 12 different human diseases. These datasets are shown in Table 2.1.

The second approach uses knock-out datasets. In this case, the exact source of perturbation is known: the specific gene being knock-out. Thus the pathways that include this gene will be truly relevant to the phenotype, since they contain the very source of the perturbation that created the phenotype. In other words, these pathways are true positives and are also considered as the target pathways in our validation.

The p-values (representing the probability of observing the given perturbations just by chance) are used to assign significance to each pathway. The list of pathways is then ranked based on these p-values.

In order to formalize and quantify the assessment, we define an "improvement factor" that will be used to compare the performance of two pathway analysis methods. If the target pathway for a given dataset goes from not significant in the results of method 1 to significant in the results of method 2, the improvement factor for this dataset will be 1 (see Figure 2.3).

	GEO ID	Pubmed	Reference	Disease	Target Pathway
1	GSE1297	14769913	[17]	Alzheimer's Disease	hsa05010
2	GSE5281	17077275	[74]	Alzheimer's Disease	hsa05010
3	GSE5281	17077275	[74]	Alzheimer's Disease	hsa05010
4	GSE5281	17077275	[74]	Alzheimer's Disease	hsa05010
5	GSE20153	20926834	[156]	Parkinson's disease	hsa05012
6	GSE20291	15965975	[155]	Parkinson's disease	hsa05012
7	GSE8762	17724341	[110]	Huntington's disease	hsa05016
8	GSE4107	17317818	[50]	Colorectal Cancer	hsa05210
9	GSE8671	18171984	[111]	Colorectal Cancer	hsa05210
10	GSE9348	20143136	[49]	Colorectal Cancer	hsa05210
11	GSE14762	19252501	[150]	Renal Cancer	hsa05211
12	GSE781	14641932	[73]	Renal Cancer	hsa05211
13	GSE15471	19260470	[2]	Pancreatic Cancer	hsa05212
14	GSE16515	19732725	[97]	Pancreatic Cancer	hsa05212
15	GSE19728	NA	NA	Glioma	hsa05214
16	GSE21354	NA	NA	Glioma	hsa05214
17	GSE6956	18245496	[147]	Prostate Cancer	hsa05215
18	GSE6956	18245496	[147]	Prostate Cancer	hsa05215
19	GSE3467	16365291	[45]	Thyroid Cancer	hsa05216
20	GSE3678	NA	NA	Thyroid Cancer	hsa05216
21	GSE9476	17910043	[120]	Acute myeloid leukemia	hsa05221
22	GSE18842	20878980	[112]	Non-Small Cell Lung Cancer	hsa05223
23	GSE19188	20421987	[51]	Non-Small Cell Lung Cancer	hsa05223
24	GSE3585	17045896	[11]	Dilated cardiomyopathy	hsa05414

Table 2.1: The twenty-four datasets from the GEO database used to evaluate the pathway analysis methods compared in this thesis. Each dataset corresponds to a disease for which there is a target pathway in KEGG.

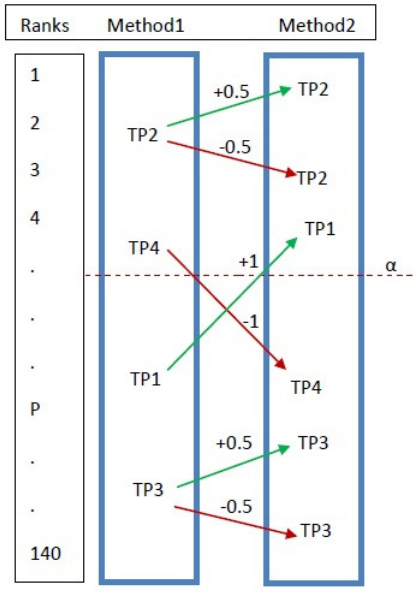


Figure 2.3: The improvement factor criterion used to assess the results. Alpha (α) represents the chosen significance threshold. The green and red arrows denote situations in which method 2 is better or worse than method 1, respectively. The number on each arrow represents the value the improvement factor in each case. If a target pathway becomes significant in the results of method 2, the improvement factor for that target pathway will be +1 (e.g. target pathway TP1); if the pathway becomes not significant, the improvement factor is considered -1 (e.g. TP4). If the target remains on the same side of the significance threshold, the improvement factor is considered +0.5 or -0.5 based on the improvement or deterioration of the rank, respectively (e.g. TP2 and TP3).

If the target pathway goes from significant to not significant, the improvement factor will be -1. If the significance of the target pathway does not change but the ranking improves, the improvement factor will be +0.5. Finally, if the significance does not change but the ranking worsens, the improvement will be -0.5. If the ranking remains the same, the improvement is zero for that dataset. The improvement of method 2 compared to method 1 is the average of improvement factors associated to each target pathway over the set of 24 different datasets. If the overall improvement is positive, then method 2 is considered to perform better than method 1 based on this validation method.

The proposed method is implemented using the R statistical programming environment [129] and is available as a Bioconductor R package (ROntoTools [145]). We used KEGG signaling pathways as input pathways. The pathways were obtained from the "SPIA" R pack-

age version 2.14.0 [128] as included in Bioconductor version 2.13, released on October 15, 2013. We selected all pathways that have at least one interaction with the type of activation, inhibition, repression or expression between their genes. This resulted in a set of 139 pathways. The results of pDis analysis (all genes) are compared to GSA, GSEA and SPIA (all genes) and the results of pDis analysis (cut-off) are compared to SPIA (cut-off). SPIA (cut-off) combines two different p-values. One is the perturbation p-value (pPERT) of a pathway. The perturbation p-value is computed based on the perturbation accumulation of the pathway, which is the sum of the perturbation factors of its genes. The other p-value of SPIA is the hypergeometric p-value, based on the number of DE genes in the pathway in a given dataset. Since the number of DE genes in each pathway does not depend on the analysis method, the hypergeometric p-value is the same in SPIA (cut-off) and the method proposed.

Each dataset was normalized by the "mar" normalization method available in the "affy" R package (version 1.38.1) [54] from Bioconductor version 2.12, released on April 4, 2013. For each gene, the probe id was mapped to gene Entrez ID. The fold change between normal and disease conditions for each probe was calculated by using the "limma" package (version 3.16.8) [118] from Bioconductor version 2.12, released on April 4, 2013. We used the log2-transform of the fold changes for each gene in our analysis. The moderated t-test was performed on each probe to compute the significance of the changes between two phenotypes. For the methods that use cut-off approach, we used a 5% threshold to select the DE genes.

2.4.1 Results of the target pathways for 24 disease datasets

The ranks and p-values of target pathways in all human disease datasets are shown in Figure. 2.4. The details of the results for the proposed and reference methods are provided in Table 2.3 (SPIA and pDis analysis (cut-off) and Table 2.4, 2.5 and 2.6 (GSEA, GSA, SPIA (all genes) and pDis analysis (all genes)). The distributions of the ranks and the p-values obtained for the target pathways in four methods are shown as boxplots in Figure. 2.4.

P-value (paired t-test p-value)	SPIA (cut-off)	GSA	GSEA	SPIA (all genes)
pDis analysis (cut-off)	**0.01**	-	-	-
pDis analysis (all genes)	-	0.07	0.074	**0.01**

Ranks (paired Wilcoxon.test p-value)	SPIA (cut-off)	GSA	GSEA	SPIA (all genes)
pDis analysis (cut-off)	0.13	-	-	-
pDis analysis (all genes)	-	0.75	0.29	**0.05**

Table 2.2: Results of the statistical tests that were performed to compare the results of the various pathway analysis methods. pDis analysis (cut-off) was compared to SPIA (cut-off). pDis analysis (all genes) was compared to GSEA, GSA and SPIA (all genes). Each p-value shows whether the ranks and the p-values of the target pathways in proposed method are significantly lower than the reference methods (at 5% significance threshold). The results show that pDis analysis (cut-off) yields significantly better p-values than SPIA (cut-off) for the target pathways. Also, pDis analysis (all genes) yields lower p-values as well as lower ranks compared to GSEA and SPIA (all genes).

The paired t-test and the paired Wilcoxon test were performed to compare the distribution of the ranks and p-values of target pathways in each method. The results are shown in Table 2.2. The statistical tests are performed as one-tail tests in order to test whether the ranks and p-values of target pathways in proposed methods are significantly lower than the reference methods. The results show that the p-values of the target pathways in pDis analysis (cut-off) are significantly lower than SPIA. Furthermore, the ranks and the p-values of the target pathways in pDis analysis (all genes) is significantly lower than GSEA. The p-values of pDis analysis (all genes) are also lower than those yielded by GSA but not significantly so (at 5%).

The pDis analysis (all genes) yields better results compared to GSEA, in term of both ranks (panel C in Figure 2.4, Wilcoxon test p-value = 0.29), as well as p-values of the target pathways (panel D in Figure 2.4, t-test p-value = 0.074). The proposed method yields significantly better results compared to SPIA (all genes) in terms of both ranks (panel C in Figure 2.4, Wilcoxon test p-value = 0.05), as well as p-values of the target pathways (panel D

	GEO ID	Target pathway	SPIA (pPERT)			pDis analysis (cut-off)			Improvement
			p-values	FDR	ranks	p-values	FDR	ranks	Compared to SPIA
1	GSE1297	Alzheimer's Disease	0.916	1.00	78.79	0.729	0.987	70.83	+0.5
2	GSE5281	Alzheimer's Disease	0.807	1.00	71.53	0.022	0.328	6.20	+0.5
3	GSE5281	Alzheimer's Disease	0.956	1.00	92.54	0.006	0.201	2.99	+0.5
4	GSE5281	Alzheimer's Disease	0.831	0.985	82.20	0.068	0.359	18.94	+0.5
5	GSE20153	Parkinson's disease	1	1.00	62.82	1	1.00	98.72	-0.5
6	GSE20291	Parkinson's disease	0.129	0.712	18.10	0.425	0.803	50.48	-0.5
7	GSE8762	Huntington's disease	1	1.00	69.49	0.425	0.524	79.66	-0.5
8	GSE4107	Colorectal Cancer	0.011	0.213	5.15	0.023	0.184	12.50	-0.5
9	GSE8671	Colorectal Cancer	0.406	0.778	50.74	0.351	0.772	44.12	+0.5
10	GSE9348	Colorectal Cancer	0.198	0.503	37.96	0.387	0.679	56.93	-0.5
11	GSE14762	Renal Cancer	0.009	0.07	12.04	0.482	0.786	61.31	-0.5
12	GSE781	Renal Cancer	0.412	1.00	36.94	0.859	0.935	91.79	-0.5
13	GSE15471	Pancreatic Cancer	0.651	0.843	76.47	0.451	0.757	59.56	+0.5
14	GSE16515	Pancreatic Cancer	0.94	1.00	88.15	0.452	0.796	55.56	+0.5
15	GSE19728	Glioma	0.979	1.00	91.24	0.235	0.485	47.45	+0.5
16	GSE21354	Glioma	0.367	0.744	49.26	0.342	0.560	61.03	-0.5
17	GSE6956	Prostate Cancer	0.21	1.00	17.28	0.771	0.981	74.26	-0.5
18	GSE6956	Prostate Cancer	0.592	1.00	54.13	0.555	0.993	41.32	+0.5
19	GSE3467	Thyroid Cancer	0.745	0.957	77.78	0.57	0.925	58.52	+0.5
20	GSE3678	Thyroid Cancer	0.59	0.987	59.70	0.313	0.706	41.04	+0.5
21	GSE9476	Acute myeloid leukemia	0.164	0.841	18.11	0.064	0.825	5.51	+0.5
22	GSE18842	Non-Small Cell Lung Cancer	0.395	0.857	44.53	0.06	0.348	16.79	+0.5
23	GSE19188	Non-Small Cell Lung Cancer	0.97	1.00	93.43	0.176	0.560	31.39	+0.5
24	GSE3585	Dilated cardiomyopathy	1	1.00	81.18	0.577	0.825	69.89	+0.5
Average			0.595	0.854	57.06	0.389	0.682	48.19	+3/24=12.5%

Table 2.3: The ranks and the p-values of the 24 target pathways for SPIA (cut-off) and pDis analysis (cut-off). The improvement factor based on Figure 2.3 is calculated for each dataset considering the 5% significance threshold using FDR-corrected p-values. The average improvement factor shows that pDis analysis (cut-off) improves 12.5% compared to SPIA (cut-off). As shown, the average p-value and rank for the target pathways are lower (i.e. better) in pDis analysis (cut-off) than in SPIA (cut-off).

in Figure 2.4, t-test p-value =0.01). The results also show that the proposed method provides more significant p-values compared to GSA, even thought the differences are not statistically significant (see Table 2.2). There is not significant difference between the ranks yielded by pDis (all genes) and GSA. The figure also shows the comparison between pDis analysis (cut-off) and SPIA (cut-off). The proposed method yields significantly better results compared to SPIA (cut-off) in terms of p-values (panel B in Figure 2.4, t-test p=0.01). The results are also better in terms of ranks, even though the difference is not statistically significant (panel A in Fig 2.4, Wilcoxon test p-value =0.13).

As some diseases are complex phenotypes involving fundamental biochemical pathways, other pathways might be significantly impacted in addition to the target pathway. Therefore, we studied the detailed results of pDis analysis (all genes) on a dataset, in order to show that our method is not limited to identifying the target pathway as significantly impacted, but it is also able to correctly report relevant fundamental biochemical mechanisms

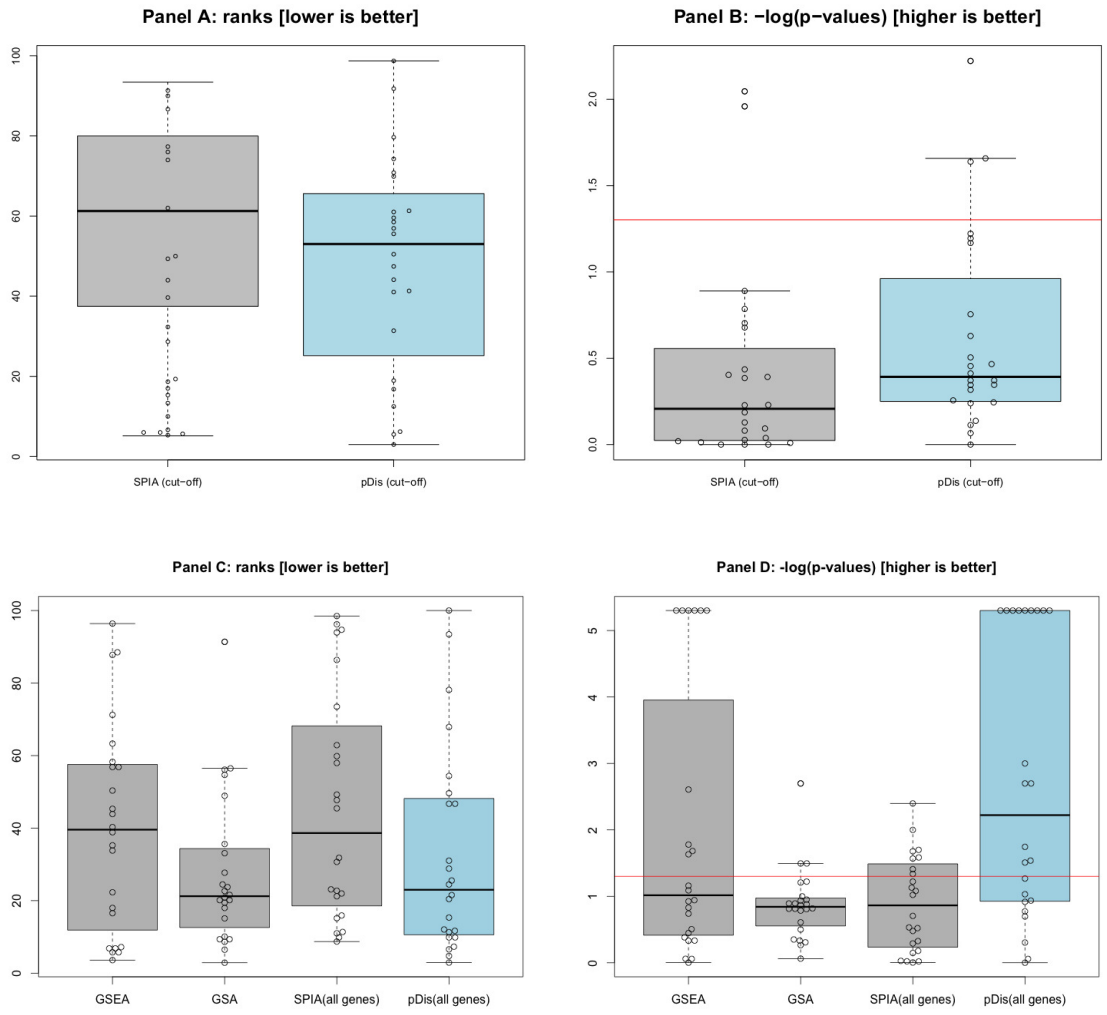


Figure 2.4: The ranks (in the left column, lower is better) and negative log of p-values of the target pathways (in the right column, higher is better) in the proposed and reference methods. The first row (panel A and panel B) shows the comparison between methods using a set of DE genes: pDis (cut-off) and SPIA (cut-off). The second row (panel C and panel D) shows the comparison between methods using all genes: GSEA, GSA and SPIA (all genes), pDis (all genes). For SPIA, the comparisons are based on the perturbation p-value (pPERT). All human signaling pathways from KEGG (139 pathways) were used in the comparisons. The data show the results obtained for the target pathways in the 24 datasets shown in Table 2.1. The bold line in the boxplots represents the median of the distribution. These distributions show that the proposed method pDis analysis (in blue) is never significantly worse than any of the existing methods, while it yields a statistically significant improvement in 5 out of the 8 comparisons (see Table 2.2).

	GEO ID	Target pathway	GSA			pDis analysis (all genes)			Improvement
			p-values	FDR	ranks	p-values	FDR	ranks	Compared to GSA
1	GSE1297	Alzheimer's Disease	0.100	0.514	19.42	5e-06	5.7e-05	4.74	+1.0
2	GSE5281	Alzheimer's Disease	0.316	0.872	33.09	5e-06	3.6e-05	7.29	+1.0
3	GSE5281	Alzheimer's Disease	0.116	0.488	23.74	5e-06	2.6e-05	9.85	+1.0
4	GSE5281	Alzheimer's Disease	0.164	0.537	27.69	5e-06	2.6e-05	9.85	+1.0
5	GSE20153	Parkinson's disease	0.542	0.885	54.67	0.002	0.008	21.53	+1.0
6	GSE20291	Parkinson's disease	0.246	0.629	35.61	5e-06	2.2e-05	11.67	+1.0
7	GSE8762	Huntington's disease	0.154	0.876	15.10	5e-06	1.6e-05	15.32	+1.0
8	GSE4107	Colorectal Cancer	0.154	0.764	20.14	0.002	0.009	20.43	+1.0
9	GSE8671	Colorectal Cancer	0.002	0.069	2.87	0.116	0.248	46.71	-0.5
10	GSE9348	Colorectal Cancer	0.032	0.342	9.35	0.054	0.172	31.02	-0.5
11	GSE14762	Renal Cancer	0.132	0.600	21.58	0.029	0.062	46.71	-0.5
12	GSE781	Renal Cancer	0.492	0.865	56.47	0.998	0.998	100	-0.5
13	GSE15471	Pancreatic Cancer	0.112	0.622	17.98	0.168	0.247	67.88	-0.5
14	GSE16515	Pancreatic Cancer	0.062	0.625	8.63	0.121	0.242	49.63	-0.5
15	GSE19728	Glioma	0.136	0.548	24.46	5e-06	2.2e-05	11.31	+1.0
16	GSE21354	Glioma	0.128	0.547	22.66	5e-06	2.1e-05	12.04	+1.0
17	GSE6956	Prostate Cancer	0.060	0.440	9.35	0.031	0.124	24.45	-0.5
18	GSE6956	Prostate Cancer	0.032	0.451	6.47	0.001	0.013	6.56	+1.0
19	GSE3467	Thyroid Cancer	0.152	0.687	20.14	0.018	0.061	28.83	-0.5
20	GSE3678	Thyroid Cancer	0.464	0.814	56.11	0.201	0.364	54.38	+0.5
21	GSE9476	Acute myeloid leukemia	0.128	0.902	10.07	5e-06	9.7e-05	2.92	+1.0
22	GSE18842	Non-Small Cell Lung Cancer	0.156	0.6992	20.86	0.496	0.635	78.10	-0.5
23	GSE19188	Non-Small Cell Lung Cancer	0.446	0.844	48.92	0.874	0.925	93.43	-0.5
24	GSE3585	Dilated cardiomyopathy	0.866	0.919	91.36	0.093	0.364	25.54	+0.5
Average			0.216	0.647	27.368	0.133	0.186	32.51	+8/24=33.3%

Table 2.4: The ranks and the p-values of the 24 target pathways for GSA and pDis analysis (all genes). The improvement factor based on Figure 2.3 is calculated for each dataset considering 5% significance threshold using FDR-corrected p-values. The average improvement factor shows that pDis analysis (all genes) improves the results 33.3% compared to GSA. Twelve target pathways were found to be significant in pDis analysis (all genes) while non of the target pathways have significant FDR-corrected p-values in GSA. As shown, the average p-value for the target pathways are lower (i.e. better) in the pDis analysis (all genes) than in GSA.

	GEO ID	Target pathway	GSEA			pDis analysis (all genes)			Improvement
			p-values	FDR	ranks	p-values	FDR	ranks	Compared to GSEA
1	GSE1297	Alzheimer's Disease	5e-06	4e-05	5.75	5e-06	5.7e-05	4.74	+0.5
2	GSE5281	Alzheimer's Disease	5e-06	4e-05	3.59	5e-06	3.6e-05	7.29	-0.5
3	GSE5281	Alzheimer's Disease	5e-06	4e-04	5.75	5e-06	2.6e-05	9.85	-0.5
4	GSE5281	Alzheimer's Disease	5e-06	4e-05	7.19	5e-06	2.6e-05	9.85	-0.5
5	GSE20153	Parkinson's disease	0.995	1	96.40	0.002	0.008	21.53	+1
6	GSE20291	Parkinson's disease	5e-06	4e-05	6.83	5e-06	2.2e-05	11.67	-0.5
7	GSE8762	Huntington's disease	5e-06	0.08	6.83	5e-06	1.6e-05	15.32	+1
8	GSE4107	Colorectal Cancer	0.081	0.171	35.25	0.002	0.009	20.43	+1
9	GSE8671	Colorectal Cancer	0.312	0.625	56.83	0.116	0.248	46.71	-0.5
10	GSE9348	Colorectal Cancer	0.118	0.283	33.81	0.054	0.172	31.02	+0.5
11	GSE14762	Renal Cancer	0.148	0.261	45.32	0.029	0.062	46.71	-0.5
12	GSE781	Renal Cancer	0.356	0.584	58.27	0.998	0.998	100	-0.5
13	GSE15471	Pancreatic Cancer	0.020	0.038	38.84	0.168	0.247	67.88	-1
14	GSE16515	Pancreatic Cancer	0.002	0.019	16.54	0.121	0.242	49.63	-1
15	GSE19728	Glioma	0.069	0.121	50.35	5e-06	2.2e-05	11.31	+1
16	GSE21354	Glioma	0.114	0.248	43.88	5e-06	2.1e-05	12.04	+1
17	GSE6956	Prostate Cancer	0.023	0.170	22.30	0.031	0.124	24.45	-0.5
18	GSE6956	Prostate Cancer	0.016	0.068	17.98	0.001	0.013	6.56	+1
19	GSE3467	Thyroid Cancer	0.463	0.682	71.22	0.018	0.061	28.83	+0.5
20	GSE3678	Thyroid Cancer	0.182	0.353	40.28	0.201	0.364	54.38	-0.5
21	GSE9476	Acute myeloid leukemia	0.4662	0.808	56.83	5e-06	9.7e-05	2.92	+1
22	GSE18842	Non-Small Cell Lung Cancer	0.414	0.727	63.30	0.496	0.635	78.10	-0.5
23	GSE19188	Non-Small Cell Lung Cancer	0.870	0.995	87.76	0.874	0.925	93.43	-0.5
24	GSE3585	Dilated cardiomyopathy	0.874	1	88.48	0.093	0.364	25.54	+0.5
Average			0.23	0.34	39.98	0.133	0.186	32.51	+2.5/24=10%

Table 2.5: The ranks and the p-values of the 24 target pathways for GSEA and pDis analysis (all genes). The improvement factor based on Figure 2.3 is calculated for each dataset considering 5% significance threshold using FDR-corrected p-values. The average improvement factor shows that pDis analysis (all genes) improves the results 10% compared to GSEA. As shown, the average p-value and rank for the target pathways are lower (i.e. better) in the pDis analysis (all genes) than in GSEA.

	GEO ID	Target pathway	SPIA (all genes)			pDis analysis (all genes)			Improvement
			p-values	FDR	ranks	p-values	FDR	ranks	Compared to SPIA (all genes)
1	GSE1297	Alzheimer's Disease	0.095	0.414	22.72	5e-06	5.7e-05	4.74	+1
2	GSE5281	Alzheimer's Disease	0.661	0.880	73.48	5e-06	3.6e-05	7.29	+1
3	GSE5281	Alzheimer's Disease	0.060	0.255	23.10	5e-06	2.6e-05	9.85	+1
4	GSE5281	Alzheimer's Disease	0.332	0.695	47.72	5e-06	2.6e-05	9.85	+1
5	GSE20153	Parkinson's disease	0.021	0.170	11.36	0.002	0.008	21.53	+1
6	GSE20291	Parkinson's disease	0.020	0.174	10.98	5e-06	2.2e-05	11.67	+1
7	GSE8762	Huntington's disease	0.955	0.992	96.21	5e-06	1.6e-05	15.32	+1
8	GSE4107	Colorectal Cancer	0.010	0.101	9.84	0.002	0.009	20.43	+1
9	GSE8671	Colorectal Cancer	0.991	1.00	98.48	0.116	0.248	46.71	+0.5
10	GSE9348	Colorectal Cancer	0.197	0.433	45.45	0.054	0.172	31.02	+0.5
11	GSE14762	Renal Cancer	0.004	0.025	15.90	0.029	0.062	46.71	-1
12	GSE781	Renal Cancer	0.074	0.338	21.21	0.998	0.998	100	-0.5
13	GSE15471	Pancreatic Cancer	0.039	0.125	30.68	0.168	0.247	67.88	-0.5
14	GSE16515	Pancreatic Cancer	0.046	0.144	31.81	0.121	0.242	49.63	-0.5
15	GSE19728	Glioma	0.301	0.502	59.84	5e-06	2.2e-05	11.31	+1
16	GSE21354	Glioma	0.026	0.118	21.96	5e-06	2.1e-05	12.04	+1
17	GSE6956	Prostate Cancer	0.083	0.543	15.15	0.031	0.124	24.45	-0.5
18	GSE6956	Prostate Cancer	0.027	0.294	8.71	0.001	0.013	6.56	+1
19	GSE3467	Thyroid Cancer	0.936	0.990	93.93	0.018	0.061	28.83	+0.5
20	GSE3678	Thyroid Cancer	0.951	0.977	94.69	0.201	0.364	54.38	+0.5
21	GSE9476	Acute myeloid leukemia	0.512	0.793	62.87	5e-06	9.7e-05	2.92	+1
22	GSE18842	Non-Small Cell Lung Cancer	0.294	0.591	49.24	0.496	0.635	78.10	-0.5
23	GSE19188	Non-Small Cell Lung Cancer	0.712	0.824	86.36	0.874	0.925	93.43	-0.5
24	GSE3585	Dilated cardiomyopathy	0.471	0.801	57.95	0.093	0.364	25.54	+0.5
Average			0.325	0.507	45.40	0.133	0.186	32.51	+11/24=43.75%

Table 2.6: The ranks and the p-values of the 24 target pathways for SPIA (all genes) and pDis analysis (all genes). The improvement factor based on Figure 2.3 is calculated for each dataset considering 5% significance threshold using FDR-corrected p-values. The average improvement factor shows that pDis analysis (all genes) improves the results 43.75% compared to SPIA (all genes). Twelve target pathways were found to be significant in pDis analysis (all genes) while only one target pathway have significant FDR-corrected p-values in SPIA (all genes). As shown, the average p-value and rank for the target pathways are lower (i.e. better) in the pDis analysis (all genes) than in SPIA (all genes).

in the condition under study. We chose to perform detailed analysis of the first neurodegenerative disease as it appears in Table 2.1. We provide the information about the p-values of all the analyzed pathways with FDR-corrected p-value lower that 5% for the dataset studying Alzheimer's disease [17] (see Table 2.7). The pathways with bold font in each table are known to be related to that disease based on existing literature. We can see that most of the significant pathways are biologically meaningful in the condition studied, showing high precision in the results. These results indicate that the proposed method is able to report the target pathways as more significant and ranked higher, compared to the state-of-the-art methods for pathway analysis, as well as it is able to report as significant the pathways that are known to be associated to a given disease.

	Name	ID	p-values	FDR	ranks	references
1	Alzheimer's disease	05010	5e-06	5e-05	4.74	
2	**Cytokine-cytokine receptor interaction**	04060	5e-06	5e-05	4.74	[109]
3	**Glutamatergic synapse**	04724	5e-06	5e-05	4.74	[38]
4	**GABAergic synapse**	04727	5e-06	5e-05	4.74	[76]
5	**Dopaminergic synapse**	04728	5e-06	5e-05	4.74	[84]
6	**Long-term depression**	04730	5e-06	5e-05	4.74	[98]
7	Endocrine and other factor-regulated calcium reabsorption	04961	5e-06	5e-05	4.74	
8	**Parkinson's disease**	05012	5e-06	5e-05	4.74	[5, 24]
9	**Amyotrophic lateral sclerosis (ALS)**	05014	5e-06	5e-05	4.74	[85]
10	**Huntington's disease**	05016	5e-06	5e-05	4.74	[24, 44]
11	Vibrio cholerae infection	05110	5e-06	5e-05	4.74	
12	Pathogenic Escherichia coli infection	05130	5e-06	5e-05	4.74	
13	Oocyte meiosis	04114	1e-03	0.01	10.95	
14	**Long-term potentiation**	04720	1e-03	0.01	10.95	[67, 138]
15	**Retrograde endocannabinoid signaling**	04723	1e-03	0.01	10.95	[89]
16	Gastric acid secretion	04971	1e-03	0.01	10.95	
17	Pancreatic secretion	04972	1e-03	0.01	10.95	
18	**VEGF signaling pathway**	04370	2e-03	0.01	13.87	[48, 105]
19	Epithelial cell signaling in Helicobacter pylori infection	05120	2e-03	0.01	13.87	
20	Systemic lupus erythematosus	05322	2e-03	0.01	13.87	
21	Salmonella infection	05132	3e-03	0.02	15.33	
22	**Calcium signaling pathway**	04020	0.01	0.03	16.79	[22, 153]
23	Salivary secretion	04970	0.01	0.03	16.79	
24	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	05412	0.01	0.03	16.79	
25	**Gap junction**	04540	0.01	0.03	18.25	[82, 124]
26	**Phosphatidylinositol signaling system**	04070	0.01	0.04	18.98	[14, 158]
27	Morphine addiction	05032	0.01	0.04	19.71	
28	**Protein processing in endoplasmic reticulum**	04141	0.01	0.04	20.80	[142]
29	Shigellosis	05131	0.01	0.04	20.80	
30	Renal cell carcinoma	05211	0.01	0.05	21.90	

Table 2.7: The resulting ranks and p-values for all the pathways with FDR-corrected p-value lower than 5% from analyzing the dataset studying Alzheimer's disease [17]. We studied the association of these top pathways to Alzheimer's disease. The pathway shown in red is the target pathway with the name corresponding to the disease under study. The bold pathways are the ones with known association with Alzheimer's disease based on existing literature. The number of bold and red pathways represents the number of true positives found by the method. Here we can see 16 true positives with FDR-corrected p-value lower than 5%.

2.4.2 Results of the target pathways for eight yeast knock-out datasets

We also validate our approach using eight datasets that come from experiments studying eight yeast knock-out genes. We obtained the KEGG signaling pathways for yeast from the "ROntoTools" R package version 1.2.0 [146] as included in Bioconductor version 2.12 released on April 4, 2013. We used all pathways that have at least one interaction of type *activation, inhibition, expression*, or *repression*. There are nine such yeast pathways in KEGG. We used the data provided by [61] as our wild type and knock-out sample. These are contained in the datasets GSE42215 [61] and GSE42527 [61], respectively. We selected eight knock-out datasets whose knock-out genes belong to at least one pathway considered in the analysis. The log2-fold changes for each knock-out sample were calculated by comparing expression levels of that sample with the wild type samples. Each dataset was processed as described above. We performed the pDis analysis (all genes), SPIA (all genes) and GSA, for each of the eight knock-out samples using the calculated log2-fold changes. The target pathways for each knock-out data are the pathways that include the knock-out genes. The ranks and p-values of the target pathways for eight yeast knock-out datasets are shown in the Tables 2.8 and 2.9. The data show an improvement of about 55% with respect to SPIA (all genes) and an improvement of about 20% with respect to GSA. The GSEA results were not included in the comparison on the knock-out datasets because GSEA analysis is not available for yeast pathways. The statistical tests are performed as one-tail in order to test whether the ranks and p-values of target pathways in proposed methods are significantly lower than the reference methods. The proposed method yields significantly better results compared to SPIA (all genes) in terms of both p-values (t-test p-value = 0.002) as well as ranks of the target pathways (Wilcoxon test p-value = 0.01). The result show that pDis (all genes) provides lower p-values (t-test p-value = 0.09) and lower ranks for the target pathways, although not significantly (Wilcoxon test p-value = 0.36) when compared to GSA.

	knock-out genes	target pathway	SPIA (all genes)			pDis analysis (all genes)			improvement factors
			p-values	FDR	ranks	p-values	FDR	ranks	compared to SPIA
1	APC9	Cell cycle	0.422	1	3	0.133	0.411	2	+0.5
		Meiosis - yeast	0.917	1	5	0.184	0.414	4	+0.5
2	TPK3	Meiosis- yeast	5e-06	4.5e-05	1	5e-06	4e-05	1	-
3	RGT2	Meiosis- yeast	0.075	0.225	3	0.001	0.009	1	+1
4	USA1	Protein processing in endoplasmic reticulum	1	1	7.5	0.092	0.27	3	+0.5
5	TIF4631	RNA transport	1	1	7.5	0.084	0.756	1	+0.5
6	URM1	Sulfur relay system	0.048	0.306	1	0.040	0.36	1	+1
7	SSM4	Protein processing in endoplasmic reticulum	1	1	7.5	0.004	0.018	2	+0.5
8	CUE1	Protein processing in endoplasmic reticulum	1	1	7.5	0.208	0.624	3	+0.5
Average			0.606	0.725	4.77	0.082	0.318	2	5/9=55%

Table 2.8: The ranks and the p-values of the target pathways for SPIA (all genes) and pDis analysis (all genes) using 8 yeast knock-out datasets. The improvement factor based on Figure 2.3 is calculated for each dataset considering 5% significance threshold using FDR-corrected p-values. The average improvement factor shows that pDis analysis (all genes) improves the results 55% compared to SPIA (all genes). Three target pathways were found to be significant after FDR-correction in pDis analysis (all genes) while only one target pathways have significant FDR-corrected p-values in SPIA (all genes). As shown, the average p-value and rank for the target pathways are lower (i.e. better) in the pDis analysis (all genes) than in SPIA (all genes). The results show that pDis analysis (all genes) yields significantly better p-values than SPIA (all genes) for the target pathways (p-value from t-test = 0.002) as well as it has significantly lower ranks for the target pathways compared to SPIA (all genes) (p-value from Wilcoxon test = 0.01).

	knock-out genes	target pathway	GSA			pDis analysis (all genes)			improvement factors
			p-values	FDR	ranks	p-values	FDR	ranks	compared to GSA
1	APC9	Cell cycle	0.05	0.28	1	0.133	0.411	2	-0.5
		Meiosis - yeast	0.06	0.28	2	0.184	0.414	4	-0.5
2	TPK3	Meiosis- yeast	0.40	0.63	5	5e-06	4e-05	1	+1
3	RGT2	Meiosis- yeast	0.98	0.98	9	0.001	0.009	1	+1
4	USA1	Protein processing in endoplasmic reticulum	0.09	0.78	1	0.092	0.27	3	-0.5
5	TIF4631	RNA transport	0.78	0.88	6	0.084	0.756	1	+0.5
6	URM1	Sulfur relay system	0.02	0.09	2	0.040	0.36	1	+0.5
7	SSM4	Protein processing in endoplasmic reticulum	0.04	0.43	1	0.004	0.018	2	+1
8	CUE1	Protein processing in endoplasmic reticulum	0.04	0.31	1	0.208	0.624	3	-0.5
Average			0.27	0.51	3.1	0.082	0.318	2	2/9=22.2%

Table 2.9: The ranks and the p-values of the target pathways for GSA and pDis analysis (all genes) using 8 yeast knock-out datasets. The improvement factor based on Figure 2.3 is calculated for each dataset considering 5% significance threshold using FDR-corrected p-values. The average improvement factor shows that pDis analysis (all genes) improves the results 22.2% compared to GSA. Three target pathways were found to be significant after FDR-corrected-correction in pDis analysis (all genes) while no target pathways have significant FDR-corrected p-values in GSA. As shown, the average p-value and rank for the target pathways are lower (i.e. better) in the pDis analysis (all genes) than in GSA. The results show that pDis analysis (all genes) yields better p-values than GSA for the target pathways (p-value from t-test = 0.09) as well as it has lower ranks for the target pathways compared to GSA (p-value from Wilcoxon test = 0.36).

2.4.3 False positives under the null hypothesis

As we have demonstrated, the proposed method produces significantly lower p-values for the target pathways compared with the existing methods, across the set of 24 datasets used in the validation. However, lower p-values for the target pathways could be produced if the new method indiscriminately lowered the p-values for *all pathways*, thus introducing many false positives.

In order to show that this is not the case, we ran a number of experiments with completely random data. In each of these experiments, a set of expression changes are assigned to the genes from a random normal distribution with mean of zero and standard deviation of 1. This was repeated 1,000 times and the p-values for the pathways were computed in each iteration. The pathways' p-values for these random datasets, produced the distribution for the p-values under the null hypothesis. Null-hypothesis distributions were also calculated for each target pathway and showed no abnormal tendencies. The distribution of the pooled p-values for all pathways over the 1,000 iterations is shown in Figure 2.5. Both the distribution of the pooled p-values, as well as all null distributions associated with each individual target pathway were uniform, demonstrating that our method does not yield more significant p-values for the target pathways by lowering all p-values. These distributions demonstrate that the proposed method does not produce any more false positives than appropriate for any significance threshold.

2.4.4 Results of the target pathways for weighted-pDis analysis

Most of the existing pathway analysis methods compute gene level statistics that are then combined into a pathway level score. One key drawback is that the same gene statistic is considered in all pathways containing that gene. This may be a suboptimal strategy since a gene can play different roles in different pathways. Different approaches for estimating the weights of fold changes of the genes are introduced in [126, 143, 144]. A recent method was proposed to detect, quantify and correct the crosstalk effect between overlapping pathways [31]. We hypothesized that the performance of the proposed approach (pDis analysis)

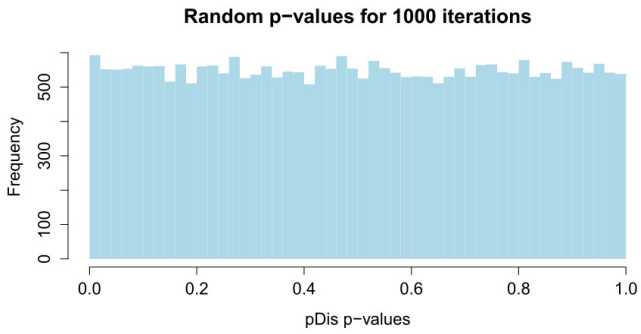


Figure 2.5: The null distribution of the p-values obtained from pDis analysis for all KEGG signaling pathways (139 pathways) in 1,000 iterations. The input gene expression values were chosen from a random normal distribution with mean of 0 and standard deviation of 1. The histogram shows the null distribution of the pooled p-values. Uniform distributions were also obtained for each individual target pathway (data not shown). The uniform distributions prove that pDis analysis does not produce any more false positives than expected.

can be improved by considering effects of the crosstalk between pathways. The contributions of the genes are calculated based on the given condition and the set of pathways. Here, we investigate eliminating such crosstalk effect on the fold change of the genes based on the proposed crosstalk coefficients, which represent the contributions of the genes in each pathway. This contribution is estimated by the number of differentially expressed (DE) genes in the overlapping pathways. A pathway is expected to have higher contribution if it has more DE genes. For more details about the exact formula see Supplementary Material in [31].

We eliminate the crosstalk effect between pathways by weighing the fold changes of the genes in different pathways, to which they belong, as follows:

$$new\Delta E_{ij} = P_{ij} \cdot \Delta E_i \tag{2.10}$$

where, P_{ij} is the weight of $gene_i$ in $pathway_j$. The weighted fold change ($new\Delta E$) of a gene in each pathway represents the corrected involvement of that specific gene in that specific pathway.

The proposed approach is used in conjunction with the proposed pathway analysis method, referred to as "weighted-pDis analysis". The results are validated using the same

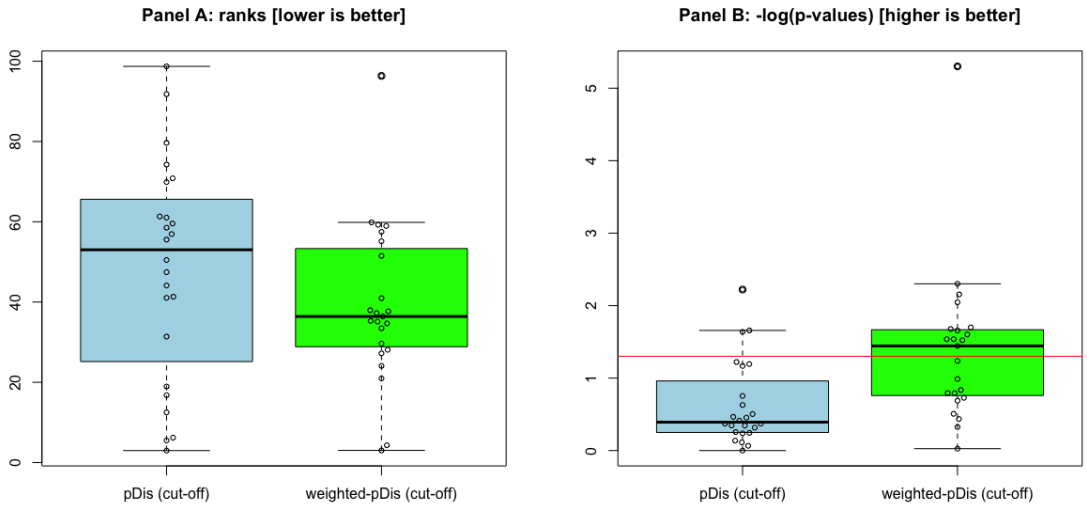


Figure 2.6: The ranks (in the left column, lower is better) and negative log of p-values of the target pathways (in the right column, higher is better) in the pDis and weighted-pDis analysis. All human signaling pathways from KEGG (139 pathways) were used in the comparisons. The box-plots show the results obtained for the target pathways in the 23 datasets shown in Table 2.1. The bold line in the box-plots represents the median of the distribution. The red line represents the 5% significance level. These distributions show that the weighted-pDis (in green) performs better than original pDis analysis.

approaches mentioned above, on 23 real datasets from experiments studying 12 different conditions and are compared to the proposed pDis analysis. We could not analyzed "GSE20153" due to a very small number of DE genes in that dataset (only 3).

The Figure 2.6 shows the distributions of the ranks and p-values of the target pathways in weighted and original pDis analysis. The paired t-test and the paired Wilcoxon test were performed to compare the distributions of the ranks and p-values of target pathways in each method (t-test p-value= 0.0008, Wilcoxon p-value= 0.05). The results show that the p-values and ranks of the target pathways in weighted-pDis analysis are significantly lower than original pDis analysis. The details of the results from each method and the improvement factors of each dataset are shown in table 2.10.

2.4.5 Results of the target pathways for integrated pDis analysis

The current pathway analysis approaches take into consideration pathways as independent entities. However, these pathways can share genes and interactions and therefore

	Target pathway	pDis (cut-off)			weighted-pDis (cut-off)			Improvement
		p-values	FDR	ranks	p-values	FDR	ranks	Compared to pDis (cut-off)
1	Alzheimer's Disease	0.73	0.99	70.83	5e-06	9e-05	3.03	+1
2	Alzheimer's Disease	0.02	0.33	6.20	1.00	1.00	100.00	-0.5
3	Alzheimer's Disease	0.01	0.20	2.99	0.18	0.29	61.19	-0.5
4	Alzheimer's Disease	0.07	0.36	18.94	0.15	0.25	60.61	-0.5
5	Parkinson's disease	0.42	0.80	50.48	0.46	0.99	36.67	+0.5
6	Huntington's disease	0.42	0.52	79.66	0.04	0.37	8.47	+0.5
7	Colorectal Cancer	0.02	0.18	12.50	0.04	0.09	38.97	-0.5
8	Colorectal Cancer	0.35	0.77	44.12	0.03	0.07	36.03	+0.5
9	Colorectal Cancer	0.39	0.68	56.93	0.02	0.06	31.75	+0.5
10	Renal Cancer	0.48	0.79	61.31	0.02	0.07	30.66	+0.5
11	Renal Cancer	0.86	0.94	91.79	0.08	0.24	33.58	+0.5
12	Pancreatic Cancer	0.45	0.76	59.56	0.01	0.05	25.00	+1
13	Pancreatic Cancer	0.45	0.80	55.56	0.03	0.09	29.63	+0.5
14	Glioma	0.23	0.49	47.45	5e-06	0.02	18.25	+1
15	Glioma	0.34	0.56	61.03	5e-06	0.02	12.50	+1
16	Prostate Cancer	0.77	0.98	74.26	0.19	0.35	52.94	+0.5
17	Prostate Cancer	0.56	0.99	41.32	0.35	0.82	32.23	+0.5
18	Thyroid Cancer	0.57	0.93	58.52	0.18	0.30	59.26	-0.5
19	Thyroid Cancer	0.31	0.76	41.04	0.17	0.32	52.24	-0.5
20	Acute myeloid leukemia	5.51	0.06	0.83	0.29	0.52	55.12	-0.5
21	Non-Small Cell Lung Cancer	16.79	0.06	0.35	0.05	0.14	36.50	-0.5
22	Non-Small Cell Lung Cancer	31.39	0.18	0.56	0.05	0.13	39.42	-0.5
23	Dilated cardiomyopathy	69.89	0.58	0.83	0.03	0.65	4.30	+0.5
	average	0.36	0.66	46.00	0.14	0.29	37.31	4.5/23=19.56%

Table 2.10: The ranks and the p-values of the 23 target pathways for pDis analysis and wighted-pDis analysis. The improvement factor based on Fig. 2.3 is calculated for each dataset considering the 5% significance threshold using FDR-corrected p-values. The average improvement factor shows that wighted-pDis analysis improves 19.56% compared to pDis analysis. As shown, the average p-value and rank for the target pathways are lower (i.e. better) in weighted-pDis analysis than in pDis analysis.

affect each other. We investigated the results of our pathway analysis approach by considering the relations between pathways. We observed that some genes are included in more than one pathway and have interactions with different genes based on the pathways in which they exist. We looked at the global network of pathways as one single network. This global network is an integration of all the genes, and their associated interactions, which includes 5,052 nodes and 27,811 interactions. This approach is introduced in [18]. The primary disregulation for each gene is computed based on this global network, and then the score is summed to calculate the score of each pathway separately. The results are validated using the same procedure and datasets mentioned above.

The Figure 2.7 shows the distributions of the ranks and p-values of the target pathways in the integrated and original pDis analysis. The paired t-test and the paired Wilcoxon test were performed to compare the distributions of the ranks and p-values of target pathways in each method (t-test p-value$= 9.579e-07$, Wilcoxon p-value$= 0.001$). The results

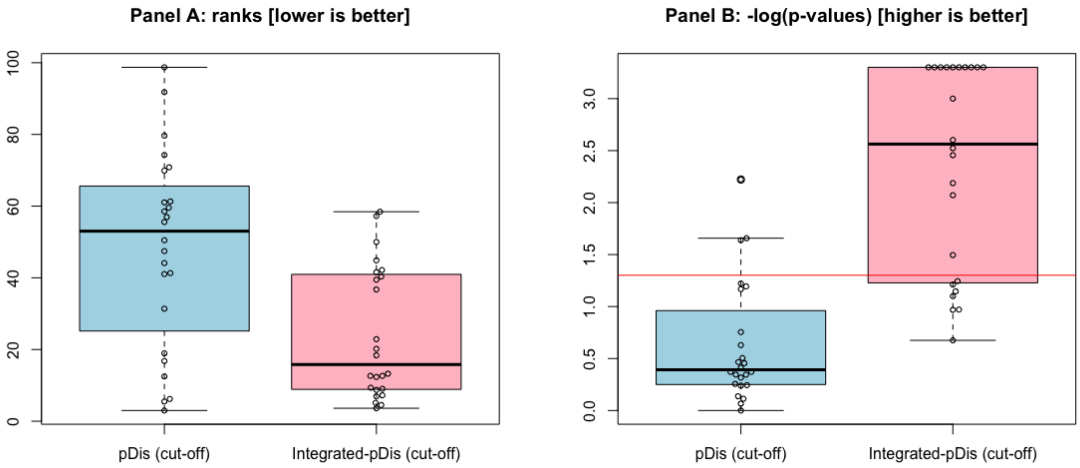


Figure 2.7: The ranks (in the left column, lower is better) and negative log of p-values of the target pathways (in the right column, higher is better) in the pDis and Integrated-pDis analysis. All human signaling pathways from KEGG (139 pathways) were used in the comparisons. The box-plots show the results obtained for the target pathways in the 24 datasets shown in Table 2.1. The bold line in the box-plots represents the median of the distribution. The red line represents the 5% significance level. These distributions show that the Integrated-pDis (in pink) performs better than original pDis analysis.

show that the p-values of the target pathways in Integrated-pDis analysis are significantly lower than original pDis analysis.

The details of the results from each method and the improvement factors of each dataset are shown in table 2.11.

2.5 Conclusion

Here, we proposed a new topological pathway analysis method based on the amount of perturbation associated with each individual gene. The proposed pDis analysis considers the dis-regulation of each gene in every pathway to calculate a p-value with respect to the distribution of the dis-regulation under the null hypothesis. The proposed method is able to use either i) a pre-selected number of DE genes, pDis analysis (cut-off), or ii) the entire list of measured expression levels, pDis analysis (all genes). The results showed that the proposed method yields significant improvements with respect to the state-of-the-art methods: SPIA, GSEA and GSA. Furthermore, we showed that the pDis analysis results are

	Target pathway	pDis (cut-off)			Integrated-pDis (cut-off)			Improvement Compared to pDis (cut-off)
		p-values	FDR	ranks	p-values	FDR	ranks	
1	Alzheimer's Disease	0.729	0.893	70.83	5e-04	0.005	5.12	+1
2	Alzheimer's Disease	0.022	0.348	6.20	5e-04	0.003	7.22	+1
3	Alzheimer's Disease	0.006	0.207	2.99	5e-04	0.005	4.51	+1
4	Alzheimer's Disease	0.068	0.842	18.94	5e-04	0.007	3.61	+1
5	Parkinson's disease	1.00	1.00	98.72	0.008	0.039	20.18	+1
6	Parkinson's disease	0.425	1.00	50.48	5e-04	0.002	12.34	+1
7	Huntington's disease	0.425	0.352	79.66	5e-04	0.003	6.92	+1
8	Colorectal Cancer	0.023	0.961	12.50	0.056	0.099	57.22	-0.5
9	Colorectal Cancer	0.351	0.217	44.12	0.031	0.070	44.87	-0.5
10	Colorectal Cancer	0.387	0.154	56.93	0.079	0.188	42.16	+0.5
11	Renal Cancer	0.482	0.147	61.31	5e-04	0.002	9.03	+1
12	Renal Cancer	0.859	1.00	91.79	0.061	0.154	39.45	+0.5
13	Pancreatic Cancer	0.451	0.573	59.56	0.0064	0.015	40.36	+1
14	Pancreatic Cancer	0.452	1.00	55.46	0.002	0.012	18.37	+1
15	Glioma	0.235	0.135	47.45	5e-04	0.002	12.56	+1
16	Glioma	0.342	0.360	61.03	5e-04	0.002	8.73	+1
17	Prostate Cancer	0.771	0.998	74.26	0.211	0.359	58.43	+0.5
18	Prostate Cancer	0.555	0.931	41.32	0.003	0.026	13.25	+1
19	Thyroid Cancer	0.570	0.790	58.52	0.106	0.213	50.00	+0.5
20	Thyroid Cancer	0.313	1.00	44.04	0.071	0.190	36.74	+0.5
21	Acute myeloid leukemia	0.064	0.795	1.20	0.002	0.80	22.89	-0.5
22	Non-Small Cell Lung Cancer	0.060	0.925	20.48	0.0009	0.93	12.65	+0.5
23	Non-Small Cell Lung Cancer	0.176	0.394	35.54	5e-04	0.39	9.33	+0.5
24	Dilated cardiomyopathy	0.577	1.00	69.89	0.107	0.258	41.56	+0.5
	Average	0.389	0.668	48.19	0.031	0.070	24.07	15.5/24=64.5%

Table 2.11: The ranks and the p-values of the 24 target pathways for pDis analysis and Integrated-pDis analysis. The improvement factor based on Fig. 2.3 is calculated for each dataset considering the 5% significance threshold using FDR-corrected p-values. The average improvement factor shows that Integrated-pDis analysis improves 64.5% compared to pDis analysis. As shown, the average p-value and rank for the target pathways are lower (i.e. better) in Integrated-pDis analysis than in pDis analysis.

significantly improved if the involvements of the gene in the pathways as well as, the relations between pathways are taken into consideration. The comparisons have been performed with a validation method that used 24 different datasets involving 12 different human diseases and eight different datasets involving eight knocked out genes in yeast.

CHAPTER 3: NEIGHBOR-NET ANALYSIS

3.1 Background

Understanding the mechanisms that cause changes in a phenotype requires identifying genes that are disrupted and relationships between them. Current technologies allow us to measure gene expression with unprecedented accuracy. Collecting such data across time enables the inferences of gene interactions (i.e., how a change in the expression of a gene affects other genes). There are many existing methods that extract information both from a single comparison (i.e., a steady state) [19, 55, 71] or multiple comparisons across time (i.e., time series) [157].

First attempts to discover gene regulatory networks started as soon as the first microarray experiments were published. Microarray technology measures the messenger RNA (mRNA) expression level of each gene. mRNAs are translated into proteins or transcription factors (TF), so the activity of a protein is estimated by the corresponding mRNA production which can be increased or decreased by the cell function. Assuming the mRNA expression of each gene is correlated to its protein product, we can use microarray data to infer the relations between genes and their products. Based on this assumption, the information from protein-protein interaction and protein-DNA interaction databases can also be used as a priori knowledge to predict the relations between genes [52, 78, 91]. These information can be obtained from different resources [12, 46, 81].

The number of proposed methods to infer network from gene expression data is growing each year, but there is no established approach to validate their results. There are some methods that use synthetic data to validate their results. The data is generated based on a known network, and the results are expected to have overlaps with the known source. However, many methods use real data to assess the proposed method. The ground truth in analyzing the real data is unknown, which makes it very difficult to evaluate the results. There are some methods that construct the gene regulatory network for simple organisms (e.g. yeast). The relationships between genes in yeast is well known, so the results will be

validated by being compared to the published literature based on how many true positives the method is able to infer. Other methods, using more complex organisms, compare the known interactions with their results and assess the methods based on the number of overlaps. Another evaluation approach, to find the accuracy of a method, is using the constructed networks as input for pathway analysis algorithms. The number of common genes between each pathway/gene set and the constructed network can be considered as a factor to score the pathways/gene sets. The most related pathways to the condition are expected to have significant number of common genes if the constructed networks are accurate.

In Table 3.12, features of different methods are summarized to show the advantages and disadvantages of their algorithms.

Reference	Type of data		Datasets	Physical interaction		Validation method	
	Steady states	Time-series		PPI	PDI	Real data	Synthetic data
Langfelder et al., 2008	√		Single			Mice	
Jiang et.al., 2008	√		Single			Human (colon cancer)	
Rhodes et.al., 2007	√		Multiple			Human	
Liu et.al., 2007	√		Single	√		Human (Type II Diabetes)	
Novershtern et.al., 2011	√	√	Single	√	√	Yeast	√
Idea et.al., 2002	√		Multiple	√	√	Yeast	
Zou et.al., 2005		√	Single			Yeast	

Table 3.12: Summary of the features in different methods for constructing gene regulatory networks. The "type of data" column shows if the method uses steady state or time-series data as an input dataset. The "datasets" column shows if the method uses single or multiple datasets for one condition to construct the network. The "physical interaction" column shows if the method uses protein-protein interaction (PPI) or protein-DNA interaction (PDI) as background knowledge. The"validation method" column shows if the method is evaluated by real or synthetic data, and "prior knowledge" column shows if the method uses known information to preselect the regulators or the target genes.

3.2 Motivation

Understanding how genes interact with each other is the key to understand the onset and evolution of a disease, for instance. Therefore, there is a need for computational algorithms able to construct network of active interactions by analyzing data. One of the drawbacks of the existing curated pathways or gene interactions databases is that each interaction is extracted from literature or experimentally validated by independent studies. However, most such interactions were found in specific tissues and/or phenotype, and not all studies employed the same tissue and/or phenotype.

Therefore, these independently identified interactions in the databases may not exist in actual phenotype or tissue studied in a subsequent experiment. Furthermore, new phenomena may be involved in the tissue or phenotype currently being studied. Utilizing only existing pathways from pathways databases or literature, limits one's ability to discover new phenomena and new interactions. In order to overcome such limitations, some existing methods try to build the regulatory networks based on the correlation or the co-expression existing in the given datasets [55, 71, 107, 157]. The networks resulting from such methods are specific to the condition under study, but the interactions identified are only based on the genes' expression level. This limitation can produce many false positives as well as false negatives, because an interaction between two genes is not necessarily reflected in the correlation between their expression levels. The interactions between genes can involve an indirect relation between them via their protein products or their transcription factors, and sometimes interactions take place on different time scales. Therefore, there is a need for computational algorithms able to construct network of active interactions by analyzing data in more sophisticated ways, by combining gene expression data with existing pathway information, as well as with data from protein-protein interactions databases.

However, the integration of high throughput datasets such as gene expression data with prior information about gene-gene interactions to find the networks specific to a pheno-

type is still an open challenge [66]. Such methods identify the network of interactions that is most relevant to a given phenotype based on the retrieved prior knowledge, referred to as "active network". This network is also known as "network hotspot" or "responsive subnetwork" [86]. The active network, as part of a global interaction network, explains the sudden changes in the genes activity or the characteristic of the phenotype in a given disease. This network is identified based on the given data and can be considered the putative mechanism involved in the given phenotype. The advantage of identifying active network is that it is specific to the condition studied, as opposed to existing curated pathways that can describe more generic knowledge, not necessarily applicable to the given condition.

The discovery of active network can lead to the identification of signature network that is associated to a given disease, rather than a set of gene biomarkers. This can lead to better understanding of the disease, diagnosis and more accurate treatment. Biomarker networks can also achieve more predictive power to classify different diseases such as diabetes or different types of cancer [86]. Furthermore, disease-specific networks can also be used to predict drug target mechanisms and to find the response of patients to the drugs.

Based on a comprehensive review published by [86] the existing approaches aiming to identify active network using prior knowledge of interactions, can be decided in three main categories, as follows. The first category is the "significant-area-search" methods. In this category the genes and interactions between them are scored based on the input data, and the algorithm tries to find the group of genes and interactions with the highest score. The very first methods in this category are jActiveModules [52] implemented in Cytoscape [116], and Gene Network Enrichment Analysis (GNEA) [78]. The second category includes "diffusion-flow" and "network-propagation" methods. The methods in this category attempt to find the flow between genes with maximum scores in the existing networks. They identify subset of genes and interactions that accumulate the highest score flows. The most widely used methods in this category are NetWalker [65, 66], HotNet [140], and Physical Module Networks [91]. The third category includes "cluster-based" methods. These methods use

biclustering algorithms to find the interactions that are active in the given conditions. The better known methods in this category are SAMBA [125] and SANDY [80].

In the following section, we propose a "network-propagation" algorithm that will identify the maximum flow between genes through their immediately connected genes. This method uses multiple steady state gene expression data that are collected from the same phenotype. The use of multiple datasets allows the proposed approach to capture changes of gene expression that might not be captured in any single dataset due for instance to the snapshot nature of gene expression data. Gene-gene interactions will be obtained from protein-protein interaction networks describing the relations between proteins and also from experimentally curated signaling pathways. This method will use the neighborhood of each gene to identify the propagation of disruption that flows through the system. The neighborhood of each gene includes the genes that are directly connected to it based on the known interaction networks.

We apply the method on multiple datasets from experiments studying colorectal cancer, renal cancer and prostate cancer. We assess the results in two ways: first, we assess the enrichment of known biological pathways in the constructed network. This validation process is similar to a gene set analysis approach introduced in [78]. In this reference [78], the number of common genes between the associated gene sets to the given phenotype and the identified network is used to validate the results. Similarly, we also consider here the number of interactions overlapping between the constructed networks and known signaling pathways that are associated to the disease investigated. Constructed networks that are significantly enriched in these interactions are considered better than those that are less enriched in such interactions. Second, we obtain a list of genes that are associated to the disease studied from DisGeNET [13, 101] and calculate the enrichment of each constructed network in such disease-associated genes. DisGeNET integrates information from multiple public data sources and literature (15 different resources) to identify gene-disease associations [101]. Networks that are significantly enriched in disease-associated genes are considered better

than those that are less enriched in such genes. We compare our result with the results of two widely used methods: NetWalker and HotNet, which also attempt to identify the relevant biological mechanisms on a global network. We also compare our results with the classical approach of considering just the union of differentially expressed (DE) genes in all considered studies. The results of these comparisons show that the proposed method performs better in constructing networks that are significantly relevant to the given diseases based on the resulting genes and interactions.

3.3 Neighbor-net analysis

The approach presented here aims at finding networks of genes that capture the mechanisms that could explain the phenotype. The algorithm requires two types of data. The first type is gene expression data that includes measured gene expression in the control samples versus the investigated disease. The proposed approach uses multiple datasets relevant to the same condition. This is important because each dataset is only a snapshot of the system captured at a given moment in time. The changes in expression levels of the genes will be better estimated by looking at multiple datasets, which represent different snapshots of the same condition, and therefore provide much more information. The second type of data is the prior information about gene-gene interactions that can be collected from different resources describing any known interactions between genes, such as protein-protein interaction, gene regulatory networks, and curated pathways. In addition, adding existing information about known interactions allows us to estimate the potential effect of a condition on groups of interacting genes that can ultimately constitute putative models of the mechanisms in action in the given condition. The integration of these two types of data allows us to overcome some of the limitations of existing methods.

The method starts with several existing datasets available for that disease. A list of differentially expressed (DE) genes is calculated as the union of all sets of DE genes from each such individual dataset.

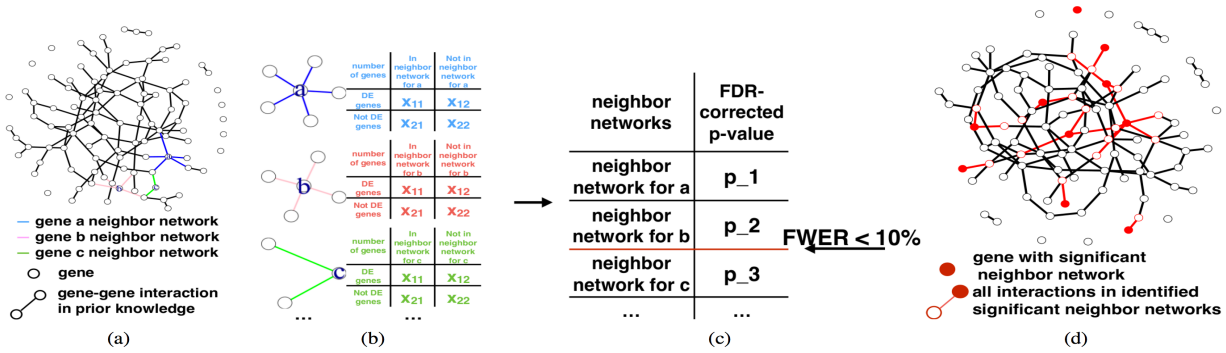


Figure 3.1: An overview of the proposed neighbor-net analysis. (a) The global network combining all the known gene-gene interactions. The colors show three sample neighbor networks for genes a, b and c. (b) The neighbor network for each gene is extracted from the global known interactions. A global list of differentially expressed (DE) genes is obtained by constructing the union of all genes found to be DE in at least one of the given datasets (based on the their calculated log fold changes and p-values adjusted for multiple comparison). The Fisher's exact test is performed on all extracted neighbor networks based on the number of DE genes they have. (c) The significant neighbor networks (FDR-corrected p-value lower than 10%) are identified. (d) The constructed network which is built by integrating all significant neighbor networks (shown in red).

The proposed approach then builds a "neighbor network" for each gene. The neighbor network associated to each gene includes the gene itself, the genes immediately connected to it and the interactions connecting them together based on known interactions from protein-protein interaction databases such as a HPRD [99, 100], as well as from pathway databases such as KEGG [58, 59]. This is done such that, even if none of the multiple datasets included in the analysis captures the effect of the gene of interest, by looking at its immediate neighborhood we can still detect changes that propagate from that gene. The neighbor network is constructed exclusively from annotation databases, independently of the DE genes. In the next step, we calculate the enrichment of each neighbor network based on the number of DE genes they contain. The hypergeometric p-value for each neighbor network is calculated based on the formula below:

$$p_o(x) = 1 - \sum_{i=0}^{x-1} \frac{\binom{M}{i} \cdot \binom{N-M}{K-i}}{\binom{N}{K}} \tag{3.1}$$

where N is the total number of genes, K is the total number of DE genes and M is the number of genes in each neighbor network. This p-value represents the probability of obtaining a number of DE genes in the neighbor network that is equal or higher than the number observed in the analysis, just by chance. Such p-values are computed for all neighbor networks and are corrected for multiple comparisons with false discovery rate (FDR) method. The significant neighbor networks, with FDR-corrected p-values lower than threshold (10%), are identified and combined together to build the final constructed network. The genes and interactions in the constructed network are the integration of all the genes and interactions extracted from all the identified significant neighbor networks. This constructed network resulted from the analysis can be considered as the active network that has the potential to capture the mechanisms involved in the given disease. A summary of the method, referred to as "neighbor-net analysis", is shown in Figure 3.1.

3.4 Discussion and results

We present the results of our analysis on three different diseases (colorectal cancer, renal cancer and prostate cancer). The method requires a set of datasets studying the same disease under the same conditions. We selected three diseases with three or more datasets available in GEO associated to each of the three cancer types. The results are compared with the results obtained with three other approaches: HotNet [140], NetWalker [65, 66], and the classical ORA [62]. The detailed results of each disease are shown in separate sections. The gene expression data in each case is obtained from Gene Expression Omnibus (GEO) [9]. Using GEO2R [113], we compute a fold change and a p-value for each gene representing the significance of the observed change in the expression levels between normal and disease samples. The information about gene interactions is obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Release 72.0) [57, 58, 59] and Human Protein database Reference (HPRD) (Release 9) [99, 100]. We chose to illustrate the proposed approach using KEGG because it is well recognized and widely used (almost 7,800 citations) but this

approach is completely independent of the pathway database used and can be used equally well with Reactome [27, 56], BioCarta [90], etc. Similarly, the approach can be used with any other protein-protein interaction database such as BioGRID [119], BIND [3], etc.

We collapsed all 173 available KEGG signaling pathways into one network with 5,052 nodes and 27,811 interactions. The protein-protein interaction (PPI) network downloaded from HPRD includes 9,672 nodes and 39,233 interactions. Combing these two networks results a global graph with 11,086 nodes and 62,934 interactions. The neighbor networks are built for each gene and the associated p-values are computed to represent the significance of their enrichment in the list of DE genes. The constructed network for each case study is the integration of all significant neighbor networks.

There is no universally accepted validation method to assess the accuracy of the constructed networks. Similar to a previous study [78], we evaluate the results by performing pathway enrichment analysis based on the edges overlapping between the constructed network and pathways to identify what known biological mechanisms are captured. Such analysis allows us to validate the parts of the constructed network that are obtained from existing pathways known to be involved in the given disease. The hypergeometric p-value for each pathway is computed based on the number of common edges it has with the constructed network. The ranked list of significant pathways for each case study will show which pathways are more enriched in the constructed network. For each disease, we consider the pathway that was created in order to explain that particular disease as the "target pathway" (e.g. *Colorectal cancer pathway* is the target pathway for colorectal cancer). We validate the constructed network using the rank of target pathway in the reported list of pathways sorted by p-values corrected for multiple comparisons. The constructed network is more relevant to the investigated disease if the rank of the target pathway is lower and the corresponding p-value is more significant [126]. Note that other truly impacted pathways may be present in this list for legitimate reasons.

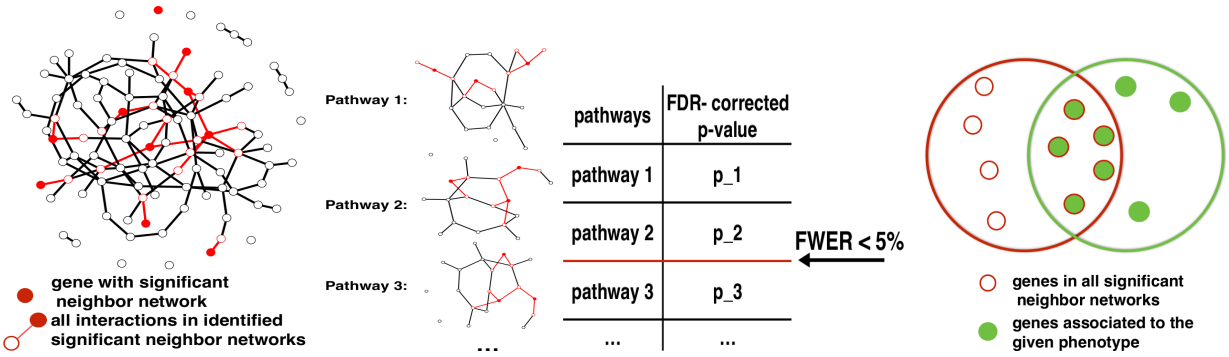


Figure 3.2: An overview of the two evaluation processes for the constructed network. (a) The constructed network which is built by integrating all significant neighbor networks (shown in red). (b) We start the evaluation process by looking at the enrichment of each KEGG pathway in edges also present in the constructed network. The red edges in each pathway represent the edges overlapping between that pathway and the constructed network. The significance of enrichment of each pathway is calculated based on the number of such edges. (c) Pathways are ranked using their enrichment p-values. (d) The second evaluation process calculates an enrichment p-value for each constructed network. This p-value characterizes the enrichment of each constructed network in genes that are known to be associated to the investigated disease based on DisGeNET. Lower p-values represent a significant enrichment of the constructed network in nodes known to be associated with the disease.

In addition, we also evaluate the results by using the DisGeNET database [13, 101] to determine how many genes in the network are known to be related to the given condition. A p-value representing the significance of the number of identified genes known to be associated to the investigated disease in the constructed network is computed. A lower p-value will mean that the method identifies genes that are more relevant to the disease and therefore, that the constructed network is more likely to describe the mechanism involved in that disease. Both evaluation approaches are summarized in Figure 3.2. The results described in the following sections show that the proposed method performs better in both evaluation processes in comparison to all reference methods.

3.4.1 Colorectal cancer

We analyze five gene expression datasets studying colorectal cancer from GEO [9] (GSE4173, GSE9348, GSE21510, GSE32323 and GSE8671). A global list of differentially expressed (DE) genes is obtained by selecting the genes with an absolute value of log2 fold

change higher than 1.5 and adjusted p-value lower than 0.01 in at least one dataset. The union of DE genes includes 2,968 genes out of 19,852 total genes in the five experiments. We perform our analysis on selected DE genes. Based on the calculated hypergeometric p-values for all neighbor networks, 20 of them are significantly enriched in the given list of DE genes. The constructed network is a global graph that integrates all the significant neighbor networks. It includes 144 genes and 251 interactions and is shown in Figure 3.3. This network can be seen as most likely to include the mechanisms involved in colorectal cancer.

The edges overlapping between each KEGG signaling pathway and the constructed network are identified. The probability of having the observed number of these edges just by chance is calculated for every KEGG pathway. The list of KEGG pathways that are significantly enriched in edges from the constructed network is shown in Table 3.13. This table shows that the constructed network includes a significant number of edges from the target pathway. In fact, the *Colorectal cancer pathway* is ranked 3rd and it has a significant number of edges in common with the constructed network. The two other pathways that are ranked higher than the *Colorectal cancer pathway* are the *Hippo signaling pathway* and the *Wnt signaling pathway*. There is a extensive evidence that both *Hippo signaling pathway* [10, 20, 94], as well as *Wnt signaling pathway* [15, 23, 106] are very important in colorectal cancer.

Figure 3.4 shows the edges from the network built by the proposed algorithm, as they appear in the context of existing KEGG pathways. Some of the gene interactions (edges) in this figure, appear multiple times in various significant KEGG pathways.

Interestingly, these edges from the network built by the proposed neighbor-net analysis describe a well known mechanism known to be involved in colorectal cancer. The β-catenin protein is a very well-known protein that has important impacts on developing colorectal cancer [88]. It is produced by gene "CTNNB1". This gene is one of the parent nodes in the green network present in 9 out of the 10 significant pathways shown in Figure 3.4. Notably, this gene is not identified as a differentially expressed gene by classical approaches

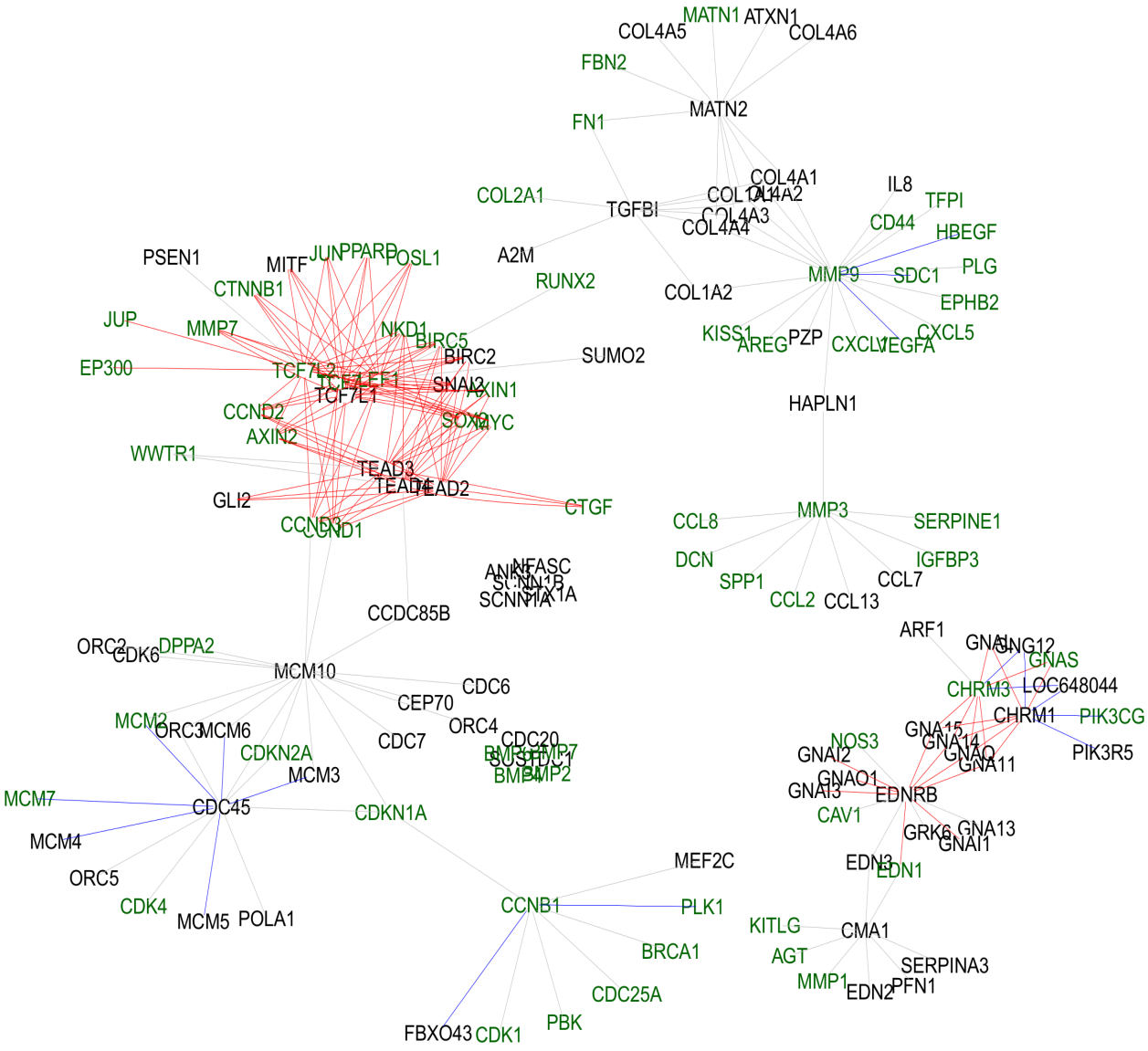
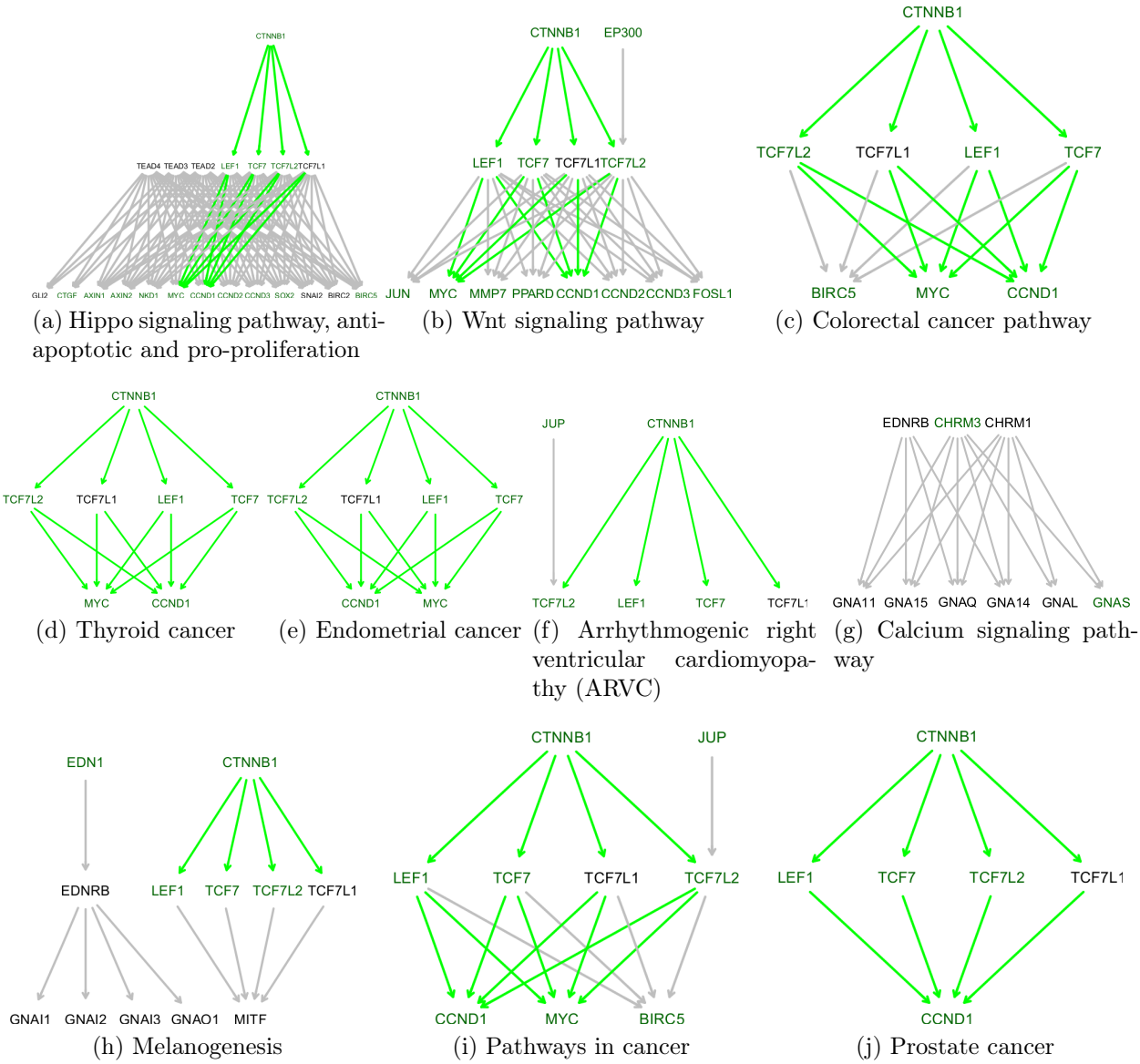


Figure 3.3: The active network that describes the putative mechanisms involved in colorectal cancer. The five subnetworks shown above include 144 nodes and 251 edges. The 130 red edges represent the interactions that exist in significantly enriched KEGG pathways. The 17 edges shown in blue are present in KEGG pathways that are not significantly enriched in edges from this network. The 70 genes shown in green are known to be associated to colorectal cancer based on the DisGeNET database.



(a) Hippo signaling pathway, anti-apoptotic and pro-proliferation

(b) Wnt signaling pathway

(c) Colorectal cancer pathway

(d) Thyroid cancer

(e) Endometrial cancer

(f) Arrhythmogenic right ventricular cardiomyopathy (ARVC)

(g) Calcium signaling pathway

(h) Melanogenesis

(i) Pathways in cancer

(j) Prostate cancer

Figure 3.4: The edges overlapping between the identified active network in colorectal cancer and each significantly enriched KEGG signaling pathways. The green genes represent the genes that are associated to colorectal cancer based on DisGeNET database and green edges represent the common edges in more than 33% of the significant pathways. The common network with gene "CTNNB1" as a parent node is found in 9 out of 10 significant pathways. This gene produces β-catenin protein which is one of the important proteins that trigger colorectal cancer.

significant pathway	FDR-corrected p-value	references
Hippo signaling pathway	2.1e-104	[10, 20, 94]
Wnt signaling pathway	5.5e-22	[15, 23, 106]
Colorectal cancer pathway	**3.7e-17**	
Thyroid cancer	1.8e-15	
Endometrial cancer	2.4e-12	
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	6.4e-08	
Calcium signaling pathway	5.7e-07	[69]
Melanogenesis	7.05e-06	
Pathways in cancer	2.8e-03	
Prostate cancer pathway	4.3e-03	

Table 3.13: A list of significantly enriched pathways (FDR-corrected p-value < 0.05). These pathways are significantly enriched in the network resulted from neighbor-net analysis. The bold pathway is the target pathway in colorectal cancer. The third column shows the references explaining the association of the respective pathways to colorectal cancer.

in any of the datasets so the classical approach of focusing on differentially express genes would not be able to identify this mechanism. The interactions between "CTNNB1" and its downstream genes,"LEF1", and "TCF7L1", are part of the network built by the proposed approach, network that is present in most of the significant KEGG pathways. These genes are immediately connected to other genes such as "MYC", "CCND1", and "BIRC5" that have important roles in the evolution of colorectal cancer through a number of cell functions (e.g. proliferation, apoptosis) [75, 88].

We also compare our result with the results produced by NetWalker and HotNet. Both methods are widely used to construct networks of genes that are meant to describe the active modules in a given phenotype. NetWalker is built based on an algorithm introduced in [65, 66]. It accepts as input a list of all genes in the analysis but it requires the selection of a specific group as "seeds". We selected as seeds the genes that are differentially expressed (fold change higher than 1.5 and adjusted p-value lower than 0.01) in at least one of the datasets. NetWalker uses multiple resources such as KEGG, REACTOME interactions, and literature based gene regulatory networks, as prior knowledge. The output of this method is a network in which nodes represent genes, and edges represent interactions between them.

This network is claimed to explain the mechanisms involved in the investigated disease. The result of NetWalker for colorectal cancer datasets is a network that includes 901 genes and 3,028 interactions. The p-value representing the significance of the number of edges overlapping between this network and *Colorectal cancer pathway* is 0.99 (see Table 3.14). This p-value shows that the network constructed by NetWalker is not overlapping in any significant way with the KEGG pathway that describes the phenomena involved in this type of cancer.

We also compare the results with HotNet [140] that also constructs a network from known protein-protein interaction (PPI) network by considering the degree (number of links) of each gene together with a gene's score that shows the significance of change in its expression level. We use the genes' negative log of p-values as their associated scores. The minimum p-value for each gene in five datasets is used to compute the scores. HotNet requires a threshold for selecting the important networks. The authors provide an algorithm that suggests 5 different thresholds for the given data. We use the minimum threshold suggested by the algorithm. The constructed network includes 215 genes and 175 interactions. The p-value representing the enrichment of the target pathway in the constructed network is 1.0 (see Table 3.14). In other words, the network constructed by HotNet does not overlap in any significant way with the colorectal cancer pathway from KEGG, suggesting that this constructed network does not capture many of the phenomena considered to be central to the colorectal cancer development by the KEGG's authors.

We also compare the enrichment of the target pathway in the network constructed from the proposed method with the results of over representation analysis (ORA) which is a classical pathway analysis method [62, 87]. ORA takes into consideration the number of DE genes observed in each pathway and calculates the probability of observing this number just by chance. The analysis is performed on the union of DE genes in all datasets considered. The rank and the p-value of the target pathway representing its enrichment in the list of DE genes is calculated and shown in Table 3.14.

Colorectal cancer		
method	rank	FDR-corrected p-value
neighbor-net analysis	**3**	**3.7e-17**
NetWalker	22	0.99
HotNet	95	1.0
ORA	96	0.37

Table 3.14: The ranks and p-values of the target pathway (*Colorectal cancer pathway*) in neighbor-net analysis and three other methods. The p-values represent the significance of the enrichment of the *Colorectal cancer pathway* in the identified active network. The comparisons show that neighbor-net analysis reports the target pathway more significant and highly ranked.

We also evaluate the results by assessing the number of nodes in each constructed network that are known to be associated to the investigated disease. We compare the genes in the identified active network with the genes known to be associated to colorectal cancer, obtained from the DisGeNET database. The number of associated genes to colorectal cancer from this database is 2,277. Based on the extracted associated genes 70 genes out of 144 genes reported by the neighbor-net analysis are known to be associated to colorectal cancer (48%). The percentage of the genes associated to the colorectal cancer in the network constructed by neighbor-net analysis is higher compared to all the reference methods. This means that the neighbor-net analysis is able to identify a higher proportion of genes related to colorectal cancer in comparison to other methods (see Table 3.15). Also, the computed p-value that represents the significance of enrichment of such genes in the constructed network in neighbor-net analysis is also highly significant (8.4e-15).

The p-values of observing the number of associated genes to colorectal cancer, just by chance, as well as the percentages of such genes in the lists of genes reported by NetWalker, HotNet and selected DE genes are shown in Table 3.15.

The lower p-values represent more significant enrichment of the genes associated to colorectal cancer in the constructed network. The p-values show that the selected DE genes and the list of genes resulted from NetWalker are also significantly enriched in the genes associated to colorectal cancer. However, the percentage of such genes in the total list of

Colorectal cancer				
method	#selected genes	#colorectal cancer genes	%	p-value
neighbor-net analysis	**144**	**70**	**48%**	**8.4e-15**
NetWalker	901	283	31%	5.5e-18
HotNet	215	46	21%	0.23
Selected DE genes	2968	552	18%	5.8e-27

Table 3.15: The statistical analysis of the results from neighbor-net analysis and all the methods compared in colorectal cancer. The columns show: the number of genes in the identified active network reported by each method, the number of associated genes to colorectal cancer based on information obtained from DisGeNET [101], the percentages of the genes known to be associated to colorectal cancer in the total number of identified genes in each constructed networks, and the corresponding p-values for the enrichment in each method. The p-value of observing the given number of genes that are associated to colorectal cancer in the constructed network is highly significant in the neighbor-net analysis. The percentage of the associated genes in the constructed network in also higher compared to all three existing methods.

identified genes by neighbor-net analysis is higher compared to all other methods. The statistical analysis shows that neighbor-net analysis performs better than the existing methods in both evaluation approaches. Essentially, the neighbor-net analysis is able to find more associated genes to colorectal cancer compared to NetWalker, HotNet and the enrichment approach, as well as more interactions relevant to colorectal cancer based on number of overlaps with the *Colorectal cancer pathway*.

3.4.2 Renal cancer

We analyze three gene expression datasets studying renal cancer from GEO (GSE14762, GSE6344, and GSE781). A list of differentially expressed genes is obtained by selecting the genes with an absolute value of log2 fold change higher than 1.5 and adjusted p-value lower than 0.01 in at least one dataset. The union of these DE genes includes 1,223 genes out of 18,113 total genes in the three experiments. We perform our analysis on selected DE genes. Based on the calculated hypergeometric p-value for all neighbor networks, 69 of them are significantly enriched in the given list of DE genes. The constructed network, which is an integration of all significant neighbor networks, includes 663 genes and 1,552 interactions (see Figure 3.5).

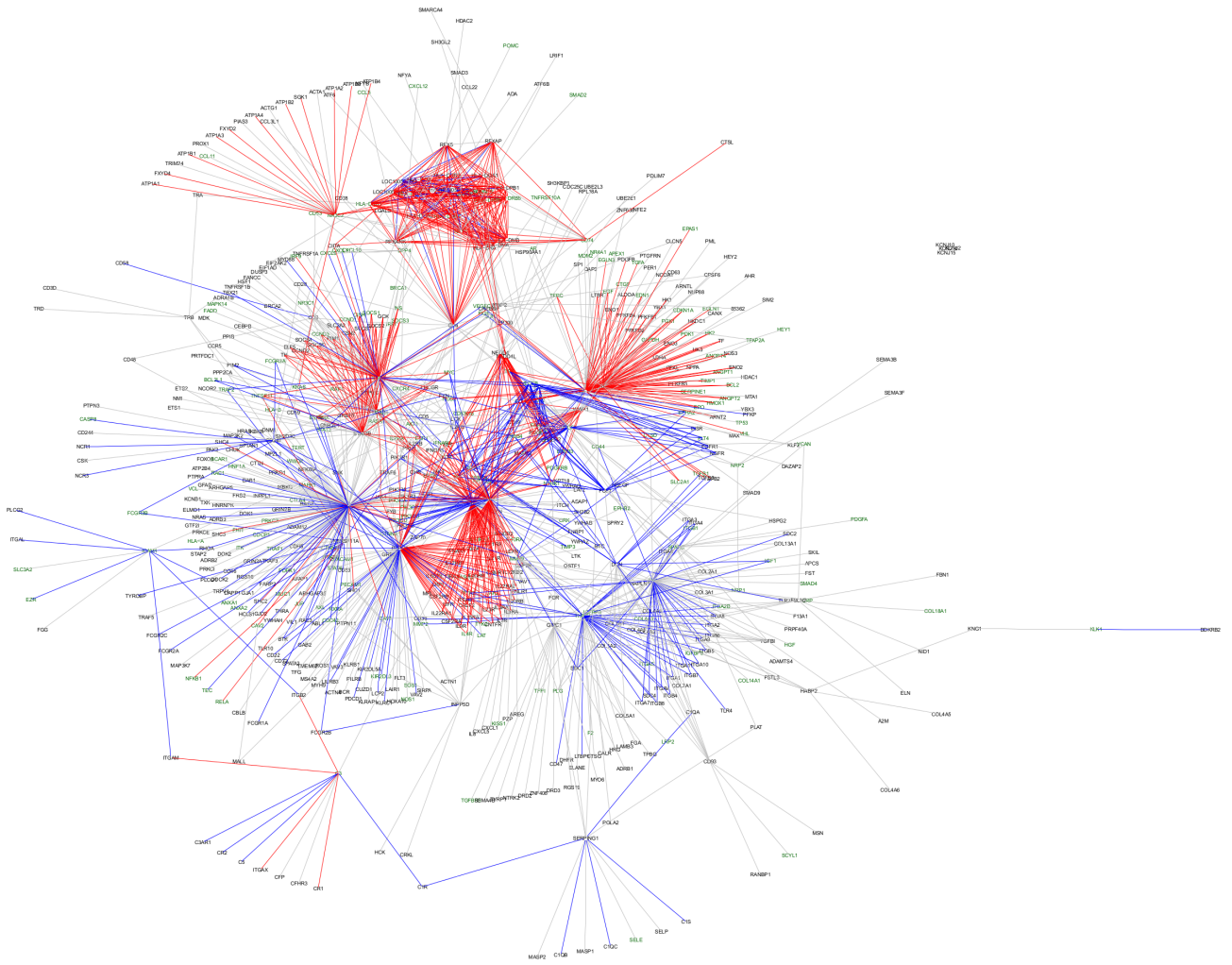
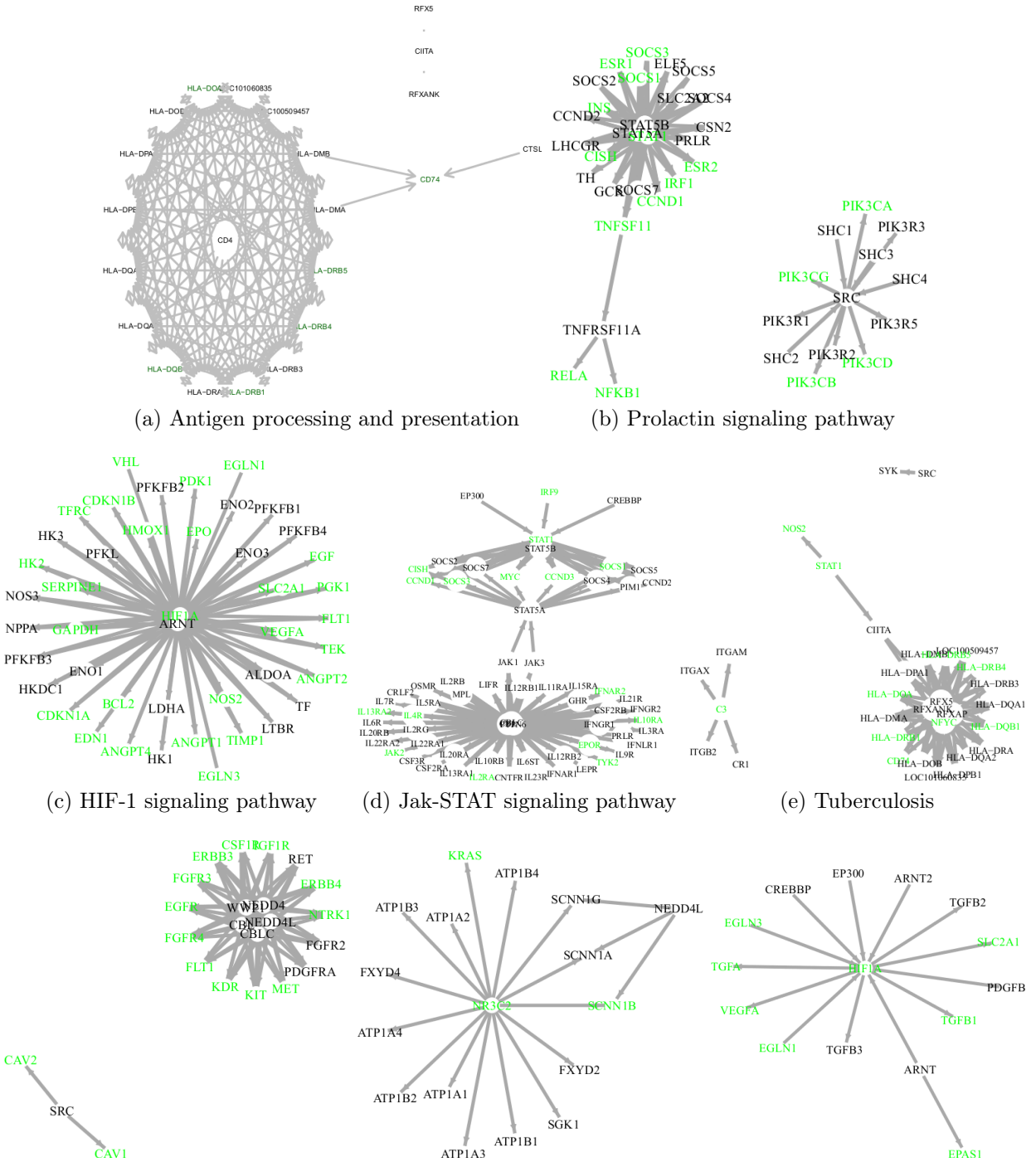


Figure 3.5: The active network that describes the putative mechanisms involved in renal cancer. The two subnetworks shown above include 663 nodes and 1,552 edges. The colored edges represent the interactions that exist in KEGG signaling pathways. The 650 red edges represent the interactions that exist in significantly enriched KEGG pathways. The 320 blue edges are present in KEGG pathways that are not significantly enriched in edges from this network. The 198 genes shown in green are known to be associated to renal cancer based on the DisGeNET database.

We apply the same evaluation approaches for this case study. The list of KEGG pathways that are significantly enriched in edges from the constructed network is shown in Table 3.16. The "Renal cell carcinoma pathway", which is the target pathway in renal cancer, is reported as significant. Most of the significant pathways that are ranked higher than the target pathway are also very important in developing renal cancer. (*Antigen processing and presentation* [1, 114], *HIF-1 signaling pathway* [77, 104], *Jak-STAT signaling pathway* [43, 123], *Endocytosis pathway* [33, 92] and *Aldosterone-regulated sodium reabsorption pathway* [8]).

Figure 3.6 shows the edges from the network built by the proposed algorithm, as they appear in the context of existing KEGG pathways. One of the hub genes in the constructed network is gene "HIF1A" that produces protein "HIF-α". This gene also appears in "Renal cell carcinoma pathway". HIF1A is recognized as having an important function in development of renal cancer [41, 139]. However, it is not identified as a differentially expressed gene by classical approach in any of the three studying datasets. The interactions between this gene and its downstream genes, which are present in the constructed network, play important roles in renal cancer [47]. These interactions also appear in two of the significant KEGG pathways. Identifying such interactions in the constructed network highlights the fact that the constructed network resulted from neighbor-net analysis is very relevant to renal cancer and can indeed capture the mechanisms involved in development of this type of cancer.

We compare the result of the neighbor-net analysis with the results of the classical approach of selecting DE genes, NetWalker, and HotNet. The ranks and the p-values representing the significance of the enrichment of the target pathway in all the methods are shown in Table 3.17. The rank of "Renal cell carcinoma pathway" is much lower (8th compared to 94th and higher), and its corresponding p-value is more significant ($2.6e-03$ compared to 0.51 and higher) in neighbor-net analysis compared to all reference methods. We also compare the genes in the constructed network with the genes known to be associated to renal cancer obtained from DisGeNET database. The number of associated genes to renal



(a) Antigen processing and presentation (b) Prolactin signaling pathway

(c) HIF-1 signaling pathway (d) Jak-STAT signaling pathway (e) Tuberculosis

(f) Endocytosis (g) Aldosterone-regulated sodium reabsorption (h) Renal cell carcinoma

Figure 3.6: The edges overlapping between the identified active network in renal cancer and each significantly enriched KEGG signaling pathways. The green genes represent the genes that are associated to renal cancer based on DisGeNET database. Gene "HIF1A", which is a hub node in the constructed network and exists in pathways *HIF-1 signaling pathway* and *Renal cell carcinoma* has a known important function in development of renal cancer. All the hub genes in each significant pathways have important impacts in renal cancer (see references in Table 3.16).

significant pathway	FDR-corrected p-value	references
Antigen processing and presentation	1.1e-244	[1, 114]
Prolactin signaling pathway	2.4e-52	
HIF-1 signaling pathway	1.1e-39	[77, 104]
Jak-STAT signaling pathway	6.8e-39	[43, 123]
Tuberculosis pathway	1.3e-23	
Endocytosis pathway	8.4e-23	[33, 92]
Aldosterone-regulated sodium reabsorption pathway	5.4e-13	[8]
Renal cell carcinoma pathway	**2.6e-03**	-

Table 3.16: A list of significantly enriched pathways (FDR-corrected p-value < 0.05) in renal cancer. These pathways are significantly enriched in the network resulted from neighbor-net analysis. The bold pathway is the target pathway in renal cancer. The last column shows the references explaining the association of the respective pathways to renal cancer.

cancer from this database is 1,220. Based on the extracted genes, 199 genes out of 663 genes in the constructed network are known to be associated to renal cancer (30%). The percentage of such genes in the total list of resulted genes in the neighbor-net analysis is higher compared to all reference methods. The p-values of observing the number of genes associated to renal cancer in the list of genes reported by each method just by chance are shown in Table 3.18. The table shows the results of the proposed method have both the most significant p-value, as well as the highest percentage of genes known to be associated with renal cancer. In summary, the comparisons show that the proposed method is able to find more known renal cancer genes in the constructed network compared to NetWalker, HotNet and classical approach. The proposed method also identifies more interactions related to renal cancer based on number of overlaps with the *Renal cancer pathway*.

3.4.3 Prostate cancer

We analyze four gene expression datasets studying prostate cancer from GEO (GSE6956 African and Caucasian, GSE55945, and GSE45016). A list of differentially expressed genes is obtained by selecting the genes with an absolute value of log2 fold change higher than 1 and adjusted p-value lower than 0.05 in at least one dataset. The union of these DE genes includes 2,305 genes out of 19,851 total genes in the four experiments. We perform

Renal cancer		
method	rank	FDR-corrected p-value)
neighbor-net analysis	**8**	**2.6e-03**
NetWalker	106	1.0
HotNet	96.5	1.0
ORA	94	0.52

Table 3.17: The ranks and p-values of the target pathway (renal cancer) in neighbor-net analysis and three other methods. The p-values represent the significance of the enrichment of the renal cancer pathway in the identified active network. The comparisons show that neighbor-net analysis reports the target pathway more significant and highly ranked.

Renal cancer				
method	#selected genes	#renal cancer genes	%	p-value
neighbor-net analysis	**663**	**198**	30%	**1.9e-41**
NetWalker	424	82	19%	6.7e-07
HotNet	329	62	18%	2.8e-05
Selected DE genes	1223	189	15%	3.2e-22

Table 3.18: The statistical analysis of the results from neighbor-net analysis and all the methods compared in renal cancer. The columns show: the number of genes in the identified active network reported by each method, the number of associated genes to renal cancer based on information obtained from DisGeNET [101], the percentages of the genes known to be associated to renal cancer in the total number of identified genes in each constructed networks, and the corresponding p-values for the enrichment in each method. The p-value of observing the given number of genes that are associated to renal cancer in the constructed network is highly significant in the neighbor-net analysis. The percentage of the associated genes in the constructed network in also higher compared to all three existing methods.

significant pathway	FDR-corrected p-value	references
Focal adhesion	4.0e-30	[68, 122]
Estrogen signaling pathway	2.4e-18	[26, 115]
Endocytosis	1.5e-17	[39, 137]
ErbB signaling pathway	4.3e-14	[42]
Adherens junction	1.5e-13	[21, 103]
Glioma	1.1e-11	
Ras signaling pathway	2.6e-09	[6, 96, 152]
Choline metabolism in cancer	2.5e-08	[28, 37]
Proteoglycans in cancer	7.6e-08	[34, 40]
Rap1 signaling pathway	1.0e-07	[4]
Prostate cancer pathway	**2.4e-07**	-

Table 3.19: Top 11 significant pathways (FDR-corrected p-value < 0.05) in prostate cancer. These pathways are significantly enriched in the network resulted from neighbor-net analysis. The bold pathway is the target pathway in prostate cancer. The third column shows the references explaining the association of the respective pathways to prostate cancer.

our analysis on selected DE genes. Based on the calculated hypergeometric p-value for all neighbor networks, 23 of them are significantly enriched in the given list of DE genes. The constructed network includes 526 genes and 807 interactions between them (see Figure 3.7).

Here, we apply the same validation approaches explained in previous sections. The list of pathways significantly enriched in the constructed network is shown in Table 3.19. The *Prostate cancer pathway*, which is the target pathway in this condition, is reported as significant. Most of the pathways that are ranked higher than *Prostate cancer pathway* have important impacts in prostate cancer (*Focal adhesion* [68, 122], *Estrogen signaling pathway* [26, 115], *Endocytosis* [39, 137], *ErbB signaling pathway* [42], *ErbB signaling pathway* , *Adherens junction* [21, 103], *Ras signaling pathway* [6, 96, 152], *Choline metabolism in cancer* [28, 37], *Proteoglycans in cancer* [34, 40], *Rap1 signaling pathway* [4]).

The common edges between the identified active network and significant KEGG pathways are shown in Figure 3.8. The interactions between "EGFR" and its downstream genes such as "MAPK" and "PIK3" are very well known mechanism to be associated to prostate cancer [30, 36, 39, 64]. Such interactions, which are shown in green in Figure 3.8, appear in 7 out of top 10 significant KEGG pathways that are ranked higher than prostate cancer.

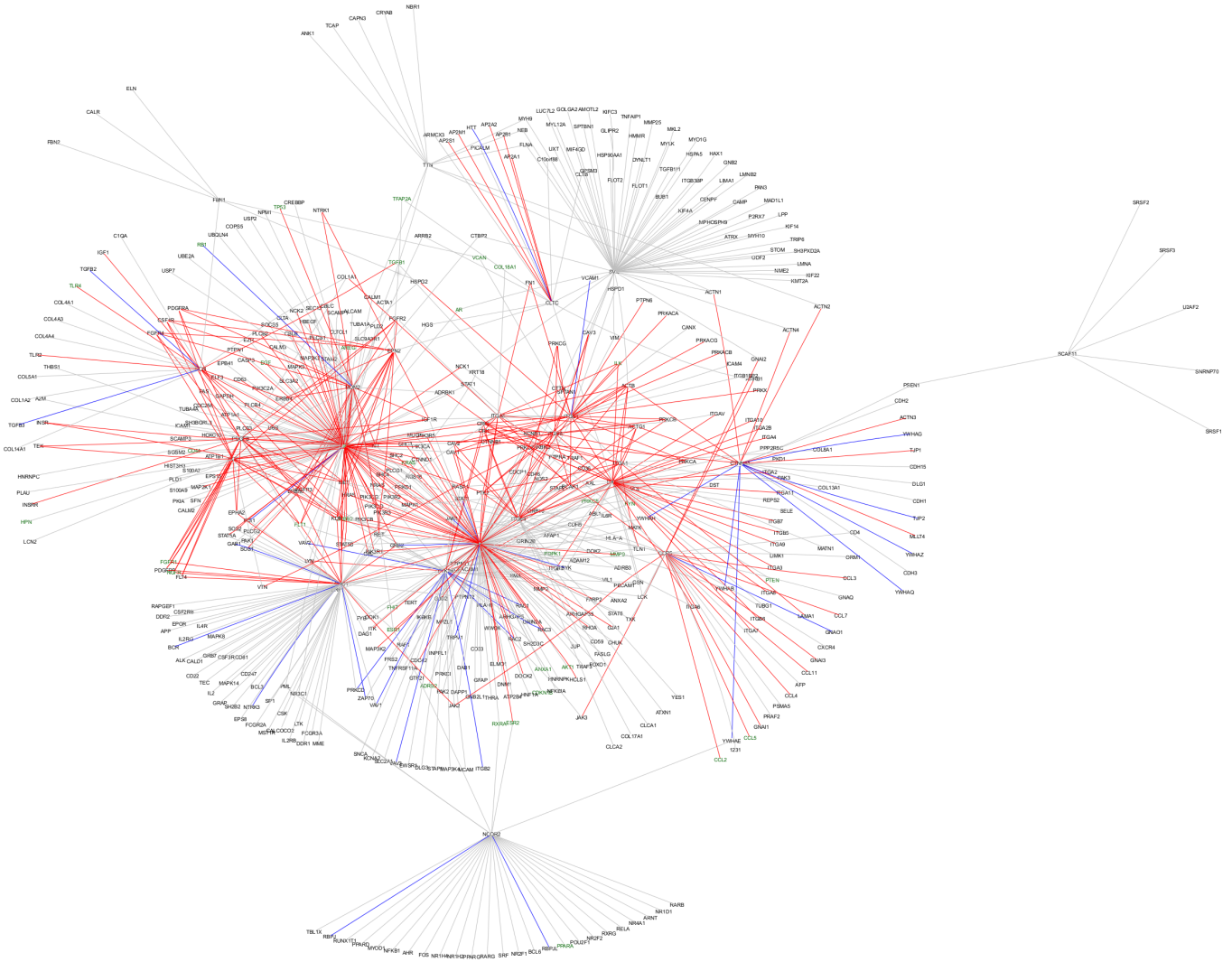
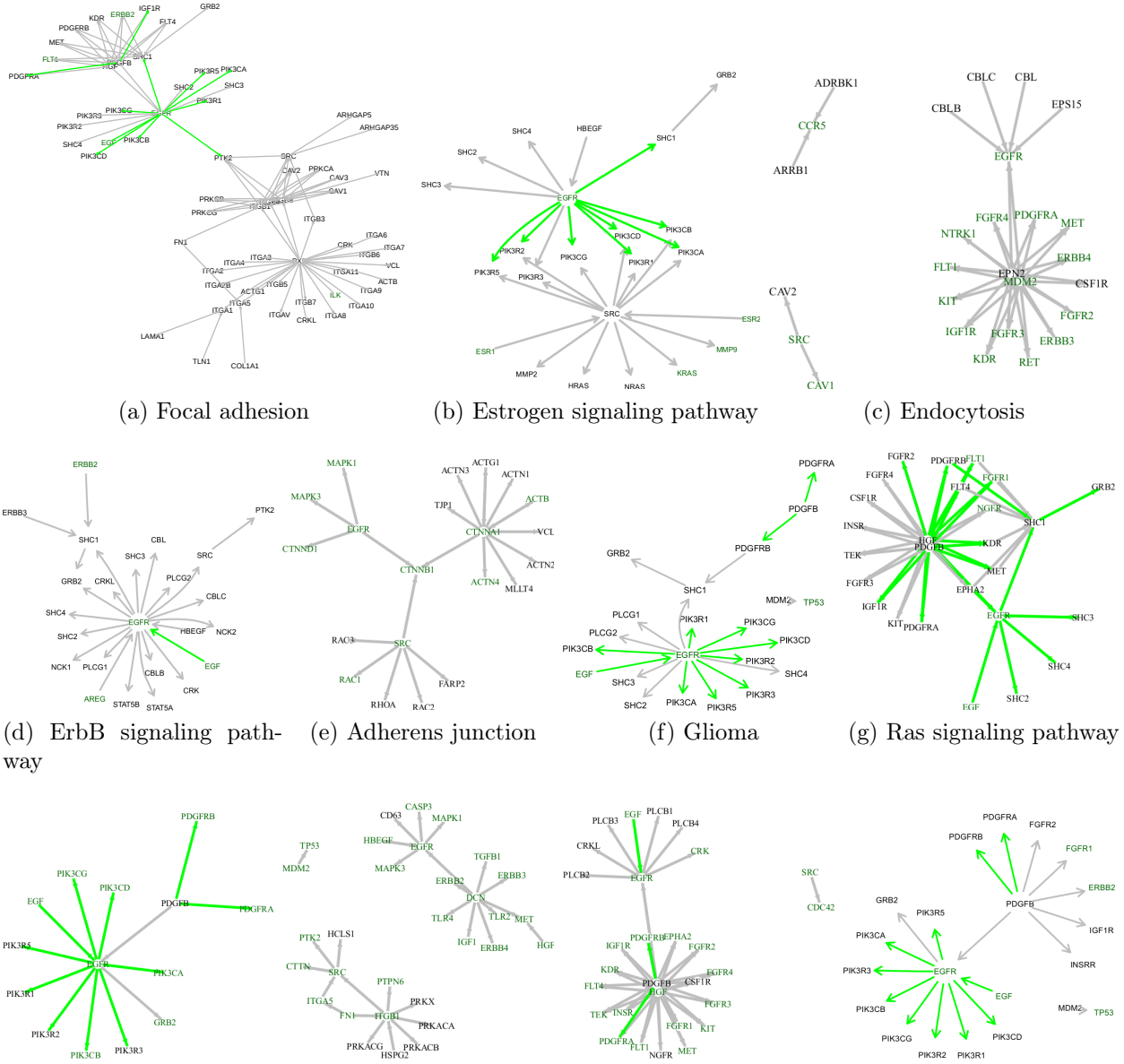


Figure 3.7: The active network that describes the putative mechanisms involved in prostate cancer. The network shown above includes 526 nodes and 807 edges. The 308 red edges represent the interactions that exist in significantly enriched KEGG pathways. The 33 blue edges are present in KEGG pathways that are not significantly enriched in edges from this network. The 35 genes shown in green are known to be associated to prostate cancer based on the DisGeNET database.



(a) Focal adhesion (b) Estrogen signaling pathway (c) Endocytosis

(d) ErbB signaling pathway (e) Adherens junction (f) Glioma (g) Ras signaling pathway

(h) Choline metabolism in cancer (i) Proteoglycans in cancer (j) Rap1 signaling pathway (k) Prostate cancer

Figure 3.8: The edges overlapping between the identified active network in prostate cancer and each significantly enriched KEGG signaling pathways. The green genes represent the genes that are associated to prostate cancer based on DisGeNET database and green edges represent the common edges in more than 33% of the significant pathways. EGFR is a hub node in all significant pathways. This gene and its identified interactions with downstream genes such as PIK3 and MAPK have very strong association with prostate cancer. Such interactions occur in most of the significant pathways.

Prostate cancer		
method	rank	FDR-corrected p-value
neighbor-net analysis	**11**	**2.4e-07**
NetWalker	107	1.0
HotNet	92.5	1.0
ORA	51	0.001

Table 3.20: The ranks and p-values of the target pathway (*Prostate cancer pathway*) in neighbor-net analysis and three other methods. The p-values represent the significance of the enrichment of the *Prostate cancer pathway* in the identified active network. The comparisons show that neighbor-net analysis reports the target pathway more significant and highly ranked.

We also compare the genes in the constructed network with the genes that are associated to prostate cancer obtained from DisGeNET database. In DisGeNET, there are 148 genes that are associated to prostate cancer. There are 35 genes out of 526 in the network constructed by the neighbor-net analysis that are associated to prostate cancer. We compare the results of the neighbor-net analysis with the results of the NetWalker, HotNet and classical approach of selecting DE genes. The details of comparison between the proposed and the reference methods are shown in Tables 3.20, 3.21. The p-value of observing the given number of genes known to be associated to prostate cancer in the constructed network is more significant, as well as the target pathway is ranked higher and the corresponding p-value is more significant in neighbor-net analysis compared to NetWalker, HotNet and classical approach. The comparisons show that the proposed method is able to report more associated genes to prostate cancer in the constructed network compared to all reference methods. It also identifies more related interactions to this disease based on number of overlaps with the known *Prostate cancer pathway*.

3.4.4 Results of neighbor-net using interactions from BioGRID

A reasonable question might be posed regarding the degree to which the results obtained by the proposed approach depends on the source of annotation, for instance, on the particular database used for protein-protein interactions. In order to investigate this, we also performed the neighbor-net analysis by using BioGRID [119] instead of HPRD, as

Prostate cancer				
method	#selected genes	#prostate cancer genes	%	p-value
neighbor-net analysis	**526**	**35**	6%	**1.8e-14**
NetWalker	379	13	3%	0.003
HotNet	283	1	0.3%	0.98
Selected DE genes	2305	37	1%	0.0004

Table 3.21: The statistical analysis of the results from neighbor-net analysis and all the methods compared in prostate cancer. The columns show: the number of genes in the identified active network reported by each method, the number of associated genes to prostate cancer based on information obtained from DisGeNET [101], the percentages of the genes known to be associated to prostate cancer in the total number of identified genes in each constructed networks, and the corresponding p-values for the enrichment in each method. The p-value of observing the given number of genes that are associated to prostate cancer in the constructed network is highly significant in the neighbor-net analysis. The low percentages is due to the fact that the total number of genes associated to prostate cancer obtained from DisGeNET is only 148; however, the neighbor-net analysis includes higher percentage of such genes in the constructed network compared to all three existing methods.

the protein-protein interactions resource. Same as before, we collapsed all 173 available KEGG signaling pathways into one network with 5,052 nodes and 27,811 interactions. The protein-protein interaction (PPI) network downloaded from BioGRID includes 2,511 nodes and 3,440 interactions. Combing these two networks results a global graph with 10,755 nodes and 35,031 interactions. The results of all three case studies (shown in Table 3.22) demonstrate that the proposed method is still able to find the relevant networks, which include known mechanisms involved in the given disease in two of the three cases. We also compared the constructed networks by neighbor-net analysis using BioGRID and HPRD and the results are shown in Table 3.23.

3.4.5 Results of neighbor-net after removing highly connected genes

Another reasonable question is whether the method might be biased towards heavily studied genes, like known cancer drivers, for which many more protein interactions are known because of study and annotation bias. In order to investigate this, we also applied the proposed approach on the same datasets, but after we excluded the genes with high connectivity in the list of differentially expressed genes for each case study. The histogram

Ranks and p-values of target pathways using interactions from BioGRID		
disease	rank of target pathway	FDR-corrected p-value of target pathway
colorectal cancer	5	2.7e-16
renal cancer	4	5.1e-06
prostate cancer	-	-

Number of disease-associated genes using interactions from BioGRID				
disease	#selected genes	#disease genes	%	p-value
colorectal cancer	46	22	47%	3.51e-07
renal cancer	218	79	36%	4.9e-27
prostate cancer	-	-	-	-

Table 3.22: Results of neighbor-net analysis using BioGRID database as protein-protein interactions resource. The ranks and the p-values of target pathway in three given diseases as well as the enrichment p-value of identified genes in obtained disease-associated genes are shown above. The neighbor-net analysis does not report any significant neighbor networks for the prostate cancer experiments in this analysis. The significant p-values for the target pathways and high enrichment of constructed network in known disease-associated genes in two out of three case studies determine that the proposed method is not significantly dependent on one specific database and is able to identify the known mechanisms involved in the given disease by using different resources.

	using HPRD	using BioGRID	intersection
disease	number of (nodes/edges)	number of (nodes/edges)	number of (nodes/edges)
colorectal cancer	144/215	46/121	39/32
renal cancer	663/1552	218/734	182/184

Table 3.23: Comparing the results of neighbor-net analysis using HPRD and BioGRID databases. The number of nodes and edges in each constructed network, and in the intersection between them in colorectal cancer and renal cancer are shown.

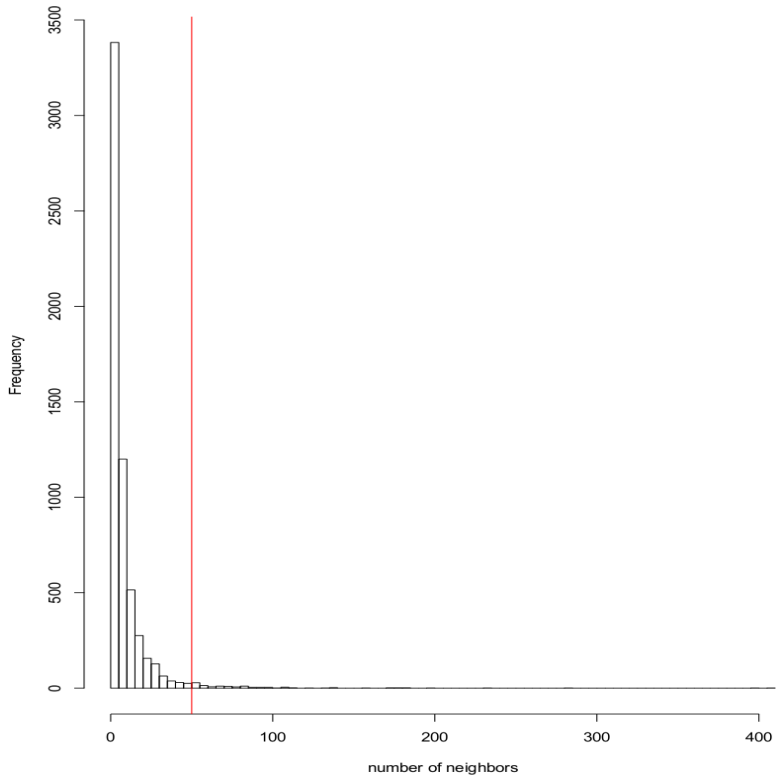


Figure 3.9: The distribution of the number of neighbors for each gene in the obtained network from KEGG and HPRD databases. The number of neighbors is also the size of the neighbor network minus one (the gene itself). The red line represents two percent of the genes, which have more than 50 neighbors. We study the results of the proposed method by excluding the genes connected to more than 50 genes from the analysis.

representing the degree of each gene is shown in Figure 3.9. Genes that are connected to more than 50 genes (Top 2%) are excluded from the list of differentially expressed genes in the analysis. The results of both evaluation approaches are included in Table 3.24. The target pathways in colorectal cancer and renal cancer case studies are still identified significant and highly ranked after this filtering. There was only one significant neighbor network in the prostate cancer study hence, all the pathways except "Proteoglycans in cancer", and 'TGF-beta signaling pathway" did not have any common interactions with the constructed network. The results show that the proposed method is not simply reporting the highly studied genes that are differentially expressed, but rather it is also able to find the mechanisms involving the less studied genes. These putative mechanisms are still significantly enriched in truly relevant pathways, as well as in disease genes from DisGeNET.

Ranks and p-values of target pathways after removing highly connected genes		
disease	rank of target pathway	FDR-corrected p-value of target pathway
colorectal cancer	3	3.9e-03
renal cancer	7	1.0e-03
prostate cancer	3	1.0

Number of disease-associated genes after removing highly connected genes				
disease	#selected genes	#disease genes	%	p-value
colorectal cancer	185	74	40%	1.0e-09
renal cancer	684	39	5%	7.4e-14
prostate cancer	20	4	20%	0.0002

Table 3.24: Results of neighbor-net analysis excluding the genes that are connected to more than 50 genes. The ranks and the p-values of target pathways in three given diseases as well as the enrichment p-values of identified genes in the obtained disease-associated genes are shown above. The significant p-values for the target pathways and high enrichment of constructed network in known disease-associated genes demonstrate that the proposed method is not dependent on the highly connected genes and is able to identify the known mechanisms involved in the given disease after excluding such genes from the analysis.

3.4.6 False positives under the null hypothesis

It is also important to investigate whether the proposed approach has the tendency to produce false positives, i.e. construct networks claimed to be representing putative mechanisms that are in fact not related to the data analyzed. In order to show that this is not the case, we applied the proposed approach on a number of randomly generated datasets. From the graph of 11,086 genes and 62,934 interactions constructed from KEGG and HPRD, we selected 1,000 genes to be "differentially expressed". The proposed approach was applied to this set of 1,000 DE genes to construct neighbor networks and calculate their significance at the 10% level. This whole process was repeated 1,000 times. In 974 cases of these 1,000 simulations (97.4%), no neighbor network was reported as significant at this significance level. In 24 cases (2.4%) there was only one neighbor network reported as significant. In one case (0.1%), there were 2 significant networks and in another one case (0.1%) there were 3 networks reported as significant. These illustrate that the proposed approach produces

substantially fewer false positives than usually accepted (10% false positives are normal for a significance level of 10%).

3.5 Conclusion

Inferring the active network involved in the investigated disease is one of the most important goals in system biology. Given the huge number of publicly available datasets, disease-specific networks can be constructed by using multiple datasets from the same condition to capture multiple states of the genes in the given phenotype. We take advantage of known information about genes interactions available in multiple databases to consider the possible disease-specific effects of genes on the genes immediately connected to them by any kind of known interaction. The networks constructed include interactions from KEGG and HPRD (and therefore it is reasonable to believe they are true). We also have shown that these networks are very significantly enriched in genes known to be involved in the respective diseases according to DisGeNET. Hence, we think it is reasonable to consider them as networks describing the putative mechanisms involved in the given phenotypes. Furthermore, the results obtained from 12 datasets involving three diseases constructed networks were shown to be able to include important genes even though they may not be differentially expressed and therefore, they would not be found by a classical approach based on DE genes.

CHAPTER 4: FUTURE WORK

In this thesis, we used the computed primary dis-regulation of each gene to rank and identify the pathways that are significantly impacted in a given condition. We hypothesize that the computed primary dis-regulation at gene level captures more information about the genes than just their fold-change. One application of these novel computed features is in sample classification or disease sub-typing.

Further more, we are interested in modifying the neighbor net analysis to quantify the effect of mutations on gene expression. Changes in the gene expression have direct effect on protein function and in turn on the progression of a disease. Quantifying these changes are the key in understanding the disease mechanism. However, more often than not, the causes of such changes are not easily identifiable. In many cases, genetic variants may cause some of the observed gene expression changes.

We are proposing to introduce a method that focuses on identifying the variants that significantly alter gene expression for an individual by integrating genetic variant data, gene expression data, as well as a priori knowledge about gene-gene interaction networks using neighborhood of the genes. As stated in chapter 3, the changes in a gene may not be captured in the expression level due to the fact that gene expression data is collected at one single time point. We are hypothesizing that the effect of variants on gene expression is better understood if the neighborhood of each gene is considered. The integration of different types of data helps to capture the changes in the system that are not likely to be captured completely in any one type of data [70, 79]. This is particularly true for complex diseases, such as cancer, which involve many phenomena that affect many levels [108, 148]. More information can be obtained if different types of data were analyzed together, thus the integration of multiple types of data has become a very important problem to solve [130, 131, 132, 133, 134, 135, 136].

Both proposed future directions, identifying variants with significant effect on gene expression using neighbor network and the primary dis-regulation of each gene, will be

helpful in identifying unknown disease subtypes. Many disease have more than one subtype due to the fact that the characteristics and progression of different diseases are the results of the interaction between the disease and the immune system of the host [25, 117, 154]. Because of this, different patients may respond differently to the same drug. Therefore, identifying different subtypes of a given phenotype is extremely important in selecting the most appropriate drug [60, 72].

REFERENCES

[1] D. Atkins, S. Ferrone, G. E. Schmahl, S. Störkel, and B. Seliger, "Down-regulation of HLA class I antigen processing molecules: An immune escape mechanism of renal cell carcinoma?" *The Journal of Urology*, vol. 171, no. 2, pp. 885–889, 2004.

[2] L. Badea, V. Herlea, S. O. Dima, T. Dumitrascu, I. Popescu *et al.*, "Combined gene expression analysis of whole-tissue and microdissected pancreatic ductal adenocarcinoma identifies genes specifically overexpressed in tumor epithelia," *Hepato-gastroenterology*, vol. 55, no. 88, p. 2016, 2008.

[3] G. D. Bader, I. Donaldson, C. Wolting, F. B. Ouellette, T. Pawson, and C. W. Hogue, "BIND–The Biomolecular Interaction Network Database," *Nucleic Acids Research*, vol. 29, no. 1, pp. 242–245, 2001.

[4] C. L. Bailey, P. Kelly, and P. J. Casey, "Activation of Rap1 promotes prostate cancer metastasis," *Cancer Research*, vol. 69, no. 12, pp. 4962–4968, 2009.

[5] R. M. Bailey, J. P. Covy, H. L. Melrose, L. Rousseau, R. Watkinson, J. Knight, S. Miles, M. J. Farrer, D. W. Dickson, B. I. Giasson *et al.*, "LRRK2 phosphorylates novel tau epitopes and promotes tauopathy," *Acta Neuropathologica*, vol. 126, no. 6, pp. 809–827, 2013.

[6] R. E. Bakin, D. Gioeli, R. A. Sikes, E. A. Bissonette, and M. J. Weber, "Constitutive activation of the Ras/mitogen-activated protein kinase signaling pathway promotes androgen hypersensitivity in LNCaP prostate cancer cells," *Cancer Research*, vol. 63, no. 8, pp. 1981–1989, 2003.

[7] M. Bansal, V. Belcastro, A. Ambesi-Impiombato, and D. Di Bernardo, "How to infer gene networks from expression profiles," *Molecular Systems Biology*, vol. 3, no. 1, 2007.

[8] H.-F. Bao, Z.-R. Zhang, Y.-Y. Liang, J. J. Ma, D. C. Eaton, and H.-P. Ma, "Ceramide mediates inhibition of the renal epithelial sodium channel by tumor necrosis factor-α through protein kinase C," *American Journal of Physiology-Renal Physiology*, vol. 293, no. 4, pp. F1178–F1186, 2007.

[9] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva, "NCBI GEO: archive for functional genomics data sets–update," *Nucleic Acids Research*, vol. 41, no. D1, pp. D991–D995, 2013.

[10] E. R. Barry and F. D. Camargo, "The hippo superhighway: Signaling crossroads converging on the hippo/yap pathway in stem cells and development," *Current Opinion in Cell Biology*, vol. 25, no. 2, pp. 247–253, 2013.

[11] A. S. Barth, R. Kuner, A. Buness, M. Ruschhaupt, S. Merk, L. Zwermann, S. Kääb, E. Kreuzer, G. Steinbeck, U. Mansmann *et al.*, "Identification of a common gene expression signature in dilated cardiomyopathy across independent microarray studies," *Journal of the American College of Cardiology*, vol. 48, no. 8, pp. 1610–1617, 2006.

[12] N. N. Batada, T. Reguly, A. Breitkreutz, L. Boucher, B.-J. Breitkreutz, L. D. Hurst, and M. Tyers, "Stratus not altocumulus: a new view of the yeast protein interaction network," *PLOS Biology*, vol. 4, no. 10, p. e317, 2006.

[13] A. Bauer-Mehren, M. Bundschus, M. Rautschka, M. A. Mayer, F. Sanz, and L. I. Furlong, "Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases," *PLOS ONE*, vol. 6, no. 6, p. e20284, 2011.

[14] K. Bhaskar, M. Miller, A. Chludzinski, K. Herrup, M. Zagorski, and B. T. Lamb, "The PI3K-Akt-mTOR pathway regulates Abeta oligomer induced neuronal cell cycle events," *Molecular Neurodegener*, vol. 4, p. 14, 2009.

[15] M. Bienz and H. Clevers, "Linking colorectal cancer to Wnt signaling," *Cell*, vol. 103, no. 2, pp. 311–320, 2000.

[16] BioCarta, "BioCarta - Charting Pathways of Life," http://www.biocarta.com.

[17] E. M. Blalock, J. W. Geddes, K. C. Chen, N. M. Porter, W. R. Markesbery, and P. W. Landfield, "Incipient Alzheimer's disease: microarray correlation analyses reveal major

transcriptional and tumor suppressor responses," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 7, pp. 2173–2178, 2004.

[18] B. Bokanizad, R. Tagett, S. Ansari, B. H. Helmi, and S. Drăghici, "SPATIAL: A System-level PAThway Impact AnaLysis approach," *Nucleic Acids Research*, vol. 44, no. 11, pp. 5034–5044, 2016.

[19] A. J. Butte and I. S. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," in *Pacific Symposium on Biocomputing*, vol. 5, 2000, pp. 418–429.

[20] J. Cai, N. Zhang, Y. Zheng, R. F. de Wilde, A. Maitra, and D. Pan, "The hippo signaling pathway restricts the oncogenic potential of an intestinal regeneration program," *Genes & Development*, vol. 24, no. 21, pp. 2383–2388, 2010.

[21] D. R. Chesire, C. M. Ewing, J. Sauvageot, G. S. Bova, and W. B. Isaacs, "Detection and analysis of β-catenin mutations in prostate cancer," *The Prostate*, vol. 45, no. 4, pp. 323–334, 2000.

[22] K.-H. Cheung, D. Shineman, M. Müller, C. Cardenas, L. Mei, J. Yang, T. Tomita, T. Iwatsubo, V. M.-Y. Lee, and J. K. Foskett, "Mechanism of Ca 2+ disruption in Alzheimer's disease by presenilin regulation of InsP 3 receptor channel gating," *Neuron*, vol. 58, no. 6, pp. 871–883, 2008.

[23] H. Clevers, "Wnt/β-catenin signaling in development and disease," *Cell*, vol. 127, no. 3, pp. 469–480, 2006.

[24] S. C. Correia, R. X. Santos, G. Perry, X. Zhu, P. I. Moreira, and M. A. Smith, "Mitochondrial importance in Alzheimer's, Huntington's and Parkinson's diseases," in *Neurodegenerative Diseases*. Springer, 2012, pp. 205–221.

[25] L. M. Coussens and Z. Werb, "Inflammation and cancer," *Nature*, vol. 420, no. 6917, pp. 860–867, 2002.

[26] N. Craft, Y. Shostak, M. Carey, and C. L. Sawyers, "A mechanism for hormone-independent prostate cancer through modulation of androgen receptor signaling by the HER-2/neu tyrosine kinase," *Nature Medicine*, vol. 5, no. 3, pp. 280–285, 1999.

[27] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, B. Jassal, S. Jupe, L. Matthews, B. May, S. Palatnik, K. Rothfels, V. Shamovsky, H. Song, M. Williams, E. Birney, H. Hermjakob, L. Stein, and P. D'Eustachio, "The Reactome pathway knowledgebase," *Nucleic Acids Research*, vol. 42, no. D1, pp. D472–D477, 2014.

[28] I. De Jong, J. Pruim, P. Elsinga, W. Vaalburg, and H. Mensink, "11C-choline positron emission tomography for the evaluation after treatment of localized prostate cancer," *European Urology*, vol. 44, no. 1, pp. 32–39, 2003.

[29] A. de la Fuente, "From "differential expression" to "differential networking"– identification of dysfunctional regulatory networks in diseases," *Trends in Genetics*, vol. 26, no. 7, pp. 326–333, 2010.

[30] G. Di Lorenzo, G. Tortora, F. P. DArmiento, G. De Rosa, S. Staibano, R. Autorino, M. DArmiento, M. De Laurentiis, S. De Placido, G. Catalano *et al.*, "Expression of epidermal growth factor receptor correlates with disease relapse and progression to androgen-independence in human prostate cancer," *Clinical Cancer Research*, vol. 8, no. 11, pp. 3438–3444, 2002.

[31] M. Donato, Z. Xu, A. Tomoiaga, J. G. Granneman, R. G. MacKenzie, R. Bao, N. G. Than, P. H. Westfall, R. Romero, and S. Drăghici, "Analysis and correction of crosstalk effects in pathway analysis," *Genome Research*, vol. 23, no. 11, pp. 1885–1893, 2013.

[32] S. Drăghici, P. Khatri, A. L. Tarca, K. Amin, A. Done, C. Voichiţa, C. Georgescu, and R. Romero, "A systems biology approach for pathway level analysis," *Genome Research*, vol. 17, no. 10, pp. 1537–1545, 2007.

[33] A. Dürrbach, E. Angevin, P. Poncet, M. Rouleau, G. Chavanel, A. Chapel, D. Thierry, A. Gorter, R. Hirsch, B. Charpentier *et al.*, "Antibody-mediated endocytosis of G 250

tumor-associated antigen allows targeted gene transfer to human renal cell carcinoma in vitro," *Cancer Gene Therapy*, vol. 6, no. 6, pp. 564–571, 1999.

[34] I. J. Edwards, "Proteoglycans in prostate cancer," *Nature Reviews Urology*, vol. 9, no. 4, pp. 196–206, 2012.

[35] B. Efron and R. Tibshirani, "On testing the significance of sets of genes," *The Annals of Applied Statistics*, vol. 1, no. 1, pp. 107–129, 2007.

[36] S. S. El Sheikh, J. Domin, P. Abel, G. Stamp, and E.-N. Lalani, "Phosphorylation of both EGFR and ErbB2 is a reliable predictor of prostate cancer cell proliferation in response to EGF," *Neoplasia*, vol. 6, no. 6, pp. 846–853, 2004.

[37] M. Farsad, R. Schiavina, P. Castellucci, C. Nanni, B. Corti, G. Martorana, R. Canini, W. Grigioni, S. Boschi, M. Marengo *et al.*, "Detection and localization of prostate cancer: correlation of 11C-choline PET/CT with histopathologic step-section analysis," *Journal of Nuclear Medicine*, vol. 46, no. 10, pp. 1642–1649, 2005.

[38] P. T. Francis, "Glutamatergic systems in Alzheimer's disease," *International Journal of Geriatric Psychiatry*, vol. 18, no. S1, pp. S15–S21, 2003.

[39] Y. Gan, C. Shi, L. Inge, M. Hibner, J. Balducci, and Y. Huang, "Differential roles of ERK and Akt pathways in regulation of EGFR-mediated signaling and motility in prostate cancer cells," *Oncogene*, vol. 29, no. 35, pp. 4947–4958, 2010.

[40] E. Glynne-Jones, M. E. Harper, L. T. Seery, R. James, I. Anglin, H. E. Morgan, K. M. Taylor, J. M. Gee, and R. I. Nicholson, "TENB2, a proteoglycan identified in prostate cancer that is associated with disease progression and androgen independence," *International Journal of Cancer*, vol. 94, no. 2, pp. 178–184, 2001.

[41] J. D. Gordan, P. Lal, V. R. Dondeti, R. Letrero, K. N. Parekh, C. E. Oquendo, R. A. Greenberg, K. T. Flaherty, W. K. Rathmell, B. Keith *et al.*, "HIF-α effects on c-Myc distinguish two subtypes of sporadic vhl-deficient clear cell renal carcinoma," *Cancer Cell*, vol. 14, no. 6, pp. 435–446, 2008.

[42] A. W. Grasso, D. Wen, C. M. Miller, J. S. Rhim, T. G. Pretlow, and H.-J. Kung, "ErbB kinases and NDF signaling in human prostate cancer cells," *Oncogene*, vol. 15, no. 22, pp. 2705–2716, 1997.

[43] H. Ha and H. B. Lee, "Reactive oxygen species amplify glucose signaling in renal cells cultured under high glucose and in diabetic kidney," *Nephrology*, vol. 10, no. s2, pp. S7–S10, 2005.

[44] L. B. Haim, K. Ceyzériat, M. A. Carrillo-de Sauvage, F. Aubry, G. Auregan, M. Guillermier, M. Ruiz, F. Petit, D. Houitte, E. Faivre *et al.*, "The JAK/STAT3 pathway is a common Inducer of astrocyte reactivity in Alzheimer's and Huntington's diseases," *The Journal of Neuroscience*, vol. 35, no. 6, pp. 2817–2829, 2015.

[45] H. He, K. Jazdzewski, W. Li, S. Liyanarachchi, R. Nagy, S. Volinia, G. A. Calin, C.-g. Liu, K. Franssila, S. Suster *et al.*, "The role of microRNA genes in papillary thyroid carcinoma," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 52, pp. 19 075–19 080, 2005.

[46] M. J. Herrgård, B.-S. Lee, V. Portnoy, and B. Ø. Palsson, "Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in Saccharomyces cerevisiae," *Genome Research*, vol. 16, no. 5, pp. 627–635, 2006.

[47] A.-C. Hoffmann, K. D. Danenberg, H. Taubert, P. V. Danenberg, and P. Wuerl, "A three-gene signature for outcome in soft tissue sarcoma," *Clinical Cancer Research*, vol. 15, no. 16, pp. 5191–5198, 2009.

[48] T. J. Hohman, S. P. Bell, and A. L. Jefferson, "The role of vascular endothelial growth factor in neurodegeneration and cognitive decline: Exploring interactions with biomarkers of alzheimer disease." *JAMA Neurology*, 2015.

[49] Y. Hong, T. Downey, K. W. Eu, P. K. Koh, and P. Y. Cheah, "A 'metastasis-prone'signature for early-stage mismatch-repair proficient sporadic colorectal cancer patients and its implications for possible therapeutics," *Clinical & Experimental Metastasis*, vol. 27, no. 2, pp. 83–90, 2010.

[50] Y. Hong, K. S. Ho, K. W. Eu, and P. Y. Cheah, "A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis," *Clinical Cancer Research*, vol. 13, no. 4, pp. 1107–1114, 2007.

[51] J. Hou, J. Aerts, B. Den Hamer, W. Van Ijcken, M. Den Bakker, P. Riegman, C. van der Leest, P. van der Spek, J. A. Foekens, H. C. Hoogsteden, F. Grosveld, and S. Philipsen, "Gene expression-based classification of non-small cell lung carcinomas and survival prediction," *PLOS One*, vol. 5, no. 4, p. e10312, 2010.

[52] T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel, "Discovering regulatory and signalling circuits in molecular interaction networks," *Bioinformatics*, vol. 18, no. suppl 1, pp. S233–S240, 2002.

[53] T. Ideker and R. Sharan, "Protein networks in disease," *Genome Research*, vol. 18, no. 4, pp. 644–652, 2008.

[54] R. A. Irizarry, L. Gautier, B. M. Bolstad, C. Miller, M. Astrand, L. M. Cope, R. Gentleman, J. Gentry, C. Halling, W. Huber, J. MacDonald, B. I. P. Rubinstein, C. Workman, and J. Zhang, *affy: Methods for Affymetrix Oligonucleotide Arrays*, 2005, R package version 1.6.7.

[55] W. Jiang, X. Li, S. Rao, L. Wang, L. Du, C. Li, C. Wu, H. Wang, Y. Wang, and B. Yang, "Constructing disease-specific gene networks using pair-wise relevance metric: Application to colon cancer identifies interleukin 8, desmin and enolase 1 as the central elements," *BMC Systems Biology*, vol. 2, no. 1, p. 72, 2008.

[56] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. Gopinath, G. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein, "REACTOME: a knowledgebase of biological pathways," *Nucleic Acids Research*, vol. 33, no. Database issue, pp. D428–432, 2005.

[57] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[58] M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya, "The KEGG databases at GenomeNet," *Nucleic Acids Research*, vol. 30, no. 1, pp. 42–46, 2002.

[59] M. Kanehisa, S. Goto, S. Kawashima, Y. Okunom, and M. Hattori, "The KEGG resource for deciphering the genome," *Nucleic Acids Research*, vol. 32, no. Database Issue, pp. 277–280, Jan 2004.

[60] G. J. Kelloff and C. C. Sigman, "Cancer biomarkers: selecting the right drug for the right patient," *Nature Reviews Drug Discovery*, vol. 11, no. 3, pp. 201–214, 2012.

[61] P. Kemmeren, K. Sameith, L. A. van de Pasch, J. J. Benschop, T. L. Lenstra, T. Margaritis, E. O'Duibhir, E. Apweiler, S. van Wageningen, C. W. Ko *et al.*, "Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors," *Cell*, vol. 157, no. 3, pp. 740–752, 2014.

[62] P. Khatri and S. Drăghici, "Ontological analysis of gene expression data: current tools, limitations, and open problems," *Bioinformatics*, vol. 21, no. 18, pp. 3587–3595, 2005. [Online]. Available: http://dblp.uni-trier.de/db/journals/bioinformatics/bioinformatics21.html#KhatriD05

[63] P. Khatri, M. Sirota, and A. J. Butte, "Ten years of pathway analysis: current approaches and outstanding challenges," *PLOS Computational Biology*, vol. 8, no. 2, p. e1002375, 2012.

[64] J.-H. Kim, C. Xu, Y.-S. Keum, B. Reddy, A. Conney, and A.-N. T. Kong, "Inhibition of EGFR signaling in human prostate cancer PC-3 cells by combination treatment with β-phenylethyl isothiocyanate and curcumin," *Carcinogenesis*, vol. 27, no. 3, pp. 475–482, 2006.

[65] K. Komurov, S. Dursun, S. Erdin, and P. T. Ram, "Netwalker: A contextual network analysis tool for functional genomics," *BMC Genomics*, vol. 13, no. 1, p. 282, 2012.

[66] K. Komurov, M. A. White, and P. T. Ram, "Use of data-biased random walks on graphs for the retrieval of context-specific networks from genomic data," *PLOS Computational Biology*, vol. 6, no. 8, p. e1000889, 2010.

[67] A. Kumar, "Long-term potentiation at CA3–CA1 hippocampal synapses with special emphasis on aging, disease, and stress," *Frontiers in Aging Neuroscience*, vol. 3, 2011.

[68] E. Kyle, L. Neckers, C. Takimoto, G. Curt, and R. Bergan, "Genistein-induced apoptosis of prostate cancer cells is preceded by a specific decrease in focal adhesion kinase activity," *Molecular Pharmacology*, vol. 51, no. 2, pp. 193–200, 1997.

[69] S. A. Lamprecht and M. Lipkin, "Chemoprevention of colon cancer by calcium, vitamin d and folate: molecular mechanisms," *Nature Reviews Cancer*, vol. 3, no. 8, pp. 601–614, 2003.

[70] G. R. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan, and W. S. Noble, "A statistical framework for genomic data fusion," *Bioinformatics*, vol. 20, no. 16, pp. 2626–2635, 2004.

[71] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *BMC Bioinformatics*, vol. 9, no. 1, p. 559, 2008.

[72] B. D. Lehmann, J. A. Bauer, X. Chen, M. E. Sanders, A. B. Chakravarthy, Y. Shyr, and J. A. Pietenpol, "Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies," *The Journal of Clinical Investigation*, vol. 121, no. 7, pp. 2750–2767, 2011.

[73] M. E. Lenburg, L. S. Liou, N. P. Gerry, G. M. Frampton, H. T. Cohen, and M. F. Christman, "Previously unidentified changes in renal cell carcinoma gene expression identified by parametric analysis of microarray data," *BMC Cancer*, vol. 3, no. 1, p. 31, 2003.

[74] W. S. Liang, T. Dunckley, T. G. Beach, A. Grover, D. Mastroeni, D. G. Walker, R. J. Caselli, W. A. Kukull, D. McKeel, J. C. Morris *et al.*, "Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain," *Physiological Genomics*, vol. 28, no. 3, pp. 311–322, 2007.

[75] B. Lifschitz-Mercer, R. Amitai, B. B.-S. Maymon, L. Shechtman, B. Czernobilsky, L. Leider-Trejo, A. Ben-Ze'ev, and B. Geiger, "Nuclear localization of β-catenin and

plakoglobin in primary and metastatic human colonic carcinomas, colonic adenomas, and normal colon," *International Journal of Surgical Pathology*, vol. 9, no. 4, pp. 273–279, 2001.

[76] A. Limon, J. M. Reyes-Ruiz, and R. Miledi, "Loss of functional gabaa receptors in the alzheimer diseased brain," *Proceedings of the National Academy of Sciences*, vol. 109, no. 25, pp. 10 071–10 076, 2012.

[77] W. M. Linehan, J. Vasselli, R. Srinivasan, M. M. Walther, M. Merino, P. Choyke, C. Vocke, L. Schmidt, J. S. Isaacs, G. Glenn *et al.*, "Genetic basis of cancer of the kidney disease-specific approaches to therapy," *Clinical Cancer Research*, vol. 10, no. 18, pp. 6282S–6289S, 2004.

[78] M. Liu, A. Liberzon, S. W. Kong, W. R. Lai, P. J. Park, I. S. Kohane, and S. Kasif, "Network-based analysis of affected biological processes in type 2 diabetes models," *PLOS Genetics*, vol. 3, no. 6, p. e96, 2007.

[79] B. Louie, P. Mork, F. Martin-Sanchez, A. Halevy, and P. Tarczy-Hornoch, "Data integration and genomic medicine," *Journal of Biomedical Informatics*, vol. 40, no. 1, pp. 5–16, 2007.

[80] N. M. Luscombe, M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann, and M. Gerstein, "Genomic analysis of regulatory network dynamics reveals large topological changes," *Nature*, vol. 431, no. 7006, pp. 308–312, 2004.

[81] K. D. MacIsaac, T. Wang, D. B. Gordon, D. K. Gifford, G. D. Stormo, and E. Fraenkel, "An improved map of conserved regulatory sites for Saccharomyces cerevisiae," *BMC Bioinformatics*, vol. 7, no. 1, p. 113, 2006.

[82] N. J. Maragakis and J. D. Rothstein, "Mechanisms of disease: astrocytes in neurodegenerative disease," *Nature Clinical Practice Neurology*, vol. 2, no. 12, pp. 679–689, 2006.

[83] A. A. Margolin and A. Califano, "Theory and limitations of genetic network inference from microarray data," *Annals of the New York Academy of Sciences*, vol. 1115, no. 1, pp. 51–72, 2007.

[84] A. Martorana and G. Koch, "Is dopamine involved in Alzheimer's disease?" *Frontiers in Aging Neuroscience*, vol. 6, 2014.

[85] M. Mhatre, R. A. Floyd, and K. Hensley, "Oxidative stress and neuroinflammation in Alzheimer's disease and amyotrophic lateral sclerosis: common links and potential therapeutic targets," *Journal of Alzheimer's Disease*, vol. 6, no. 2, pp. 147–157, 2004.

[86] K. Mitra, A.-R. Carvunis, S. K. Ramesh, and T. Ideker, "Integrative approaches for finding modular structure in biological networks," *Nature Reviews Genetics*, vol. 14, no. 10, pp. 719–732, 2013.

[87] C. Mitrea, Z. Taghavi, B. Bokanizad, S. Hanoudi, R. Tagett, M. Donato, C. Voichiţa, and S. Drăghici, "Methods and approaches in the topology-based analysis of biological pathways," *Frontiers in Physiology*, vol. 4, p. 278, 2013.

[88] P. J. Morin, A. B. Sparks, V. Korinek, N. Barker, H. Clevers, B. Vogelstein, and K. W. Kinzler, "Activation of β-catenin-Tcf signaling in colon cancer by mutations in β-catenin or apc," *Science*, vol. 275, no. 5307, pp. 1787–1790, 1997.

[89] J. Mulder, M. Zilberter, S. J. Pasquaré, A. Alpár, G. Schulte, S. G. Ferreira, A. Köfalvi, A. M. Martín-Moreno, E. Keimpema, H. Tanila *et al.*, "Molecular reorganization of endocannabinoid signalling in Alzheimer's disease," *Brain*, vol. 134, no. 4, pp. 1041–1060, 2011.

[90] D. Nishimura, "Biocarta," *Biotech Software & Internet Report: The Computer Software Journal for Scient*, vol. 2, no. 3, pp. 117–120, 2001.

[91] N. Novershtern, A. Regev, and N. Friedman, "Physical Module Networks: an integrative approach for reconstructing transcription regulation," *Bioinformatics*, vol. 27, no. 13, pp. i177–i185, 2011.

[92] A. Nykjaer, D. Dragun, D. Walther, H. Vorum, C. Jacobsen, J. Herz, F. Melsen, E. I. Christensen, and T. E. Willnow, "An endocytic pathway essential for renal uptake and activation of the steroid 25-(OH) vitamin D 3," *Cell*, vol. 96, no. 4, pp. 507–515, 1999.

[93] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 27, no. 1, pp. 29–34, 1999.

[94] D. Pan, "The hippo signaling pathway in development and cancer," *Developmental Cell*, vol. 19, no. 4, pp. 491–505, 2010.

[95] K.-H. Pan, C.-J. Lih, and S. N. Cohen, "Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 25, pp. 8961–8965, 2005.

[96] B.-J. Park, J.-I. Park, D.-S. Byun, J.-H. Park, and S.-G. Chi, "Mitogenic conversion of transforming growth factor-β1 effect by oncogenic Ha-Ras-induced activation of the mitogen-activated protein kinase signaling pathway in human prostate cancer," *Cancer Research*, vol. 60, no. 11, pp. 3031–3038, 2000.

[97] H. Pei, L. Li, B. L. Fridley, G. D. Jenkins, K. R. Kalari, W. Lingle, G. Petersen, Z. Lou, and L. Wang, "FKBP51 affects cancer cell response to chemotherapy by negatively regulating Akt," *Cancer Cell*, vol. 16, no. 3, pp. 259–266, 2009.

[98] L. D. Pellegrino, M. E. Peters, C. G. Lyketsos, and C. M. Marano, "Depression in cognitive impairment," *Current Psychiatry Reports*, vol. 15, no. 9, pp. 1–8, 2013.

[99] S. Peri, J. D. Navarro, R. Amanchy, T. Z. Kristiansen, C. K. Jonnalagadda, V. Surendranath, V. Niranjan, B. Muthusamy, T. Gandhi, M. Gronborg *et al.*, "Development of human protein reference database as an initial platform for approaching systems biology in humans," *Genome Research*, vol. 13, no. 10, pp. 2363–2371, 2003.

[100] S. Peri, J. D. Navarro, T. Z. Kristiansen, R. Amanchy, V. Surendranath, B. Muthusamy, T. Gandhi, K. Chandrika, N. Deshpande, S. Suresh, B. Rashmi,

K. Shanker, N. Padma, V. Niranjan, H. Harsha, N. Talreja, B. Vrushabendra, M. Ramya, A. Yatish, M. Joy, H. Shivashankar, M. Kavitha, M. Menezes, D. R. Choudhury, N. Ghosh, R. Saravana, S. Chandran, S. Mohan, C. K. Jonnalagadda, C. Prasad, C. Kumar-Sinha, K. S. Deshpande, and A. Pandey, "Human protein reference database as a discovery resource for proteomics," *Nucleic Acids Research*, vol. 32, no. Suppl 1, pp. D497–D501, 2004.

[101] J. Piñero, N. Queralt-Rosinach, À. Bravo, J. Deu-Pons, A. Bauer-Mehren, M. Baron, F. Sanz, and L. I. Furlong, "DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes," *Database*, vol. 2015, p. bav028, 2015.

[102] J. R. Pollack, C. M. Perou, A. A. Alizadeh, M. B. Eisen, A. Pergamenschikov, C. F. Williams, S. S. Jeffrey, D. Botstein, and P. O. Brown, "Genome-wide analysis of DNA copy-number changes using cDNA microarrays," *Nature Genetics*, vol. 23, pp. 41–46, 1999.

[103] A. Ramteke, H. Ting, C. Agarwal, S. Mateen, R. Somasagara, A. Hussain, M. Graner, B. Frederick, R. Agarwal, and G. Deep, "Exosomes secreted under hypoxia enhance invasiveness and stemness of prostate cancer cells by targeting adherens junction molecules," *Molecular Carcinogenesis*, vol. 54, no. 7, p. 554, 2015.

[104] R. R. Raval, K. W. Lau, M. G. Tran, H. M. Sowter, S. J. Mandriota, J.-L. Li, C. W. Pugh, P. H. Maxwell, A. L. Harris, and P. J. Ratcliffe, "Contrasting properties of hypoxia-inducible factor 1 (HIF-1) and HIF-2 in von hippel-lindau-associated renal cell carcinoma," *Molecular and Cellular Biology*, vol. 25, no. 13, pp. 5675–5686, 2005.

[105] P. Religa, R. Cao, D. Religa, Y. Xue, N. Bogdanovic, D. Westaway, H. H. Marti, B. Winblad, and Y. Cao, "VEGF significantly restores impaired memory behavior in Alzheimer's mice by improvement of vascular survival," *Scientific Reports*, vol. 3, 2013.

[106] T. Reya and H. Clevers, "Wnt signalling in stem cells and cancer," *Nature*, vol. 434, no. 7035, pp. 843–850, 2005.

[107] D. R. Rhodes, S. Kalyana-Sundaram, V. Mahavisno, R. Varambally, J. Yu, B. B. Briggs, T. R. Barrette, M. J. Anstet, C. Kincead-Beal, P. Kulkarni *et al.*, "Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles," *Neoplasia (New York, NY)*, vol. 9, no. 2, p. 166, 2007.

[108] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim, "Methods of integrating data to uncover genotype-phenotype interactions," *Nature Reviews Genetics*, vol. 16, no. 2, pp. 85–97, 2015.

[109] J. M. Rubio-Perez and J. M. Morillas-Ruiz, "A review: inflammatory process in alzheimer's disease, role of cytokines," *The Scientific World Journal*, vol. 2012, 2012.

[110] H. Runne, A. Kuhn, E. J. Wild, W. Pratyaksha, M. Kristiansen, J. D. Isaacs, E. Régulier, M. Delorenzi, S. J. Tabrizi, and R. Luthi-Carter, "Analysis of potential transcriptomic biomarkers for Huntington's disease in peripheral blood," *Proceedings of the National Academy of Sciences*, vol. 104, no. 36, pp. 14 424–14 429, 2007.

[111] J. Sabates-Bellver, L. G. Van der Flier, M. de Palo, E. Cattaneo, C. Maake, H. Rehrauer, E. Laczko, M. A. Kurowski, J. M. Bujnicki, M. Menigatti *et al.*, "Transcriptome profile of human colorectal adenomas," *Molecular Cancer Research*, vol. 5, no. 12, pp. 1263–1275, 2007.

[112] A. Sanchez-Palencia, M. Gomez-Morales, J. A. Gomez-Capilla, V. Pedraza, L. Boyero, R. Rosell, and M. E. Fárez-Vidal, "Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer," *International Journal of Cancer*, vol. 129, no. 2, pp. 355–364, 2011.

[113] E. W. Sayers, T. Barrett, D. A. Benson, E. Bolton, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, S. Federhen *et al.*, "Database resources of the national center for biotechnology information," *Nucleic acids research*, vol. 39, no. suppl 1, pp. D38–D51, 2011.

[114] B. Seliger, D. Atkins, M. Bock, U. Ritz, S. Ferrone, C. Huber, and S. Störkel, "Characterization of human lymphocyte antigen class I antigen-processing machinery defects

in renal cell carcinoma lesions with special emphasis on transporter-associated with antigen-processing down-regulation," *Clinical Cancer Research*, vol. 9, no. 5, pp. 1721–1727, 2003.

[115] S. R. Setlur, K. D. Mertz, Y. Hoshida, F. Demichelis, M. Lupien, S. Perner, A. Sboner, Y. Pawitan, O. Andrén, L. A. Johnson *et al.*, "Estrogen-dependent signaling in a molecularly distinct subclass of aggressive prostate cancer," *Journal of the National Cancer Institute*, vol. 100, no. 11, pp. 815–825, 2008.

[116] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: A software environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[117] S. V. Sharma and J. Settleman, "Oncogene addiction: setting the stage for molecularly targeted cancer therapy," *Genes & Development*, vol. 21, no. 24, pp. 3214–3231, 2007.

[118] G. K. Smyth, *Limma: linear models for microarray data.* New York: Springer, 2005, pp. 397–420.

[119] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," *Nucleic Acids Research*, vol. 34, no. Suppl 1, pp. D535–D539, 2006.

[120] D. L. Stirewalt, S. Meshinchi, K. J. Kopecky, W. Fan, E. L. Pogosova-Agadjanyan, J. H. Engel, M. R. Cronk, K. S. Dorcy, A. R. McQuary, D. Hockenbery *et al.*, "Identification of genes with abnormal expression changes in acute myeloid leukemia," *Genes, Chromosomes and Cancer*, vol. 47, no. 1, pp. 8–20, 2008.

[121] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceeding of The National Academy of Sciences of the Unites States of America*, vol. 102, no. 43, pp. 15 545–15 550, 2005.

[122] M. Sumitomo, R. Shen, M. Walburg, J. Dai, Y. Geng, D. Navarro, G. Boileau, C. N. Papandreou, F. G. Giancotti, B. Knudsen *et al.*, "Neutral endopeptidase inhibits prostate cancer cell migration by blocking focal adhesion kinase signaling," *The Journal of Clinical Investigation*, vol. 106, no. 11, pp. 1399–1407, 2000.

[123] A. Takakura, E. A. Nelson, N. Haque, B. D. Humphreys, K. Zandi-Nejad, D. A. Frank, and J. Zhou, "Pyrimethamine inhibits adult polycystic kidney disease by modulating STAT signaling pathways," *Human Molecular Genetics*, vol. 20, no. 21, pp. 4143–4154, 2011.

[124] H. Takeuchi, H. Mizoguchi, Y. Doi, S. Jin, M. Noda, J. Liang, H. Li, Y. Zhou, R. Mori, S. Yasuoka *et al.*, "Blockade of gap junction hemichannel suppresses disease progression in mouse models of amyotrophic lateral sclerosis and Alzheimer's disease," *PLOS One*, vol. 6, no. 6, p. e21108, 2011.

[125] A. Tanay, R. Sharan, M. Kupiec, and R. Shamir, "Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 9, pp. 2981–2986, 2004.

[126] A. L. Tarca, S. Drăghici, G. Bhatti, and R. Romero, "Down-weighting overlapping genes improves gene set analysis," *BMC Bioinformatics*, vol. 13, no. 1, p. 136, 2012.

[127] A. L. Tarca, S. Drăghici, P. Khatri, S. S. Hassan, P. Mittal, J.-S. Kim, C. J. Kim, J. P. Kusanovic, and R. Romero, "A novel signaling pathway impact analysis (SPIA)," *Bioinformatics*, vol. 25, no. 1, pp. 75–82, 2009.

[128] A. L. Tarca, P. Khatri, and S. Draghici, *SPIA: Signaling Pathway Impact Analysis (SPIA) using combined evidence of pathway over-representation and unusual signaling perturbations*, 2013, R package version 2.14.0. [Online]. Available: http://bioinformatics.oxfordjournals.org/cgi/reprint/btn577v1

[129] R. D. C. Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2005. [Online]. Available: http://www.r-project.org

[130] The Cancer Genome Atlas Research Network, "Integrated genomic analyses of ovarian carcinoma," *Nature*, vol. 474, no. 7353, pp. 609–615, 2011.

[131] ——, "Comprehensive genomic characterization of squamous cell lung cancers," *Nature*, vol. 489, no. 7417, pp. 519–525, 2012.

[132] ——, "Comprehensive molecular characterization of human colon and rectal cancer," *Nature*, vol. 487, no. 7407, pp. 330–337, 2012.

[133] ——, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, 2012.

[134] ——, "Integrated genomic characterization of endometrial carcinoma," *Nature*, vol. 497, no. 7447, pp. 67–73, 2013.

[135] ——, "Integrated genomic characterization of papillary thyroid carcinoma," *Cell*, vol. 159, no. 3, pp. 676–690, 2014.

[136] ——, "Comprehensive genomic characterization of head and neck squamous cell carcinomas," *Nature*, vol. 517, no. 7536, pp. 576–582, 2015.

[137] K. T. Thurn, H. Arora, T. Paunesku, A. Wu, E. M. Brown, C. Doty, J. Kremer, and G. Woloschak, "Endocytosis of titanium dioxide nanoparticles in prostate cancer PC-3M cells," *Nanomedicine: Nanotechnology, Biology and Medicine*, vol. 7, no. 2, pp. 123–130, 2011.

[138] S. Tu, S.-i. Okamoto, S. A. Lipton, and H. Xu, "Oligomeric Aβ-induced synaptic dysfunction in Alzheimer's disease," *Molecular Neurodegeneration*, vol. 9, no. 1, p. 48, 2014.

[139] K. J. Turner, J. W. Moore, A. Jones, C. F. Taylor, D. Cuthbert-Heavens, C. Han, R. D. Leek, K. C. Gatter, P. H. Maxwell, P. J. Ratcliffe *et al.*, "Expression of hypoxia-

inducible factors in human renal cancer relationship to angiogenesis and to the von hippel-lindau gene mutation," *Cancer Research*, vol. 62, no. 10, pp. 2957–2961, 2002.

[140] F. Vandin, E. Upfal, and B. J. Raphael, "Algorithms for detecting significantly mutated pathways in cancer," *Journal of Computational Biology*, vol. 18, no. 3, pp. 507–522, 2011.

[141] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, and R. Sharan, "Associating genes and protein complexes with disease via network propagation," *PLOS Computational Biology*, vol. 6, no. 1, p. e1000641, 2010.

[142] R. J. Viana, A. F. Nunes, and C. M. Rodrigues, "Endoplasmic reticulum enrollment in Alzheimer's disease," *Molecular Neurobiology*, vol. 46, no. 2, pp. 522–534, 2012.

[143] C. Voichiţa, M. Donato, and S. Drăghici, "A genetic algorithms framework for estimating individual gene contributions in signaling pathways," in *Evolutionary Computation (CEC), 2013 IEEE Congress on*. Cancun, Mexico: IEEE, 20-23 Jun. 2013, pp. 650–657.

[144] C. Voichiţa, M. Donato, and S. Drăghici, "Incorporating gene significance in the impact analysis of signaling pathways," in *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, vol. 1. Boca Raton, FL, USA: IEEE, 12-15 Dec. 2012, pp. 126–131.

[145] C. Voichita, S. Ansari, and S. Draghici, *ROntoTools: R Onto-Tools suite*, 2016, R package version 2.0.0. [Online]. Available: http://www.bioconductor.org

[146] C. Voichita and S. Draghici, *ROntoTools: R Onto-Tools suite*, 2013, R package. [Online]. Available: http://www.bioconductor.org

[147] T. A. Wallace, R. L. Prueitt, M. Yi, T. M. Howe, J. W. Gillespie, H. G. Yfantis, R. M. Stephens, N. E. Caporaso, C. A. Loffredo, and S. Ambs, "Tumor immunobiological differences in prostate cancer between African-American and European-American men," *Cancer Research*, vol. 68, no. 3, pp. 927–936, 2008.

[148] L. Wang, F. Li, J. Sheng, and S. T. Wong, "A computational method for clinically relevant cancer stratification and driver mutation module discovery using personal genomics profiles," *BMC Genomics*, vol. 16, no. 7, p. S6, 2015.

[149] P. I. Wang and E. M. Marcotte, "It's the machine that matters: predicting gene function and phenotype from protein networks," *Journal of Proteomics*, vol. 73, no. 11, pp. 2277–2289, 2010.

[150] Y. Wang, O. Roche, M. S. Yan, G. Finak, A. J. Evans, J. L. Metcalf, B. E. Hast, S. C. Hanna, B. Wondergem, K. A. Furge *et al.*, "Regulation of endocytosis via the oxygen-sensing pathway," *Nature Medicine*, vol. 15, no. 3, pp. 319–324, 2009.

[151] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[152] M. J. Weber and D. Gioeli, "Ras signaling in prostate cancer progression," *Journal of Cellular Biochemistry*, vol. 91, no. 1, pp. 13–25, 2004.

[153] N. K. Woods and J. Padmanabhan, "Neuronal calcium signaling and Alzheimer's disease," in *Calcium Signaling*. Springer, 2012, pp. 1193–1217.

[154] H. Yu, M. Kortylewski, and D. Pardoll, "Crosstalk between cancer and immune cells: role of STAT3 in the tumour microenvironment," *Nature Reviews Immunology*, vol. 7, no. 1, pp. 41–51, 2007.

[155] Y. Zhang, M. James, F. A. Middleton, and R. L. Davis, "Transcriptional analysis of multiple brain regions in Parkinson's disease supports the involvement of specific protein processing, energy metabolism, and signaling pathways, and suggests novel disease mechanisms," *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, vol. 137, no. 1, pp. 5–16, 2005.

[156] B. Zheng, Z. Liao, J. J. Locascio, K. A. Lesniak, S. S. Roderick, M. L. Watt, A. C. Eklund, Y. Zhang-James, P. D. Kim, M. A. Hauser *et al.*, "PGC-1α, a potential therapeutic target for early intervention in Parkinson's disease," *Science Translational Medicine*, vol. 2, no. 52, p. 52ra73, 2010.

[157] M. Zou and S. D. Conzen, "A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data," *Bioinformatics*, vol. 21, no. 1, pp. 71–79, 2005.

[158] G. S. Zubenko, J. S. Stiffler, H. B. Hughes, and A. J. Martinez, "Reductions in brain phosphatidylinositol kinase activities in Alzheimer's disease," *Biological Psychiatry*, vol. 45, no. 6, pp. 731–736, 1999.

ABSTRACT

NETWORK-BASED APPROACHES TO IDENTIFY THE IMPACTED GENES AND ACTIVE INTERACTIONS

by

SAHAR ANSARI

August 2017

Advisor: Dr. Sorin Draghici

Major: Computer Science

Degree: Doctor of Philosophy

A very important step in system biology is the identification of the networks that are most impacted in the given phenotype. Such networks explain where the target genes are affected by some other genes, and therefore describe the mechanisms involved in a biological process. The identified networks are used to: 1) predict the disease or the responses of the system to a specific impact, 2) find the subset of genes that interact with each other and play an important role in the condition of interest, and 3) understand the mechanisms involved in that condition. In this thesis, we propose an approach that takes advantage of pre-defined pathways obtained from existing databases to identify the impact of a phenotype studied on such pathways. Next, we introduce a method able to build a network that captures the putative mechanisms at play in the given condition, by using datasets from multiple experiments studying the same phenotype. This method takes advantage of known interactions extracted from multiple sources such as protein-protein interactions and curated biological pathways. Based on such prior knowledge, we overcome the drawbacks of snap-shot data by considering the possible effects of each gene on its neighbors.

AUTOBIOGRAPHICAL STATEMENT

Sahar Ansari

Education

- Ph.D. Computer Science, Wayne State University, Detroit MI, USA, Expected 2017.
- B.S. Electrical Engineering, Sharif University of Technology, Tehran, Iran, June 2011.

Peer review publications

- **S. Ansari**, C. Voichita, M. Donato, R. Tagett, and S. Draghici, "A novel pathway analysis approach based on the unexplained disregulation of genes," Proceedings of the IEEE, vol. PP, no. 99, pp. 1-14, March 2016. [Online]. Available: http://dx.doi.org/10.1109/JPROC.2016.2531000
- **S. Ansari**, M. Donato, N. Saberian, and S. Draghici, "An approach to infer putative disease-specific mechanisms using neighboring gene networks," Bioinformatics, p. btx097, 2017.
- B. Bokanizad, R. Tagett, **S. Ansari**, B. H. Helmi, and S. Draghici, "SPATIAL: A System-level PAThway Impact AnaLysis approach," Nucleic Acids Research, vol. 44, no. 11, pp. 5034-5044, 2016.
- N. Saberian, A. Peyvandipour, M. Donato, **S. Ansari**, S. Draghici, "A new computational drug repurposing method using established disease-drug pair knowledg," *Under review*

Software packages

- C. Voichita, **S. Ansari**, and S. Draghici, ROntoTools: R Onto-Tools suite, 2016, R package version 2.0.0. [Online]. Available: http://www.bioconductor.org