HÖGSKOLAN SKÖVDE

Master Thesis in BIOINFORMATICS

Ranking the Relevance of Genes Targeted by Cancer-Associated MiRNAs

Author:
Jörg Linde

Supervisor:
Dr. Zelmina Lubovac

School of Humanities and Informatics
University of Skövde
July 1, 2008

Abstract

MicroRNAs control the expression of their target genes by translational repression. They are involved in various biological processes including cancer progression. To uncover the biological role of microRNAs it is necessary to identify their target genes. The small number of experimentally validated target genes makes computer prediction methods very important. However, state of the art prediction tools result in a great number of putative targets. The number of false positives among those putative targets is unknown. This report proposes, investigates and analyses two ways of ranking the biological relevance of putative targets of miRNAs which are associated with breast cancer.

One approach characterises values of network properties of the putative microRNA targets in the human Protein-Protein Interaction network and compares them to network property values of validated microRNA targets. Using these results we suggest a simple approach for ranking the relevance of putative targets. The approach consists of testing if a network property value of a putative target differs from the mean value of the network. In addition we study which network property contributes most to ranking using this approach.

The second approach identifies commonly overrepresented Gene Ontology categories among putative microRNA targets, validated targets and known breast cancer genes. We investigate possibilities to use the occurrence of a putative target in these categories to rank its biological relevance.

Finally we present a number of genes with interesting features considering both approaches. These genes might play a role in breast cancer progression and might be worth investigating further.

Contents

Nomenclature

BP	Subontology Biological Process
cancersupported	Overrepresented GO categories for putative targets which are also overrepresented for cancer genes
CC	Subontology Cellular Component
copPut	Dataset of properties of putative targets within whole PPIN
copVal	Dataset of properties of validated targets within whole PPIN
doublesupported	Overrepresented GO categories for putative targets which are also overrepresented for validated targets and cancer genes
fullNet	Whole human Protein-Protein-Interaction Network
GO	Gene Ontology
MF	Subontology Molecular Function
miRNA	MicroRNA
MWW	Unpaired Mann-Whitney-Wilcoxon test
PPIN	Human Protein-Protein-Interaction Network
putNet	Subnetwork of putative targets and their first neigbours
valNet	Subnetwork of validated targets and their first neigbours
valsupported	Overrepresented GO categories for putative targets which are also overrepresented for validated targets

1 Introduction

1.1 Problem Definition

MicroRNAs (miRNAs) are small RNAs which suppress the translation of their target messengerRNAs (mRNAs), and in this way control the expression of the target genes. Within recent years a lot of research has been done which aims to discover the biological roles of miRNAs. One task is to find the target genes which are suppressed by a distinct miRNA. Because there exists no highthroughput experimental method, which finds miRNA target genes, the number of experimentally validated targets is still quite small [1]. This makes computer prediction methods necessary. Most of the prediction methods are based on sequence similarity or thermodynamical features (see chapter 2.5). However, for a single miRNA these tools predict a large number of putative target genes with an unknown number of false positives. Therefore this report tries to define, investigate and compare different approaches to characterise the putative target genes according to their biological relevance. To do so we use more biological background information of the putative target genes than only sequence based or thermodynamical features. Furthermore it has been shown that miRNAs play a role in cancer development [2, 3, 4].Iorio et al identified miRNAs which are differentially expressed between normal tissue and human breast cancer tissue [5]. The data set used in this work consists of putative target genes of those miRNAs. While some of those genes have a known function in human breast cancer, most of the about 700 genes, are not known to play an important role in this disease. Thus this work tries to find those putative target genes which seem to play a major role in cancer development by using biological and cancer related background information.

This work uses two different approaches to characterise the putative target genes. The first one ranks them according to their network property values within the human Protein- Protein Interaction Network which differ statistically significantly from the mean property values within this network [6, 7]. The second approach uses overrepresented Gene Ontology (GO) [8] categories to rank the putative target genes.

1.2 Motivation

MiRNAs are involved in several cellular processes and in a number of diseases including cancer [9, 2]. To uncover their role the prediction of target genes is an important step. However, state of the art prediction tools result in a large number of putative target genes. The small number of experimentally validated target genes [1] makes it difficult to determine the number of false positives and false negatives [10]. But even if we had a highly specific algorithm we could not be sure which target gene is important in a biological sense. For example a target might only be seen as relevant if the target gene and the miRNA are temporarily and locally similarly expressed. This project investigates different properties of putative miRNA target genes that may be useful to suggest possible ranking of these genes according to their biological relevance. Most existing prediction tools use sequence based or three dimensional complex analysis to find target genes and evolutionary sequence conservation to reduce the number of false positives. As a complement to this we try to use more biological background information to find the most important genes among these putative target genes. In this ways, we aim to find highly biologically relevant target genes. This may help us to understand the biological roles of miRNAs. The use of a data set consisting of cancer relevant miRNAs may furthermore lead to the finding of target genes which play an important role in cancer. This could finally improve both the diagnosis and the treatment of cancer. It has been suggested that miRNAs and their targets might be used as both biomarkers and drug targets [2]. We try to create a list of target genes which then could be further studied and experimentally validated.

1.3 Aims and Objectives

This report uses two different methods to characterise the biological relevance of genes which are putative targets of miRNAs which are differentially expressed between human breast cancer tissue and normal tissue proposed by Iorio et al [5].

At first we use network properties of the human Protein-Protein Interaction Networks (PPINs). It has been shown that the values of some network properties of putative miRNA targets differ significantly from mean values of the whole human PPIN [6, 7]. The use of network properties to characterise cancer related proteins is also supported by the finding that cancer related proteins have a greater mean degree value than the

mean value of the human PPIN [11]. So we characterise these properties of the vertices from the putative target genes and compare them to the mean values of the network. In this way we investigate different ways of ranking the putative target genes according to their role within the PPIN. To evaluate different approaches we use a data set of experimentally validated miRNA target genes [1] and a data set of genes which are known to be involved in breast cancer [12].

The second approach uses the GO [8]. It has been shown that some GO categories are overrepresented for miRNA target genes [13]. We compare the set of overrepresented GO categories for the putative target genes to both: 1) overrepresented categories for experimentally validated target genes [1] and 2) overrepresented categories for genes known to be involved in breast cancer [12]. In this way we try to find new ways to characterise the relevance of putative target genes. Overrepresented GO categories have been used before to characterise the biological role of miRNA targets [13]. However, our approach tries to use commonly overrepresented GO categories for the putative targets, validated targets and cancer genes to rank the relevance of the putative target genes. This has never been done before.

1.4 Overview of this Report

The first chapter of this report introduces the topic of this work. The second chapter explains basic terms and presents related work. The third chapter introduces the used data sets and presents the methods in detail. The fourth chapter shows the results of the different methods while the fifth chapter discusses these results and proposes further work.

2 Background and Related Work

2.1 Biological Terms

This report uses a variety of molecular biological terms such as: genes, DNA, RNA, proteins etc. A definition of these terms can, for example, be found in Albert et al [14]. For convenience, we sometimes use the terms protein and gene interchangeably. Especially if we talk about a "target gene" in the Protein-Protein Interaction Network we refer to the product of this gene. We use Gene Symbols as identifiers for the proteins.

2.2 MicroRNA

MicroRNAs (miRNA) are small non-coding RNAs which play a role in posttranscriptional gene expression control. The first microRNA was found in 1993 [15] and since then miRNAs have been widely studied. MiRNA genes are frequently a part of the introns of other genes but may also be found in other regions of the genome [2] . Little is known about the control of the expression of miRNA genes themselves. Because they are often located in introns their expression is coupled to the genes in whose introns they are located. Furthermore it has been suggested that miRNAs are frequently controlled in an epigenetic way, for example by methylation of CpG islands [13]. This is supported by the finding that miRNAs are often located at fragile chromosomal sites [3]. After transcription, the pri-miRNA is cut into an about 70 base pairs long precursor miRNA. After transporting out from the nucleus this RNA is cut again into the about 22 base pair long miRNA. This miRNA can then silence target genes in two ways, either by cleavage of the mRNA or by repressing of its translation. The extent of the base pairing between the small miRNA and the mRNA seems to determine whether the miRNA cleaves its target after forming an RNA-induced silencing complex or represses translation by binding to the 3'UTR [9]. MiRNAs were found in 76 different species including vertebrates, plants and viruses [1] [16]. Together with the evolutionary conservation of miRNA sequences [17] this finding supports the important role of miRNAs in gene expression control. About one-third of all human genes is expected to be regulated by miRNAs [18].

[1] miRBASE, release 10.1 April 2008

2.3 MicroRNA targets

To uncover the biological role of miRNAs, researchers should first discover which genes are targets of a distinct miRNA and are thus repressed by it. 541 human miRNAs have been experimentally discovered [2] [16] and more than 800 are predicted [19]. One miRNA can regulate several targets and one target can have several target sites and thus be regulated by several miRNAs. This implies that about 30% of human genes are miRNA targets [17] and supports the important role of miRNA regulated gene expression control. However, the number of experimentally validated targets is still small. TarBase [3] [1] includes 570 experimentally validated targets for 128 miRNAs for 8 species. This shows that for a lot of miRNAs not a single validated target is know. This database makes a difference between "direct support" of targets and "indirect support". The difference is the experiment which is used to validate a target. While the first one tests all target sites individually using an in vitro gene assay [20], indirect support is given by simultaneously testing multiple target sites by MicroArray gene expression analysis (since a repression of a microRNA should lead to higher expression of its target genes) [21]. The number of directly supported targets is considerably smaller.

Still, little is known about the general function of miRNAs and their targets. Stark et al [13] reported that genes which are not targets are involved in basic cellular processes and that miRNAs and their target genes are often expressed in neighbouring tissue. This leads to the proposal that microRNA may play a role in developmental processes. Furthermore they found a statistically significant overrepresentation of Gene Ontology categories [8] for microRNA targets.

2.4 The Role of MicroRNAs in Cancer

Cancer is a disease more and more people in the western world suffer from [22]. Although there are different types of cancer, this disease has some common characteristics: the loss of cellular identity, an uncontrolled cell growth and changes in the cell death controlling system [2]. Cancer shows a number of genetic and epigenetic changes. Recent studies have shown that miRNAs play a role in the development of cancer [9, 2, 5]. MicroRNAs can act as both tumor suppressor genes and oncogenes [2]. As tumor suppressor genes

[2] miRBASE, release 10.1 April 2008
[3] April 2008

they repress the translation of genes which cause tumors. On the other hand miRNAs can also repress the translation of tumor suppressor genes and thus act as oncogenes. This demonstrates how important the knowledge of miRNA targets is to uncover the function of miRNAs. Lu et al [23] showed that MicroRNA expression profiles can be used to classify human cancer states. Iorio et al [5] found miRNA genes differentially expressed between human breast cancer tissue and normal tissue. They even showed that it is possible to distinguish between normal tissue and breast cancer tissue by analysing the expression profile of a set of miRNAs. For the five most differentially expressed miRNAs they predicted 719 putative target genes. The aim of this work is to investigate different properties of the miRNA targets that may be useful to rank these putative target genes according to their relevance. Furthermore it has been shown that miRNAs are often located in chromosomal fragile sites which are associated to cancer [3] which supports the important role of miRNAs in cancer. MiRNAs and their target genes are possible drug targets and may be used to diagnose cancer [2, 3]. The finding that the expression values of 15 deregulated miRNAs could be used to discriminate between normal tissue and human breast cancer tissue supports this proposal [5].

2.5 Prediction Methods

The small number of experimentally validated targets makes computer based prediction methods necessary. The first tools were supported by the high complementarity between miRNAs and experimentally supported binding sites. For plant miRNAs and their targets are highly complementary which makes the prediction easier. However, animal microRNAs and their targets show only partial complementarity [24].

There are a number of different methods to search for miRNA targets. Some of them are especially created to be applied to a certain organism [10]. Most of them can be broadly divided into two types; the first searches for complementary target genes while the second uses thermodynamical features of the miRNA:targetmRNA complex to find target genes.

The first approach relies on the finding that most animal miRNAs bind to the 3' untranslated region of the target genes. In experiments researchers identified a "seed region" near the 5' end of the miRNA which is exactly complementary to the target site and three different ways how animal miRNA bind to the target site [25]. This knowledge

is used by the tools which scan the sequences for putative target sites. However, due to the shortness of miRNAs there are a lot of hits and algorithms (like MiRanda [26] and TargetScan [27]) try to reduce the number of false positives by using evolutionary conservation of putative target sites and thermodynamical features of the complexes.

The second approach starts directly with thermodynamical features and calculates an energy term for the miRNA in complex with the putative target gene by scanning the genome with a given window size. But the problem here is to find the right cutoff for the energy term and to identify false positives. This approach is used by the tools RNAhybrid [28] and PicTar [20] for instance. RNAhybrid also uses statistical methods to receive information for the background distribution of the energy term and multiple binding sites.

An alternative way to find miRNA target genes is to use motif search in the 3'UTR of the genes. A motif is a short sequence which occurs more frequently than expected in random sequences. In this way functional elements of the 3'UTR can be found and it has been shown that they are putative miRNA targets [29]. A further idea is to test for correlated expression [30]. A repression of microRNAs should lead to higher expression values of their target genes and vice verca. This can be tested with MicroArray Expression Profiles. Note that similar approaches are also used to validate target genes (see chapter 2.3). But since this method need experiments we can not use it for every microRNA and target since the expression data might not be available yet.

Recently developed methods try to use machine learning techniques to find target genes. They use different combinations of properties from known miRNA:targetmRNA complexes to learn from them. For example Kim et al [31] used a support vector machine. These approaches may improve with an increasing number of experimentally validated target sites and complexes.

The problem with all approaches is that the number of putative target genes for a distinct miRNA is quite large and the number of false positives and false negatives is difficult to calculate [10]. A reason for this is the shortness of miRNAs which leads to a large number of complementary sequences and possible binding partners. Most of the tools use randomly shuffled miRNA sequences to guess false positive rates. However, this is only a computational control using no biological background knowledge. Sethupathy et al [10] compared the most popular tools and showed that they differ in sensitivity

and specificity. A problem in this study was the definition of the false positive rate since the number of genes known to not be regulated regulated by a miRNA is small [1]. For this reason they decided to use the total number of predicted targets instead as quality measurement which we use in our report as well.

2.6 Protein-Protein Interaction Networks

A Protein-Protein Interaction Network (PPIN) is a Graph $G = (V, E)$ where the set of vertices V represents the set of proteins and the set of edges E (which are unsorted pairs of vertices) represents the set of interactions between two proteins. Two proteins "interact" if they physically bind (mostly temporary) . With the help of high throughput experiments, such as yeast two hybrid systems it was possible to construct the PPIN for the whole human proteome. A PPIN is a biological network which has three main features [32, 33]:

1. The degree values are power law distributed
2. The mean shortest path length is shorter than in random networks (they are so called small world networks)
3. There occur more motifs in such networks than in random networks

   The degree is the number of direct neighbours of a vertex. Biological networks usually consist of a few vertices with a great degree value (so called "hubs") while most of them have a quite small one. This makes the network more robust against random changes. A path is an ordered sequence of connected vertices from a start vertex to an end point.

   The length of a path is the number of edges used. The shortest path for two vertices has a minimal number of edges among all possible paths. Biological networks have a small mean shortest path length which contributes to a fast information flux. A network motif consists of a set of connected vertices which occur in this connected way more often than expected. Proteins within one motif are believed to work somehow together in the cell.

   A number of properties can be calculated for a network. It has been shown that the property values of miRNA regulated proteins differ significantly from the mean

property values of the whole PPIN [6, 7] for the four properties: degree, vertex betweenness centrality, closeness centrality and clustering coefficient. Furthermore it has been shown that cancer associated proteins have greater mean degree and clustering coefficient values than the mean values of the PPIN [11]. For those reasons our report will focus on those four properties.

a) The degree value of a vertex is the number of directed neighbours of the vertex [34].

b) The vertex betweenness centrality value measures the centrality of it within the network and in this way it indicates how important the vertex is [34]. It is the sum over all pairwise vertices of the fraction of shortest paths crossing the vertex divided by the number of all shortest path, and is defined as:

$$B(v) = \sum_{s \neq t \neq v \in V} \frac{\delta_{st}(v)}{\delta_{st}}$$

In this formula, δ_{st} is the number of shortest paths from s to t and $\delta_{st}(v)$ is the number of shortest paths from s to t that cross v.

c) Another centrality measure is closeness centrality which is defined as the mean shortest path length between a vertex and all other vertices reachable from it [34]. It is defined as:

$$G(v) = \frac{\sum_{t \in V - v} d(v,t)}{n-1}$$

where $d(v,t)$ is the shortest distance between v and t, and n is the number of reachable vertices from v. vertices which have mean short distances to other ones in the network have greater closeness centrality values.

d) The clustering coefficient value is the fraction of connected neighbours of a vertex i divided by the number of all possible connections of the neighbours of i [35]. In other words it measures how close a vertex and its neighbours are to be totally connected and thus form a "clique". Many cliques contribute to a small world network because you can reach every vertex on a short path [35]. The clustering coefficient of an undirected graph is defined as two times the number of edges between the neighbours divided by the number of all

possible edges between the neighbours:

$$C(v) = \frac{2 * e(N(v))}{n_v * (n_v - 1)}$$

Where $e(N(v))$ is the number of edges between the neighbours of v, and n_v is the number of neighbours of v.

2.7 Gene Ontology

The Gene Ontology [8] is a database which provides a controlled and annotated vocabulary to describe the function of genes and gene products. There are three different subontologies: "Cellular Component", "Biological Process" and "Molecular Function", and for each of these a number of categories are used to describe the gene. The first subontology describes where in the cell a protein is located, while the second one states in which processes it is involved, and the last one describes the function of the protein. In mathematical terms GO is a directed acyclic graph where the vertices represent categories describing the genes whereas edges display either an "is-a" or "part of" relationship between the categories. For example, if a proteins is inside the cell nucleus, as denoted by the "Cellular Component" subontology, it is as well inside the cell which is the ancestor of the category "nucleus".

Calculation of overrepresentation of GO categories has been used to study the function of a set of miRNA target genes [26, 13]. Stark et al [13] also showed that some categories are significantly overrepresented for miRNA targets. We used the tool BiNGO [4] [36] to test for GO category overrepresentation because it is an easy to use plugin for Cytoscape [5] [37] which we also used to construct subnetworks. Another advantage of BiNGO is that it also results in graphs which represent the connections of the overrepresented GO categories within the GO hierarchy. In this way it is easy to see which categories are overrepresented because of inheritance, i.e. because their children vertices (representing categories in the graph) are overrepresented.

[4] version 2.0 downloaded March 2008
[5] version 2.5.2 downloaded March 2008

3 Materials and Methods

3.1 Datasets

3.1.1 Putative Target Genes of Deregulated MicroRNAs in Breast Cancer

Iorio et al [5] identified 29 significantly deregulated miRNAs in human breast cancer. They also showed that 15 of these miRNAs were sufficient to discriminate between normal tissue and breast cancer tissue by only analysing the miRNA expression profiles. The five most consistently differently expressed miRNAs are miR10b,miR125b, miR145,miR21 and miR155. While the first three are down-regulated, the last two are up-regulated. Because of that Iorio et al [5] assumed that the first three act as tumor suppressor genes while the last two are oncogenes. Consequently, the first three should repress targets which are oncogenes and the last two tumor suppressor genes. To generate a list of putative targets they used the three tools miRanda [26], Pictar [20] and TargetScan [27]. For more reliable results they intersected the lists generated by the tools, i.e they only accepted a target if it was predicted by at least two of the tools. Thus they generated a list of 719 putative target genes for these five miRNAs [6]. We had to exclude some of these genes for either of the following reasons: they were from the mouse or rat genome and no human homologue was known, they are not yet included in the human PPIN, or they do not have any GO annotation

3.1.2 Experimentally Validated MicroRNA Target Genes

We used TarBase [1] to access experimentally validated miRNA targets in order to compare them to putative targets. In March 2008 TarBase contained 461 validated human miRNA target sites from 418 different genes. Out of those 418, 213 are member of the human PPIN. For the five most consistently differentially expressed miRNAs identified by Iorio et al [5] a TarBase search resulted in 19 experimentally validated target genes. Out of these 19 five are from the human genome and "directly supported" (see chapter 2.3) while 15 are from the mouse genome and "indirectly supported" (one gene is from both genomes). For miR155 and miR145 one target was found while there exists two targets for miR21. The 15 "indirectly supported" targets are for miR-125b

[6] available at the supplementary material of Iorio et al [5]

while for miR-10b not a single target was found. The list of putative target genes generated by Iorio et al [5] does not consist of any of these validated target genes for these five miRNAs. However, out of the 418 validated target genes 22 can be found in this list. Note that these are targets of other miRNAs than those five most consistently deregulated.

3.1.3 Protein-Protein Interaction Networks

The human PPIN was downloaded from the Human Protein Reference Database [7] [38] which contains manually curated Protein-Protein Interactions. As identifiers we used Gene Symbols. From now on we will refer to this network as "fullPIN". This network includes 9162 vertices. We used Cytoscape [37] to find the putative target genes for deregulated miRNAs in breast cancer within "fullPIN". Out of 719 putative target genes 465 were found within "fullPIN". We want to investigate if it is better to use network properties of the putative targets in the whole human PPIN to rank these genes or if it is better to construct a subnetwork only of the putative targets. The advantage of the second way could be that we can characterise how only the putative targets are connected to each other (f.e. how fast information can be brought from one target to another target using only the targets). However, most of the network properties consider the neighbourhood of a vertex and for this reason we constructed a subnetwork including all 465 putative targets and their neighbours. This subnetwork consists of 2793 proteins. We will refer to this subnetwork as "putNet". We also tried to generate a subnetwork which consists of the neighbours of putNet and thus includes the neighbours of the neighbours of the putative target genes. However, this subnetwork consists of more than 7000 vertices and would have been too similar to "fullPIN".

For the 418 experimentally validated miRNA target genes we constructed a subnetwork including themselves and their direct neighbours. Out of those 418 genes, 213 were found within the human PPIN. The constructed subnetwork consists of 1963 vertices and we will refer to it as "valNet". The properties of this subnetwork are used to compare them to the properties of "putNet". Furthermore we directly copied the values of the properties of the 465 putative targets from "fullPIN" to another data set. In this way we want to investigate the network properties of the putative targets within the whole human PPIN and see if they differ from those of the subnetworks ("putNet"

[7] accessed September 2007

and "valNet") and from the mean properties of "fullPIN". We refer to this data set as "copPut". To compare this to the validated targets we copied the properties of the 213 validated targets from "fullNet" to another data set which we call "copVal".

3.1.4 Breast Cancer Genes

We downloaded genes which are known to be involved in breast cancer from the Breast Cancer Gene Database [8] [12]. The database is manually curated and the data is extracted from literature search. We downloaded all 72 genes which we use to evaluate our results. However, we could not use all of them because they were not included in the human PPIN or in the GO. Out of those 72 genes, 27 are located in the human PPIN but only one is a member of the 465 putative target genes we used for the PPIN analysis.

3.1.5 Gene Ontology

For GO analysis we used the BiNGO [36] plugin for Cytoscape [37]. Version 2.0 of BiNGO uses default human GO annotation files which are parsed directly from the GO web page. However, the plugin does not download and parse the annotation files "on demand", rather they are a static part of the BiNGO package. Our GO category analysis used GO annotation files from August 2007. We tested three different data sets against the three subontologies "Biological Process", "Molecular Function" and "Cellular Location": the list of 465 putative targets used for the PPIN analysis, the 418 experimentally validated targets and the 72 genes known to be involved in breast cancer. Because not all of these genes are annotated with GO categories, the number of included genes varied between the three subontologies. This is summarised in table 1. Each of the nine tests resulted in both a list with overrepresented GO categories and genes in these categories, and a graph representing the connections of overrepresented GO categories. BiNGO [36] performs a hypergeometric test. The result of this test is a p value which is the probability of sampling X genes of the test set (putative-, or validated targets, breast cancer genes) against N genes of the reference set (whole subontologies) which results in x or more genes sharing the GO category G which is shared by n out of the N genes. We used a $p < 0.05$ as significance threshold. Thus, the p value states the

[8] accessed in April 2008

probability that the overrepresentation occurred randomly. However, some categories might be overrepresented because they are parent categories of overrepresented children categories in the GO hierarchy. That is why the authors of BiNGO state that the most relevant categories might be those which are overrepresented and most specific in the hierarchy [36].

Reference Set	Molecular Function	Biological Process	Cellular Component
putative targets	439	425	418
validated targets	331	321	330
breast cancer genes	57	55	53

Table 1: Number of used Genes for GO Analysis

3.2 Ranking Methods with the Help of PPIN Properties

In order to analyse network properties of the five data sets "fullNet", "putNet", "valNet", "copPut" and "copVal", we used the "igraph" library [9] [39] of the statistical computing language R [10] [40]. The three networks ("fullNet", "putNet" and "valNet") were analysed and the values of the following properties were calculated for every vertex: degree, betweenness centrality, clustering coefficient and closeness centrality. With the help of a PERL script, network properties of the 465 putative and the 213 validated targets were copied from the results of the whole human PPIN "fullNet" which generated the data sets "copPut" and "copVal" as described in chapter 3.1.3. To characterise mean property values of the data sets we calculated the mean values for each of these four properties in every of the five data sets.

To investigate if the sets of properties differ significantly we carried out statistical tests. We know that the degree distribution follows a power law distribution [32] but for the other properties we do not know which kind of distribution they follow. We applied the unpaired Mann-Whitney-Wilcoxon (MWW) signed rank test [41] to test statistical significance. This test allows to determine significant differences between non Gaussian distributed data sets. Since we used different networks, our data sets are unpaired. The test assesses whether two samples come from the same distribution (Null Hypothesis) or from different distributions (Alternative Hypothesis). We did a directed

[9] version 0.5
[10] version 2.4.1

test in the direction indicated by the results of the comparison of network properties with the corresponding mean values (table 4). For all five data sets and all four properties we ran MWW in a pairwise manner and in this way tested if the differences or similarities between those five data sets are significant. The result of the test is a p value which states the probability that the Null Hypothesis is true.

In the next step we performed a simple way of ranking. For every vertex and every property it was tested if the value is either "greater than" or "smaller than" the mean value of the data set with the help of a PERL script. For every property and every data set we chose either the "greater than" or the "smaller than" criteria depending on how the mean values of the data sets differ from the mean values of "fullNet" which can be seen in table 4. For example since the mean betweenness centrality value of "putNet" is smaller than the corresponding value for "fullNet" we tested for every vertex in "putNet" if it has a smaller betweenness centrality value than the mean value of "putNet". However, since the mean betweenness centrality value of "copPut" is greater than "fullNet", we used "greater than" in this case. This idea is supported by a former study which proposes that vertices with greater degree values than the mean values are so called "hubs" and are thus important [42]. Also, proteins with greater degree values than the mean degree value are shown to be involved in cancer [11].

Furthermore, we studied, for each of the 16 possible combinations of network properties how much it contributes to the ranking. With the same PERL script we counted for how many proteins the "greater than" or the "smaller than" criteria applies to every property and every combination of properties.

In order to study if it is better to use subnetworks or the properties within the whole human PPIN we used this PERL script for the data sets "putNet" and "copPut". These results are compared to the validated data sets "valNet" and "copVal". For "copPut" we used both sets of mean values, the mean values of the data sets themselves, and the mean values of "fullNet" since they are from the same network. For "putNet" we only used mean property values of the data set itself since these properties are from a different network than "fullNet".

The output of this PERL script is a list which shows a Boolean value indicating for every vertex and every property if it is "greater than" or "smaller than" the mean value. An example of this list is shown in table 2. Furthermore the script prints a number that

shows how many vertices have different values than their mean values for every property and every combination of properties. An example of output is the number of vertices having a greater degree value, smaller betweenness centrality value and greater closeness centrality value. Using these numbers we calculated for each of the 16 combinations of properties the fraction of the number of properties having this combination divided by the number of all vertices in the network. In this way we want to analyse how much every combination of properties contributes to the ranking and if there is a similar pattern between the putative and the validated targets. The same procedure was performed for all four data sets and for "copPut" twice, once with the mean values of the data set itself and once with the mean values of "fullNet".

In order to visualise the results, and enable an easier interpretation, we used the tables to generate Venn diagrams which show the overlaps between the 16 possible combinations of network properties for "putNet" and "copPut". We used the tool "Venny" [11] [43]. Since for clustering coefficient the mean values differ between the validated data sets and the putative data sets we generated four Venn diagrams. Two diagrams based on the "smaller than" criteria for clustering coefficient values and two diagrams using "greater than", respectively for "putNet" and "copPut".

name	deg	bet	clust	clos	points	<deg	>bet	<clust	<clos	val
mean values	9.67	4624.08	0.12	0.04						
BDNF	10	5082.76	0.04	0.04	2	1	0	0	1	1
RAB11FIP2	9	2008.94	0.13	0.04	3	0	1	1	1	1
TRIM2	2	31.86	0.00	0.04	1	0	1	0	0	1
CFL2	4	56.1	0.33	0.04	3	0	1	1	1	1
SSFA2	1	0.0	0.00	0.04	1	0	1	0	0	1
SRF	51	34360.07	0.04	0.04	2	1	0	0	1	1
PLAG1	2	15.94	0.0	0.04	2	0	1	0	1	1
ARCN1	6	388.25	0.4	0.04	2	0	1	1	0	1
KIAA1598	1	0.0	0.0	0.04	1	0	1	0	0	1
PDCD4	4	136.3	0.0	0.04	2	0	1	0	1	1

Table 2: Example of results for testing if properties are "smaller than" or "greater than" mean values. The table shows a part of the results for "putNet". For every vertex it was tested if the property values differ from the mean values in the way indicated at the top of the table which resulted in boolean values. The last column indicates if this is a validated target.

[11] used April 2008

3.3 Ranking Methods with the Help of GO Analysis

The GO analysis resulted in three lists for each of the reference sets (putative targets, validated targets and cancer genes) used respectively. One list is generated for each the subontologies "Biological Process", "Molecular Function" and "Cellular Location". Furthermore for each subontology and each reference set a graph was produced. We want to know which overrepresented categories of the putative targets are also overrepresented in the validated targets and breast cancer genes. That is why we calculated which GO categories are commonly overrepresented between the three reference lists, for each subontology respectively. We will call a category, overrepresented in the list of putative targets, "supported" by another reference list if it is also overrepresented in this list. In this way we produced four subsets of the overrepresented categories for the putative target genes which "support" the categories overrepresented for putative targets in different ways. We created one subset of categories of the putative targets which are also overrepresented among the breast cancer genes and the validated targets. Since these lists of overrepresented categories for putative targets are also overrepresented for cancer genes and validated targets we will refer to these lists as "doublesupported". We created two subsets of categories of the putative targets which are also overrepresented by either the breast cancer genes or the validated targets (we will refer to this as "cancersupported" and "valsupported" respectively). We used one subset of overrepresented categories of the putative targets which are not overrepresented in any of the other lists.

Having the lists of shared overrepresented GO categories we converted them into lists of genes which are "doublesupported", "cancersupported" or "valsupported". However, some categories might just be overrepresented because their children categories are overrepresented. Also in former studies these kinds of categories were removed from the list of overrepresented categories [13] and the authors of BiNGO [36] state that the most important categories are the overrepresented ones which are placed at a low level in the hierarchy. However, our approach differs in the way that we are more interested in the genes of the commonly overrepresented categories than in the categories themselves. It may happen that a common category (for example a "doublesupported" one) is just overrepresented because its children categories are overrepresented but the children categories are not in the list of supported categories. For this reason copying all genes of the parent category into the list of supported genes would be wrong if this category is just overrepresented because of not supported children categories. That is why we

removed all genes from a parent category which are also members of overrepresented children categories. For this purpose we used the graphs created by Cytoscape [37] which show the hierarchy of overrepresented categories. We looped top-down through the hierarchy of shared overrepresented categories and removed those genes from a parent category which were overrepresented in its children categories. Figure 1 illustrates our approach. In this way we created three lists of genes for the two ontologies "Molecular Function" and "Cellular Component" respectively. One list of "doublesupported", one list of "cancersupported" and one list of "valsupported" genes.

However, for the subontology "Biological Process" all three reference lists (putative genes, validated genes and cancer genes) resulted in a great number of overrepresented categories. For example there are 292 overrepresented categories for the putative targets. A total of 44 categories were commonly overrepresented by all three reference lists. This list contains all putative target genes several times. Thus using our half manual approach to remove inherited overrepresentation was too time consuming. For this reason we decided to take only the ten best hits (smallest p value) for the overrepresented categories of putative, validated and cancer genes, respectively. Using the ten best rated categories we characterised which of them are common between the putative validated and cancer genes. Using this we created lists of "supported" genes were we removed genes which are members of categories which are overrepresented because of their children, exactly as described for the other two ontologies.

Using these lists we tested for the 465 putative target genes in which list they occur and ranked them according to that. Looking at the results obtained so far (see chapter 4.2) it seems that "Molecular Function" might be the most useful for ranking since there is a clear difference between the "doublesupported" categories being more general and the single supported or not supported being more specific. For this reason we decided to give higher score if a gene is a member of these categories. Since the process in which a gene is involved is more important than where it is located we gave more points for "Biological Process" than for "Cellular Component". Furthermore if a gene is in a "doublesupported" list of any subontology we gave higher score than if it is only single supported. Our ranking matrix is shown in table 3.

Finally we created tables which show how many points a putative target gene has and how it was supported.

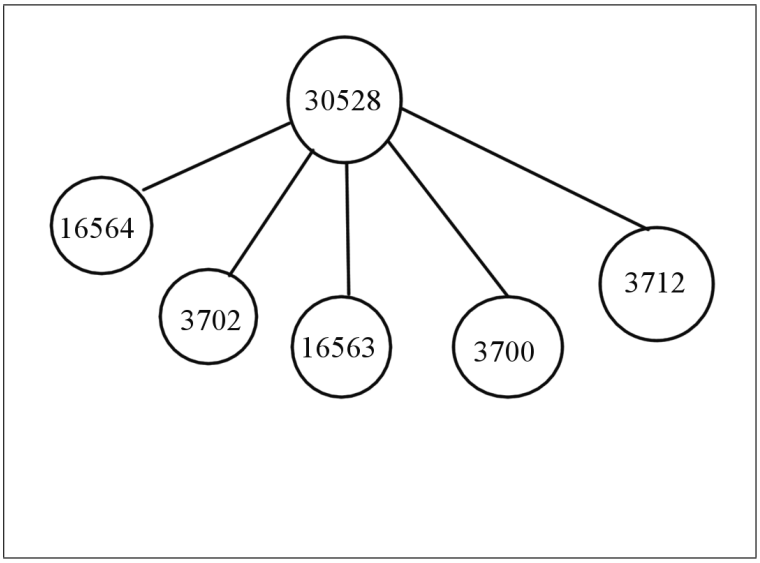


Figure 1: Illustrating our approach for removing genes of children categories which lead to overrepresented parent categories. The figure shows a part of the hierarchy of the subontology "Molecular Function". Category 30528 was overrepresented within the putative targets, the cancer genes and the validated targets and is thus "doublesupported". However, its children 16564 and 3702 are only overrepresented within the putative targets while 3712 is also overrepresented within the cancer genes. 3700 and 16563 are "doublesupported" as well. Our approach removes all genes from 30528 which are members of the children categories and copies the remaining genes into the list of "doublesupported " genes. Then it steps down in the hierarchy and goes on with 3700 and 16563.

Category	Points
doublesupported_MF	4
cancersupported_MF	2
doublesupported_CC	2
validatedsupported_CC	1
cancersupported_CC	1
validatedsupported_BP	2

Table 3: Ranking matrix used for GO Analysis. The matrix shows how many points were given to a putative target which occurred in one of the shown categories after removing inherited genes. Supported lists without any common categories are not shown.

3.4 Combining both Methods

To compare the results of PPIN analysis and GO analysis with each other we characterised the properties of the putative targets, which are best ranked in the GO analysis, within "fullNet". We tested in which subset of combinations of network properties these putative targets are located as described in 3.2. Furthermore we did the same for the putative targets which are only supported by the subontology "Biological Process", since it hast been shown before that they have a greater degree value than the mean value of the human PPIN [7].

3.5 Validation

3.5.1 Properties of Validated MiRNA Targets

We checked for every putative target in "copPut" and "putNet" if it is one of the 22 validated targets, which are members of the 465 putative targets, or the one breast cancer gene and thus checked which combinations of network properties apply for these validated genes. The last column of table 2 shows if this gene is one of the 22 validated (1 if it is validated otherwise 0). Furthermore we located the 22 validated targets and the one gene known to be involved in breast cancer in the Venn diagrams.

The tables created with the help of GO category analysis were evaluated by looking for the 22 validated targets within these lists and by literature research.

3.5.2 Predicted Targets Using latest Tools

MiRNA target prediction tools have been constantly changing and improving. We want to compare our results with putative targets predicted by the latest versions of commonly used tools. In this way we want to study how the changes in the tools change the network property values of the predicted targets. It has been assumed [10] that using the intersection of MiRanda [26], TargetScan [27] and PicTar [20] might be the way of predicting targets with the highest specificity. That is why we decided to use the intersection of these tools to predict putative targets with a small number of expected

false positives. Using a tool developed by Sethupathy et al [12] [10], we predicted targets of the five deregulated miRNAs using the intersection of the mentioned tools, i.e. the tool only predicts a target if it is predicted by all three prediction tools (note that Iorio et al [5] accepted a target if it was predicted by at least two out of three tools). This resulted in a list of 143 putative targets. Out of these 143, 139 are located in the PPIN. Interestingly, only 99 of those putative targets are also putative targets suggested by Iorio et al [5] whereas 44 putative targets are new.

We characterised the network properties of the 139 putative targets predicted by using the latest versions of commonly used prediction tools within "fullPIN". As described in chapter 3.2 we calculated mean network property values for these putative targets within "fullPIN" and checked for every vertex if its property value differs from the mean value of "fullPIN". Using this we also calculated the fraction of how many vertices have different network property values or combinations of them divided by the total number of used genes (139). Exactly the same was also done for those 44 putative targets out of the 139 which were not predicted by Iorio et al [5].

3.5.3 Literature Research

Literature research was done for the following sets of genes:

1. The ten best ranked genes of "copPut" using the "smaller than" criterion for clustering coefficient values.
2. The ten best ranked genes of "copPut" using the "greater than" criterion for clustering coefficient values.
3. The ten worst ranked genes of "copPut" using the "smaller than" criterion for clustering coefficient values.
4. The three genes which were best ranked using GO analysis and network property analysis.
5. The seven further genes best ranked using only GO analysis.
6. The ten worst ranked Genes using GO analysis.

[12] used May 2008

We used "copPut" because the betweenness values of the subnetworks are small because of the way the networks were constructed (see chapter 4.1). For 1) and 2) best ranked means that for those genes all four criteria for network properties apply. Since this is the case for more than ten genes, we decided for the genes with the highest degree values because the validated and putative targets constantly have greater degree values than the mean of "fullNet" which has been reported by other studies [6, 7, 11]. To compare the results of the "smaller than" criterion for clustering coefficient values with the results of the "greater than" criterion we did literature research for both. The ten worst ranked genes are those for which the smallest number of criteria apply, i.e. for four of them none applies and for 6 only the criterion for closeness centrality applies.

For GO we used those three genes which are best ranked for GO analysis and using criterion 1). The further seven genes were chosen randomly out of the twenty best ranked. We used the ten worst ranked which are members of none of the commonly overrepresented categories.

We used the "GeneCards" [13] [44] Database to gain information about genes and diseases in which they are involved. For every gene under consideration we saved its function, a maximum of three cancer related diseases in which the gene is involved, the number of articles in which the gene is mentioned together with breast cancer and the three best ranked out of those articles. The "GeneCards" database gives many information about the function of a gene. We summarized them with a few words.

For every gene the database shows a table which summarises results of literature research. For 92 diseases it is given a score of the relevance of the disease to this gene which is based on the analysis of co-occurrences of the gene and the disease in PubMed articles. It uses a hypergeometric distribution of the number of articles in which the gene and the disease occur together and the number where they occur independently. For every gene under consideration we considered the ten highest scored diseases. If cancer related diseases were among those ten, we saved the three highest scored of them.

Furthermore the table shows the number of articles in which the gene and the words "breast cancer" are mentioned together in one sentence. We saved this number and the three best ranked articles for every gene under consideration (here the score is simply the number of sentences in the article in which the gene and "breast cancer" are mentioned

[13] accessed June 2008

together).

4 Results

4.1 Ranking with the Help of PPIN Analysis

In order to investigate methods to rank putative target genes with the help of PPIN properties we first characterised a set of properties for the (sub)networks "fullNet", "putNet" and "valNet", as well as for the copied properties "copPut" and "copVal". For every data set we calculated the mean degree, betweenness centrality, clustering coefficient and closeness centrality values. Table 4 summarises the results.

All four data sets have greater mean degree values than "fullNet" which proposes this property to be most consistently different from the whole human PPIN. Since the subnetworks "putNet" and "valNet" include all neighbours of the putative or validated targets the mean degree values of these data sets should be the same as the mean degree values of "copPut" and "copVal" respectively. The small differences seem to be due to rounding problems.

The mean closeness centrality value is also greater for all four data sets than in "fullNet". However, the values of the subnetworks are more than 10 times greater than the mean value of "fullNet" while the other mean values are only slightly greater.

The mean betweenness centrality value is smaller for the subnetworks but greater in the copied data sets than in "fullNet". The reason for this could be the construction of the subnetworks which only include the first neighbours of the putative and validated targets. Because of this the number of unique paths passing through a vertex in these networks is smaller and this leads to smaller betweenness centrality values. Even though this would suggest that using subnetworks is not a good way to rank we used these data sets in further steps to prove or disprove this proposal.

The mean clustering coefficient value is greater for the validated target sets but smaller for the putative ones than the mean values of "fullNet". This result is quite interesting since it has been proposed before that miRNA regulated proteins have a smaller clustering coefficient than the mean degree value of the human PPIN [7].

To test the significance of these results the Mann-Whitney-Wilcoxon (MWW) signed rank test [41] was used for all four properties and all five data sets in a pairwise way.

Property/Network	Degree	Betweenness	Clust. Coef.	Closenes
fullNet	7.48	14185	0.1047	0.003829
putNet	9.76	4624	0.1210	0.0412
valNet	13.49	5756	0.0825	0.1022
copPut	9.67	18164	0.1210	0.003895
copVal	13.51	34389	0.0825	0.0039

Table 4: Mean Values of Network Properties

The result is a p value which is the probability of observing these different mean values by chance. Pairwise results are shown in table 5. The results show that the mean degree values of all four data sets differ significantly from "fullNet" while there are no significant differences between the putative targets and validated targets in both the subnetworks and copied data sets.

The closeness centrality values show the most consistent differences between all data sets. However, for the copied data sets "copPut" and "copVal" closeness centrality values do not differ significantly between the putative and validated data sets.

For the betweenness centrality values the differences between the "fullNet" and validated targets in the subnetwork are not significant while all the others are.

The differences for the mean clustering coefficient values are only significant between "fullNet" and the data sets of putative targets. However, there are no significant differences between all the other data sets even though the mean values of the putative targets are greater and those of the validated targets are smaller than the mean value of "fullNet".

The differences between "putNet" and "copPut" on the one hand are only significant for closeness centrality values and degree values. The differences between "valNet " and "copVal" on the other hand are only significant for betweenness centrality values and degree values. Based on this we can not decide if using a subnetwork or not is a better way. Furthermore there seem to be only few significant differences between the validated and the putative targets. However, since both differ strongly from the whole PPIN it should be possible to find a combination of network properties which contributes mostly to these differences.

In the next step we studied which network properties or combinations of network

Samples	Degree	Betweenness	Clust. Coef.	Closeness
FullNet-putNet	9.36e-8	0.00110	2.98e-6	2.2e-16
FullNet-valNet	3.30e-8	**0.4133**	**0.9857**	2.2e-16
FullNet-copPut	9.366-8	0.00010	2.98e-6	1.7e-10
FullNet-copVal	3.02e-8	2.73e-8	**0.9857**	7.48e-8
PutNet-valnet	0.04870	0.02952	**0.1523**	2.2e-16
PutNet-copVal	0.04703	5.0e-12	**0.1523**	2.2e-16
PutNet-copPut	**0.5000**	1.75e-7	**0.4952**	2.2e-16
ValNet-copVal	**0.4953**	2.49e-6	**0.5002**	2.2e-16
ValNet-copPut	0.04870	0.00655	**0.1523**	2.2e-16
CopVal-CopPut	0.04703	0.00400	**0.1523**	**0.1754**

Table 5: Significance tests for Network Properties. The p value indicates the probability of observing the different mean values per chance. All tests are two tailed. Bold values indicate a non significant result by using $p = 0.05$

properties contribute the most to ranking. Furthermore we compared "putNet" and "copPut" with each other and their validated data sets in order to answer the question if it is better to use subnetworks or not for ranking. To do so we calculated how many vertices have greater/smaller values (depending on how the mean property differs from the mean of "fullNet", see chapter 3.2) than the mean values for all four network properties and all 16 combinations of them. Table 6 shows the fractions of how many vertices have greater/smaller (combinations of) properties than the mean values for "putNet", "copPut","valNet" and "copVal" respectively. Using the data set "copPut" we calculated these data for two sets of the four mean network properties. Once for the mean properties of "copPut" itself (fifth column of table 4) and once for the mean properties of "fullNet" (second column of table 4). Since the mean values of "copPut" itself were generally greater than these from "fullNet" all fractions were smaller (data not shown). This approach seems to be too restrictive.

Table 6 does not show a similar "pattern" between the validated and putative target sets but some conclusion can be drawn using it. Generally, the more properties you combine, the smaller the fractions get, which indicates that the method becomes more restrictive. Especially in the total intersection ("all") the values are quite small, but these putative targets might be the most interesting.

The betweenness centrality data sets show quite different values for "putNet" and "valNet" on the one hand, and "copPut" and "copVal" on the other hand. The reason

for this is that the mean value of the subnetworks was smaller than the one of "fullNet" while the mean value of the copied data sets was greater. That is why we tested "smaller than" in the first case and "greater than" in the second.

The results for clustering coefficient differ largely between the validated and the putative target sets. Here the reason is that we used for the "smaller than" criterion for the validated data sets and the "greater than" criterion for the putative data sets. The values for smaller clustering coefficient are 0.73 and 0.80 for "putNet" and "valNet" , respectively, and thus quite similar. The reason for the high number in the total intersection of "copVal" is also the usage of the "smaller than" criterion for clustering coefficient values in this data set. With using a "greater than" criterion the fraction in the total intersection is 0.03.

Calculation of closeness centrality resulted in high values for all data sets. This indicates that it might not be very helpful to rank since most of the vertices have a greater closeness centrality value than the mean value even though the data sets for closeness centrality differ most significantly from each other. Combinations of closeness centrality with any other property often leads to the same result as using this property alone which indicates that in these cases closeness does not bring any more information. For example the combination of degree and closeness centrality leads to exactly the same results as using only degree.

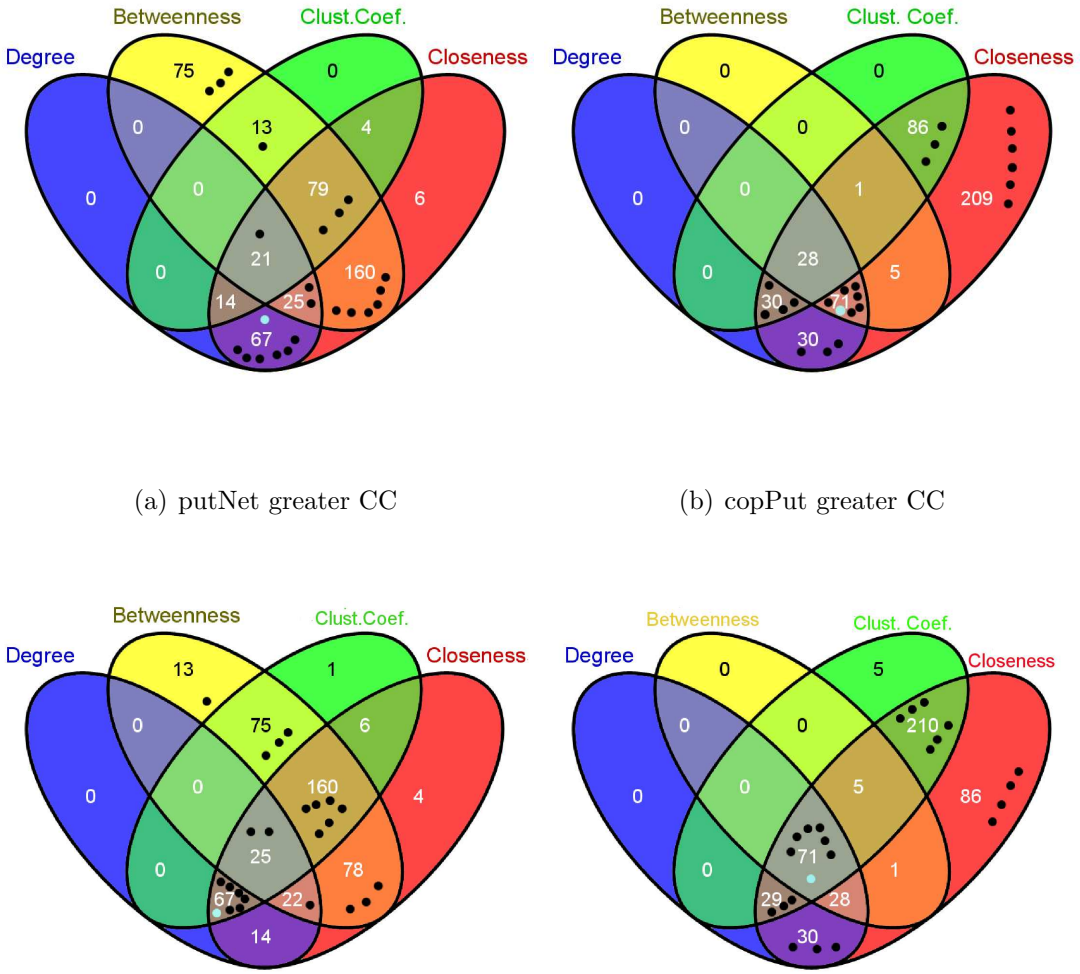
Almost the same data is shown in figure 2 for "putNet" and "copPut" but for easier interpretation it is visualised with the help of Venn Diagrams. The only difference is that putative targets which are not members of any of the sets are not shown (column "none" in table 6) since we would need a fifth set for these. The number of these putative targets is 4 and 5 for "putNet" and "copPut", respectively. Venn Diagrams are shown for both the "greater than" criterion for clustering coefficient (figure 2(a) and 2(b)) which was suggested by the putative target sets and the "smaller than" criterion (figure 2(c) and 2(d)) which was suggested by the validated data sets.

For further evaluation we located the 22 validated miRNA targets and the one known breast cancer gene which are members of the 456 putative targets in the Venn diagrams. Figure 2 shows in which subsets these proteins are located.

The diagrams show that the putative targets are not members of a distinct subset of a combination of network properties. Rather you can find them in different kinds of

Dataset/Prop.	D	B.	U	O	none	
PutNet	0.273	0.802	0.281	0.808	0.002	
CopPut	0.340	0.226	0.312	0.989	0.011	
ValNet	0.230	0.812	0.718	0.591	0.000	
CopVal	0.403	0.291	0.751	0.995	0.000	
pairs	D&B	D&U	D&O	B&U	B&O	U&O
PutNet	0.098	0.075	0.273	0.243	0.612	0.253
CopPut	0.213	0.125	0.340	0.062	0.226	0.312
ValNet	0.079	0.159	0.230	0.568	0.403	0.380
CopVal	0.281	0.300	0.403	0.248	0.291	0.746
triples	D&B&U	D&B&O	D&U&O	B&U&O	all	
PutNet	0.045	0.098	0.075	0.215	0.045	
CopPut	0.060	0.213	0.125	0.062	0.060	
ValNet	0.032	0.079	0.159	0.230	0.032	
CopVal	0.239	0.281	0.300	0.248	0.239	

Table 6: Fractions of how many vertices have greater/smaller values of properties (and combinations) than mean values of the properties. Properties of vertices from "copPut" and "copVal" were compared with mean values of "fullNet". Abbreviations: D=Degree, B=Betweenness, U=Clustering Coefficient, O = Closeness.



(a) putNet greater CC

(b) copPut greater CC

(c) putNet smaller CC

(d) copPut smaller CC

Figure 2: Venn Diagrams showing overlapping of Network Properties. The diagrams show for how many putative targets combinations of the criteria for network properties apply. For example, there are 67 targets for which the criteria for degree and closeness centrality (and only those) apply in subfigure a). For a) and b) "greater than" was used for clustering coefficient (CC) while in c) and d) "smaller than" was used. "CopPut" values were compared with mean values of "fullNet". The black dots show in which subsets the 22 validated targets are located and the cyan dot shows were the known breast cancer genes is.

subsets. However, generally the more properties you combine the less putative targets can be found. Although the validated targets are not members of a distinct subset they are more often in the subsets of combined properties than in a set where only one criterion based on one property is satisfied.

In all four cases there is a great number of both putative and validated targets within the subsets generated based on closeness centrality values. In "copPut" all, or all but five putative targets are located in the subset generated based on closeness centrality. Again, closeness centrality does not seem to contribute a lot to ranking.

In all four cases there are 11 validated out of 158 putative targets within the sets of degree. Thus degree maximizes the ratio of maximum validated targets and minimum putative targets.

The main difference between "putNet" and "copPut" is that for the first ones we used the "smaller than" criterion for betweenness centrality values while we used the "greater than" criterion for the second ones because the mean value was greater than the one of "fullNet" in these cases. For "putNet" there are 16 validated targets included in the subsets generated based on betweenness centrality values while in "copPut" there are only seven but the number of putative targets is 105 and 348 respectively. However, it seems that the "greater than" criterion for betweenness centrality prevents some validated targets to be more "central" in the case of "putNet". For example there are seven validated targets in the intersection of degree, closeness centrality and clustering coefficient in figure 2(c) which move to the total intersection in figure 2(d).

Using the "smaller than" criterion or the "greater than" criterion for clustering coefficient changes also both the number of validated and the number of putative targets. The number of putative targets having a greater clustering coefficient is around 130 in both data sets but with a smaller one there are more than 300. This also results in a greater number of validated targets having a smaller clustering coefficient. The most interesting difference might be the subset based on degree, closeness centrality and betweenness centrality. There are seven validated targets out of 71 in "copPut" located in this subset (figure 2(b)). Using a smaller clustering coefficient exactly those genes switch places with those genes in the total intersection which can be seen in figure 2(d).

The mean network property values of the putative targets predicted by the latest versions of commonly used tools (see chapter 3.5.2) within "fullNet" are mostly similar

to those reported for "copPut", i.e. the mean degree, clustering coefficient and closeness centrality values are slightly greater than those of "copPut". The mean betweenness centrality values is instead 31065 and thus almost as great as for "copVal". This might indicate that the better the tools the greater become the mean between centrality values of the putative targets, which is supported by the validated targets having a great mean betweenness centrality value. This is also proposed by the 44 putative targets which were predicted by the latest versions of the tools but not by Iorio et al [5]. These putative targets have similar mean network property values as "copPut" as well but a mean betweenness centrality value of 47047. The mean network property values are slightly greater for the 139 genes than in "copPut". Because of this the fractions of vertices having a greater network property value than the mean of "fullNet" divided by the number of all used genes (139) are also slightly greater than in "copPut". For example the fraction of the total intersection where the criteria for all network properties apply is 0.071.

Table 7 summarises results of the literature research. The first and the second part of the table show results for the top ranked (see chapter 3.5.3) genes of "copPut" for the "smaller than" and "greater than" criterion respectively. The third part shows results for the ten worst ranked genes. The top ten putative targets using the "smaller than" criterion are mostly involved in functions as transcription regulation or signal transduction. For almost all of them there are cancer related articles among the top ten articles of diseases in which the genes are mentioned. Moreover breast cancer is often among the top three of these types of cancer related articles and for almost every gene there was breast cancer related literature found in the database. Especially for the genes SHC1 and IRS1 there were 40 and 96 relevant articles found respectively. This might imply that this approach is capable to find breast cancer related genes. Only for the protein NEDD9 there was no cancer related article found.

However, only for five of the ten best ranked putative targets using the "smaller than" criterion cancer related article were among the top ten disease related articles and only for one gene a breast cancer related article was found.

For the ten worst ranked genes was no breast cancer related article found and only for one of them cancer related articles are among the best ten ranked articles. This shows that our approach is able to rank genes which are involved in breast cancer highly while genes which may not be involved in breast cancer are lowly ranked.

Gene	Function	Cancer	Nb	Ref
SHC1	signaling adapter	men 2a,breast cancer,men 3	40	[45],[46],[47]
CRK	transforming activity	myeloid leukemia chronic	1	[48]
IRS1	insulin receptor substrate	breast cancer	96	[48],[49],[50]
CEBPB	transcript. activ.	choriocarcinoma,tumors,leukemia	1	[51]
SOCS1	regulates signal transd.	carcinoma,colorectal cancer,tumors	1	[52],[53],[54]
NEDD9	docking protein		0	
NCOA6	nucl. receptor coact.	cancer	0	
NRIP1	transcript. Act.	carcinoma embryonal,breast cancer,cancer	8	[55],[56],[57]
ETS1	transcription factor	leukemia,tumors,leukemia t-cell	16	[56],[58],[59]
BTRC	mediates ubiquitination	tumors,cancer,colorectal cancer	6	[60],[61]
YES1	phosphatase	colon can.,mammary tumor,colon carci.	0	
RPS6KA3	serine/threonine kinase		0	
RHOQ	GTPase		0	
GJA1	gap junction protein	carcinoma giant cell,tumors	0	
MAP3K10	act. JUN N-term. pathway		0	
MAP3K11	act. JUN N-term. pathway	tumors	0	
KPNA1	nuclear protein import		0	
ACVR2A	ligand binding	colon cancer,pancreatic cancer,tumors	0	
SDC1	surface proteoglycan	carcinoma,breast cancer	1	[62]
RGS7	inhibits signal transd.		0	
ATRX	transcript. regul.		0	
SMNDC1	spliceosome assemb.		0	
PURB	DNA binding		0	
GGTL3	cleaving		0	
PRPF4B	splicing		0	
SSFA2	actin binding		0	
POMT2	transfers mannosyl		0	
HCN4	ion channel		0	
GRHL1	transcript. factor		0	
ABCG1	transport		0	
MAT2A	catalyzis	carcinoma,cancer,tumor	0	

Table 7: Literature research results for PPIN property analysis. The first part shows information for the ten best ranked putative targets of "copPut" using the "greater than" criterion for clustering coefficient values and the second part the same for using the "smaller than" criterion. The last part shows results for the ten worst ranked genes. Nb=Number of articles found.

4.2 Ranking with the Help of GO Analysis

With the help of BiNGO [36] we calculated which GO categories are overrepresented in the reference lists of the putative targets the validated targets and the cancer genes for the subontologies "Molecular Function", "Biological Process" and "Cellular Component". Then we studied which categories are commonly overrepresented between the given reference lists for all three subontologies. Results are shown in table 8, table 9 and table 10 for the subontologies "Molecular Function", "Cellular Component" and "Biological Process" respectively. Note that for the last one we only used the ten best hits for each reference set to generate this list as described in 3.3.

The most common categories for the subontology "Molecular Function" are related to transcription and binding. It seems that the miRNA targets themselves regulate the expression of other genes. While the commonly overrepresented categories are very high in the hierarchy those overrepresented by the cancer genes and putative targets and the "notsupported" categories are mostly special cases of the "doublesupported" ones. We removed the genes from those "doublesupported" categories which are only overrepresented because of their children categories. For example 94% of the genes of category 30528 are members of children categories so this category may be overrepresented because of inheritance. However, the children categories 3702 and 16564 are in the list of "notsupported" categories. For this reason we removed those genes from the list of "doublesupported" categories. But there do also exist some categories in the "cancersupported" list which are not children of the "doublesupported". It seems that cancer genes and putative targets both play a role in signal transduction since both have overrepresented categories like Kinase Activity of Receptor Activity and so on. This kind of regulatory activity function is not overrepresented for the validated targets. The "notsupported" categories are mostly special cases of the "doublesupported" and "cancersupported" ones which suggests that the putative targets may have a more specific function but do not have a totally different function.

For the subontology "Cellular Component" (table 9) it is not the case that the highest categories in the hierarchy are "doublesupported" while the more special ones are either single or not supported. Here the "doublesupported" categories are either in the nucleus or a part of the membrane. While the "validatedsupport" are mostly intracellular the "cancersupported" are mostly parts of the lumen. For this subontology the non supported categories are not directly specialized categories of supported ones.

GO ID	Name
doublesupported	
8134	transcription factor binding
3700	transcription factor activity
5488	binding
16563	transcription activator activity
5515	protein binding
30528	transcription regulator activity
46983	protein dimerization activity
cancersupported	
16772	transferase activity
4713	protein-tyrosine kinase activity
16773	phosphotransferase activity
166	nucleotide binding
19899	enzyme binding
42802	identical protein binding
16301	kinase activity
5057	receptor signaling protein activity
50222	protein kinase activity
17076	purine nucleotide binding
3712	transcription cofactor activity
3677	DNA binding
notsupported	
4674	protein serine/threonine kinase activity
3702	RNA polymerase II transcription factor activity
43565	sequence-specific DNA binding
3676	nucleic acid binding
5112	Notch binding
3714	transcription corepressor activity
3723	RNA binding
16564	transcription repressor activity

Table 8: Supported GO categories for "Molecular Function". Shown are GO categories which are overrepresented within the list of putative targets, validated targets and cancer genes ("doublesupported"), those which are commonly overrepresented for the putative targets and the cancer genes ("cancersupported") and those which only are overrepresented for the putative genes ("notsupported"). There is no category overrepresented for the putative targets and the validated ones.

Either there are some categories in between or there is no connection to the supported categories at all. However, some of them are higher in the hierarchy than the supported ones.

For the subontology "Biological Process" analysis we took the ten best hits for every reference list since there simply were too many. Within the best ten there were only common categories between the putative and validated targets. Especially regulatory processes are overrepresented here as it was shown in previous studies [7].

Table 11 shows the twenty best ranked putative targets using GO analysis. The second row indicates how many putative targets are in the list of any commonly supported category used. The last column indicates if the gene was one of the 22 validated. Only three of them have 11 points while the others are not that high ranked. However, mostly they are not supported by "cancer_MF" or "cancer_CC" which is not surprising since they are not targets of those five miRNAs involved in cancer. All together there is one putative target having 12 points and 37 having 11 points. All of them occur in the same lists as the 20 shown. These genes might be quite interesting since they share common features with both validated targets and breast cancer genes.

Table 12 shows results of the literature research. It indicates that for the three genes which are best ranked in GO and PPIN analysis cancer related articles were among the top ten of disease related articles and that for all of them breast cancer related articles exist. Moreover the further seven genes are also involved in cancer related diseases which are among the best ranked and for CTLA4 breast cancer related articles exist. Only two out of the ten worst ranked genes are involved in cancer related diseases which are among the ten best ranked and for only one breast cancer related articles were found. This indicates that this approach is able to discriminate between cancer related putative targets and none cancer related ones.

4.3 Combining both Methods

We characterised network properties of the 20 best ranked putative targets in the GO analysis within "fullNet". It is not the case that these targets are members of a distinct subset of combinations of network properties which differ from the mean values. All of them have greater closeness centrality values since almost all putative targets have

GO ID	Name
doublesupported	
43231	intracellular membrane-bound organelle
43227	membrane-bound organelle
5654	nucleoplasm
5634	nucleus
validatedsupported	
5737	cytoplasm
44451	nucleoplasm part
5622	intracellular
43229	intracellular organelle
43226	organelle
44424	intracellular part
cancersupported	
43233	organelle lumen
31974	membrane-enclosed lumen
44428	nuclear part
31981	nuclear lumen
notsupported	
14069	postsynaptic density
32991	macromolecular complex
16281	eukaryotic translation initiation factor 4F complex
5802	trans-Golgi network
5667	transcription factor complex
44464	cell part
5623	cell
43234	protein complex

Table 9: Supported GO categories for "Cellular Component". Shown are GO categories which are overrepresented within the list of putative targets, validated targets and cancer genes ("doublesupported"), those which are commonly overrepresented for the putative targets and the cancer genes ("cancersupported"), those commonly supported for putative targets and validated targets ("valsupported")and those which are only overrepresented for the putative genes ("notsupported").

GO ID	Name
validatedsupported	
6357	regulation of transcription from RNA polymerase II promoter
50791	regulation of biological process
51244	regulation of cellular process
6366	transcription from RNA polymerase II promoter
65007	biological regulation

Table 10: Supported GO categories for Molecular Function. Shown are the GO categories which are overrepresented within the list of putative targets and validated targets. Since we only used the best ten overrepresented categories for every reference list there is no "doublesupported", "cancersupported" or "notsupported" category.

greater closeness centrality values in this set. However, there is no significant connection between the best ranked GO results and genes which have a greater degree, betweenness centrality or clustering coefficient value. Our results do not suggest any connection between GO and PPIN properties. However, there are three genes which are in the total intersection of all properties (see figure 2(b)) and which are highest rated. The literature research indicates that those three genes are involved in cancer (see table 12).

gene	points	doubMF	canMF	doubCC	valCC	canCC	valBP	validated
number		320	163	189	335	22	219	22
MADD	12	1	1	1	1	1	1	0
HDAC4	11	1	1	1	1	0	1	1
SP1	11	1	1	1	1	0	1	1
MXD4	11	1	1	1	1	0	1	1
VPS4B	11	1	1	1	1	0	1	0
HIC2	11	1	1	1	1	0	1	0
SMARCD2	11	1	1	1	1	0	1	0
SOX9	11	1	1	1	1	0	1	0
SMARCA4	11	1	1	1	1	0	1	0
CEBPB	11	1	1	1	1	0	1	0
EWSR1	11	1	1	1	1	0	1	0
WEE1	11	1	1	1	1	0	1	0
NFE2L1	11	1	1	1	1	0	1	0
NCOA6	11	1	1	1	1	0	1	0
ARNT	11	1	1	1	1	0	1	0
LDB1	11	1	1	1	1	0	1	0
SCG2	11	1	1	1	1	0	1	0
SMARCD1	11	1	1	1	1	0	1	0
MEF2D	11	1	1	1	1	0	1	0
TCEB3	11	1	1	1	1	0	1	0

Table 11: The 20 top ranked putative targets. The table shows the top 20 ranked putative targets using GO. For each gene the total points and the occurrence in every of the supported lists is shown. The last column indicates if the gene is one of the 22 validated genes. Note that supported lists without any common categories are not shown.

Gene	Function	Cancer	Nb	Ref
CEBPB	transcription activ.	choriocarcinoma,tumors,leukemia	1	[51]
NCOA6	nucl. receptor coact.	cancer	0	
SP3	trancription factor	retinoblastoma,breast cancer,leukemia t-cell	17	[63],[64],[65]
MADD	regul. cell prolifer.	tumor,cancer	0	
HDAC4	deacetyl. lysine resid.	leukemia ,cancer,tumors	0	
SP1	gene activ.	erythroleukemia,retinoblastoma,tumors	51	[66],[67],[63]
MXD4	transcription repr.	tumors	0	
VPS4B	protein transport		0	
HIC2	transcription repr.	leukemia lymphocytic	0	
SMARCD2	transcription activ.		0	
SOX9	skeletal development		0	
FBXO28	binds proteins		0	
FZD7	receptor	hepatocellular carcinoma,gastric	0	
SLC38A2	amino acid transp.		0	
CTLA4	t-cell activ.	melanoma	10	[68],[69],[70]
IL1RAPL1	neurotransm. releasing*		0	
CTDSPL	phosphatase*		0	
C10ORF12	unknown			
PPP1R3A	glycogen-targeting*		0	
PLEKHA8	membrane transport		0	
TP53INP1	promotes phosphoryl.		0	

Table 12: Literature research results for the ranking approach using GO. The first three genes are best ranked in both GO analysis and PPIN analysis. The next seven are best ranked in GO analysis. The last ten are worst ranked in GO analysis. An asterik indicates that the stated function is speculative. Nb=Number of articles found.

5 Conclusions

5.1 Contributions

It has been shown that miRNA regulated proteins have significantly different network property values than the mean values of the human PPIN [6, 7]. The same has been shown for cancer associated proteins [11]. This report studies the possibility of using these differences to rank putative miRNA targets according to their biological relevance. We can show that, using our simple approach, different network properties contribute differently to the ranking. Furthermore for the best ranked putative targets we can often find breast cancer related literature while for the worst ranked putative targets only a few articles exist. Among the best ranked putative targets might be genes which may play a role in cancer which is not known yet.

Overrepresented GO categories of putative miRNA targets have been identified in former studies e.g. by Stark et al [13]. However, we are making the first study of approaches to use these categories to rank putative miRNA targets. For the first time this study identifies commonly overrepresented categories of putative miRNA targets, validated miRNA targets and genes involved in breast cancer. We try to use these categories for ranking. Again among the best ranked putative targets exist more cancer and breast cancer related articles than among the worst ranked.

5.2 Main Conclusions

5.2.1 Conclusions using PPIN Properties

This report describes a first attempt to use network properties of putative miRNA target genes to rank the biological relevance of the putative targets. We characterised a set of properties within the whole human PPIN and compared it to sets of validated and putative targets. Using these results we tried to find ranking methods based on the simple boolean criteria whether the values of the properties are "greater than" or "smaller than" the mean values of the data sets and studied how much this contributes to the ranking for every property and every combination of properties.

The mean network property values of both the putative and the validated targets differ significantly from those of the whole human PPIN. This should make it possible to use these differences in order to rank the relevance of a putative target gene.

The mean degree values differ most consistently between the whole human PPIN and all other data sets. Furthermore using the "greater than" criteria for degree leads to the greatest ratio of validated targets in the data set divided by the total number of putative targets. This may suggest that using degree for ranking has a high specificity.

Another measurement whose data sets differ consistently and significantly from the whole human PPIN is closeness centrality. The mean values suggest that the putative and validated targets have greater mean closeness centrality values. However, only a very small number of vertices have smaller closeness centrality values which makes it very difficult to rank based on a "greater than" criterion for the mean value.

The mean clustering coefficient value is greater for the validated targets but smaller for the putative ones. This is the biggest difference between the putative data sets and the validated ones and might propose a way to rank putative targets. For the latest versions of prediction tools the mean clustering coefficient value is slightly greater than for the older data sets which might support this idea. However, the number of putative targets having a smaller clustering coefficient than the mean one is high which might indicate a small specificity. Using the "smaller than" criterion for clustering coefficient in combination with "greater than" for degree also leads to a great number of validated targets but a small number of putative targets. The literature research suggests that using the "smaller than" criterion you can find more cancer related articles for the best ranked putative targets than using the "greater than" criterion.

Using a subnetwork to characterise the network properties of putative targets affects the betweenness centrality values because it reduces the number of unique paths crossing a vertex. This might indicate that you should not use this measurement in combination with subnetworks. Furthermore the mean betweenness centrality values within the whole human PPIN is greater for the putative targets and even greater for the validated targets than the mean value of the whole human PPIN. For the predicted targets of the latest tools it is also greater than for the older list. This may suggest that improving of the tools leads to more putative targets with greater betweenness centrality values which is supported by the results of validated tools.

The literature research shows that this approach is able to find breast cancer related genes and rank them best. This may help to identify genes among the best ranked which are involved in cancer but which are not known yet. Moreover we can also show that genes which are worst ranked may not be involved in breast cancer.

5.2.2 Conclusions of GO Analysis

This report used GO category analysis to rank the biological relevance of putative miRNA targets. We used GO categories which are commonly overrepresented between the putative targets, validated targets and known breast cancer genes. Using this we created lists of genes which are members of commonly overrepresented targets.

For every subontology there is a number of categories which are overrepresented for the putative targets and for the cancer genes and/or the validated targets as well. For the subontology "Molecular Function" you can see a clear distinction between the categories which are commonly overrepresented for all reference lists being more general and those which are overrepresented for two reference lists or only for putative targets being more specific. Furthermore we identified a set of categories which is shared by the putative targets and the cancer genes and plays a role in signal transduction. Using the subontology "Cellular Component" there is no clear distinction between categories which are overrepresented for all reference lists and those only for one or two.

This study showed that some categories of the putative targets are only overrepresented because of over representation of their children categories. If those children categories are not overrepresented for the validated targets and/or cancer genes it is important to remove the genes from the supported lists. We tried this in the way that we removed the genes of categories which lead to over representation because of inheritance in a top-down approach. However, for the subontology "Biological Process" this semi manual method is time consuming.

Also for this approach we can show that genes which seem to be involved in cancer are highly ranked and genes which seem not to be involved are worst ranked. This approach is able to discriminate between genes which are involved in breast cancer and genes which may not be involved.

5.3 Discussion

A possible source of problems for this report could be the fact that genes which are involved in cancer are more studied and for this reason they might have a greater network property value or occur in a more specific GO category. For instance the number of interaction partners for a well studied gene might be quite large but for a gene that is not well-studied, there may only be a few known interacting partners. In this case our approach might just indicate known genes. However, with the increasing number of highthroughput experiments proteins with interesting network properties or functions, which are not well studied yet, might be found and our approaches may identify them to be important.

Our study compared the mean values of network properties from the whole human PPIN to validated miRNA targets and putative miRNA targets. Furthermore we compared the network property values of every vertex to the mean values of the data sets with "greater than" and "smaller than" criterion. These are very simple criteria and might not always be the best choice. Especially for closeness centrality most vertices have greater values than the mean value which might just be because of outliers. Here it might be better to use other statistical values such as median. The results of the significance analysis indicated that the differences between the data sets of clustering coefficient values are not significant which may explain the differences between our study and another study which found a greater mean clustering coefficient value for putative targets [7].

We chose the unpaired Mann-Whitney-Wilcoxon (MWW) signed rank test [41] to test the statistical significance of the results. Even though this test works with different sizes of samples the big differences in the sample sizes here might lead to some problems. While the whole human PPIN consists of more than 9000 vertices, the validated and putative targets have several hundreds. Another way to test would be to choose random samples from the whole PPIN and compare them to the other samples.

Two other studies have shown that miRNA targets have greater degree values than the mean value of the whole PPIN [6, 7]. Taken this together with our results it seems that miRNA targets interact with more partners than other proteins. A reason for this could be that those highly interactive proteins are vital for many important cell functions [71] and thus their expression level needs to be controlled with the help of

different mechanisms to make sure they are expressed at the right time and place [7]. Furthermore proteins involved in cancer also have a greater degree value than the mean value [11] which supports our findings.

The other network properties of miRNA targets are not that well studied. Liang et al [7] reported that miRNA targets are frequently so called "intramodular hubs" which connect different modules and thus have a high degree value but small clustering coefficient value. In the case of validated targets we report the same results but not for the putative ones. This might indicate a difference between validated and putative targets. However, it might also be because of the small data set of predicted targets used in this report since Liang et al [7] predicted more targets.

We used BiNGO [36] to calculate overrepresented GO categories. There is a number of different tools for this purpose which calculate the p values in different ways [8]. A general problem are categories which are overrepresented because of the hierarchy of the GO. There are tools which do not result in these kinds of categories and remove inherited over representation directly, for example "GOStats" [72]. However, this may lead to fewer commonly overrepresented categories and thus the number of genes in the common lists will be smaller. We decided to use the graphs generated by BiNGO [36] and to remove the genes which lead to over representation instead of the categories. However, for the subontology "Biological Process" this semi manual approach is time consuming. One idea would be to use a smaller p value for this subontology. It has been shown that there is a correlation between predicted miRNA targets having a greater degree value than the mean value of the whole PPIN and being a member of distinct overrepresented GO categories [7]. Our data sets do not show this correlation which might be due to the fact that we only used the ten best overrepresented GO categories for every reference list for this subontology.

Given the important roles which miRNA targets seem to play in the cell and especially in cancer development the prediction of miRNA targets will be an important issue in the future. One problem is the number of false positives which is difficult to determine since we can not be sure that a gene is not a target without a wet-lab experiment. Given that there are only 20 genes experimentally validated not to be targets [1] it has been suggested to use the total number of predicted miRNAs to address the false positive rate of a prediction tool [10]. Our approach may help to find a new way of ranking the importance of a predicted target.

The results of the literature research for the PPIN analysis indicate that using the "smaller than" criterion for "copProp" identifies top ranked genes, for which most often breast cancer related articles exist. On the other hand using the "greater than" criterion the top ranked putative targets do not include many genes for which breast cancer related articles exist. Moreover we show that for the worst ranked genes no breast cancer related articles are found. Among the top ranked genes IRS1 is mentioned in the most articles related to breast cancer. Insulin is a hormone which is known to control the blood sugar level but which also acts as growth factor. It has been shown that a high insulin level in combination with changes in the insulin signaling pathway is associated with breast cancer [73]. IRS1 is involved in this pathway. 40 breast cancer related articles exist which mention the gene SHC1. This gene is a second messenger. Changes in this gene affect the normal metastasis which is associated with breast cancer [46]. Isoforms of SHC1 are also used as drug targets. This shows that our approach is able to find genes which may play an important role in breast cancer. Also for the GO analysis the literature research shows that our approach is able to discriminate between genes which are involved in cancer and those which not seem to be involved. Here the gene GALA4 seems to be an exception since there are breast cancer related articles but it is worst ranked. The problem might be that this genes is hardly annotated in the GO. There is f.e. no annotation for a "Molecular Process". However, all the literature research results might be biased due to the way how we chose a subset of the best ranked putative targets and the literature research is far from being complete.

Iorio et al [5] identified a number of genes which play a role in cancer among the predicted miRNA targets of the five most consistently deregulated miRNAs in human breast cancer. Out of those genes 19 are included in the PPIN. Out of those 15 have greater degree values than the mean value of the whole human PPIN. Interestingly out of these 15 genes, 13 have smaller clustering coefficient values than the mean values as it was suggested by Liang et [7]. Looking at the subontology "Molecular Function" 16 out of the 19 genes identified by Iorio et al are in the list of genes which are members of overrepresented categories for the putative, validated targets and breast cancer genes. The remaining three are supported by cancer genes. MiR 21 is involved in various kinds of cancer and it has been shown that it plays a role in signaling pathways [3] which supports the result that the used putative targets and cancer genes commonly have a function in signaling pathways. Esquela-Kerscher et al [18] reported that miR21, miR155, miR125 and miR145 are involved in different kinds of cancer and not only

breast cancer. This may make our findings also usable for other kinds of cancer and in this way help to uncover the role of miRNAs in cancer.

5.4 Future Work

Our report shows that network properties of miRNA regulated proteins differ from those of the whole human PPIN. We propose a simple way to use these differences for ranking. However, a more in depth analysis of how the values of network properties differ between the putative targets, validated targets and the whole PPIN could be useful. For example it should be studied if it is better to use other statistical measures, such as the median, instead of the mean. Furthermore using the distance to the mean values of a network property instead of a "greater than" or "smaller than" criterion might contribute more to ranking.

State of the art target prediction tools use evolutionary conservation of target sites to reduce the number of false positives and to score their putative targets [25]. It has been shown that there is a correlation between the evolutionary rate and the degree value of a protein [74]. This might propose one way of using network properties within the scoring system of miRNA target prediction tools. Additionally, other network properties might be included to find a new scoring system and thus improve the tools.

For the GO category analysis it could be studied how other ways of removing over-represented categories because of hierarchy contribute to ranking. This may make an analysis of common categories in the subontology "Biological Process" easier.

There may also other ways to study, which might help to rank the relevance of putative target genes. Genes involved in cancer should normally be involved in the cell cycle, apoptosis, proliferation and other typical processes which are affected by cancer [75]. So one idea would be to map the putative target gene onto pathways in which they are involved and rank them according to their occurrence in cancer related pathways. Since we know that miRNAs are often located in fragile chromosom regions [3] another way would be to study if a chromosomal change which affects a miRNA also affects its putative target genes.

This report resulted in a list of putative target genes which differ in their network properties from the whole human PPIN in the same way validated targets do, as well

as in a list of genes which are members of overrepresented GO categories which are also overrepresented for validated targets and /or breast cancer genes. For the best ranked genes in both lists it could be studied if they play a role in cancer. More literature research might help to uncover the role of the best ranked targets in breast cancer and furthermore it might help to evaluate the results of different ranking approaches shown here. Furthermore these putative targets might be validated using wet-lab experiments and in this way both the prediction tools and the reported approaches in this study could be validated and improved. Moreover it could be studied if similar approaches also work for other cancers or other diseases.

References

[1] Sethupathy P, Corda B, Hatzigeorgiou AG. TarBase: A comprehensive database of experimentally supported animal microRNA targets. RNA. 2006 Feb;12(2):192–197.

[2] Espinosa CES, Slack FJ. The role of microRNAs in cancer. Yale J Biol Med. 2006 Dec;79(3-4):131–140.

[3] Calin GA, Croce CM. Chromosomal rearrangements and microRNAs: a new cancer link with clinical implications. J Clin Invest. 2007 Aug;117(8):2059–2066.

[4] Wang V, Wu W. MicroRNA: A New Player in Breast Cancer Development. Journal of Cancer Molecules. 2007;3(5):133–138.

[5] Iorio MV, Ferracin M, Liu CG, Veronese A, Spizzo R, Sabbioni S, et al. MicroRNA gene expression deregulation in human breast cancer. Cancer Res. 2005 Aug;65(16):7065–7070.

[6] Hsu CW. Characterization of microRNA-Regulated Protein-Protein Interaction Network. In: The eighth international conference on Systems Biology. Institute of Biomedical Informatics, National Yang-Ming University, Taipei, Taiwan; 2007. p. 1.

[7] Liang H, Li WH. MicroRNA regulation of human protein protein interaction network. RNA. 2007 Sep;13(9):1402–1408.

[8] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25–29.

[9] Hutvgner G, Zamore PD. A microRNA in a multiple-turnover RNAi enzyme complex. Science. 2002 Sep;297(5589):2056–2060.

[10] Sethupathy P, Megraw M, Hatzigeorgiou AG. A guide through present computational approaches for the identification of mammalian microRNA targets. Nat Methods. 2006 Nov;3(11):881–886.

[11] Jonsson PF, Bates PA. Global topological features of cancer proteins in the human interactome. Bioinformatics. 2006 Sep;22(18):2291–2297.

[12] Baasiri RA, Glasser SR, Steffen DL, Wheeler DA. The breast cancer gene database: a collaborative information resource. Oncogene. 1999 Dec;18(56):7958–7965.

[13] Stark A, Brennecke J, Bushati N, Russell RB, Cohen SM. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. Cell. 2005 Dec;123(6):1133–1146.

[14] Alberts B, Roberts K, Lewis J, Hopkin K, Johnson A, Walter P, et al. Essential Cell Biology: An introduction to the Molecular Biology of the Cell, 2nd edition. Garland Pub; 2003.

[15] Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell. 1993 Dec;75(5):843–854.

[16] Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. miRBase: tools for microRNA genomics. Nucleic Acids Res. 2008 Jan;36(Database issue):D154–D158.

[17] Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005 Jan;120(1):15–20.

[18] Esquela-Kerscher A, Slack FJ. Oncomirs - microRNAs with a role in cancer. Nat Rev Cancer. 2006 Apr;6(4):259–269.

[19] Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O, et al. Identification of hundreds of conserved and nonconserved human microRNAs. Nat Genet. 2005 Jul;37(7):766–770.

[20] Krek A, Grn D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, et al. Combinatorial microRNA target predictions. Nat Genet. 2005 May;37(5):495–500.

[21] Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, et al. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature. 2005 Feb;433(7027):769–773.

[22] World Health Organization International Agency for Research on Cancer. World Cancer Report; 2003.

[23] Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, et al. MicroRNA expression profiles classify human cancers. Nature. 2005 Jun;435(7043):834–838.

[24] Kiriakidou M, Nelson PT, Kouranov A, Fitziev P, Bouyioukos C, Mourelatos Z, et al. A combined computational-experimental approach predicts human microRNA targets. Genes Dev. 2004 May;18(10):1165–1178.

[25] Mazire P, Enright AJ. Prediction of microRNA targets. Drug Discov Today. 2007 Jun;12(11-12):452–458.

[26] John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. Human MicroRNA targets. PLoS Biol. 2004 Nov;2(11):e363.

[27] Lewis BP, hung Shih I, Jones-Rhoades MW, Bartel DP, Burge CB. Prediction of mammalian microRNA targets. Cell. 2003 Dec;115(7):787–798.

[28] Krger J, Rehmsmeier M. RNAhybrid: microRNA target prediction easy, fast and flexible. Nucleic Acids Res. 2006 Jul;34(Web Server issue):W451–W454.

[29] Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, et al. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature. 2005 Mar;434(7031):338–345.

[30] Giraldez AJ, Mishima Y, Rihel J, Grocock RJ, Dongen SV, Inoue K, et al. Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. Science. 2006 Apr;312(5770):75–79.

[31] Kim SK, Nam JW, Rhee JK, Lee WJ, Zhang BT. miTarget: microRNA target gene prediction using a support vector machine. BMC Bioinformatics. 2006;7:411.

[32] Yook SH, Oltvai ZN, Barabsi AL. Functional and topological characterization of protein interaction networks. Proteomics. 2004 Apr;4(4):928–942.

[33] Vzquez A, Dobrin R, Sergi D, Eckmann JP, Oltvai ZN, Barabsi AL. The topological relationship between the large-scale attributes and local interaction patterns of complex networks. Proc Natl Acad Sci U S A. 2004 Dec;101(52):17940–17945.

[34] Newman MEJ. Mathematics of Networks. In: Durlauf SN, Blume LE, editors. The New Palgrave Dictionary of Economics, 2nd Edition. Palgrave Macmillan; 2008. p. 1.

[35] Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature. 1998 Jun;393(6684):440–442.

[36] Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics. 2005 Aug;21(16):3448–3449.

[37] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003 Nov;13(11):2498–2504.

[38] Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, et al. Human protein reference database–2006 update. Nucleic Acids Res. 2006 Jan;34:D411–D414.

[39] Csrdi G, Nepusz T. The igraph software package for complex network research. Complex Systems. 2006;1695.

[40] R Development Core Team. R: A language and Environment for Statistical Computing; 2006.

[41] Mann, Whitney. On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics. 1947;18:50–60.

[42] Han JDJ, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature. 2004 Jul;430(6995):88–93.

[43] Oliveros JC. VENNY. An interactive tool for comparing lists with Venn Diagrams. http://bioinfogp.cnb.csic.es/tools/venny/index.html; 2007.

[44] Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: integrating information about genes, proteins and diseases. Trends Genet. 1997 Apr;13(4):163.

[45] Xie Y, Hung MC. p66Shc isoform down-regulated and not required for HER-2/neu signaling pathway in human breast cancer cell lines with HER-2/neu overexpression. Biochem Biophys Res Commun. 1996 Apr;221(1):140–145.

[46] Jackson JG, Yoneda T, Clark GM, Yee D. Elevated levels of p66 Shc are found in breast cancer cell lines and primary tumors with high metastatic potential. Clin Cancer Res. 2000 Mar;6(3):1135–1139.

[47] Frackelton AR, Lu L, Davol PA, Bagdasaryan R, Hafer LJ, Sgroi DC. p66 Shc and tyrosine-phosphorylated Shc in primary breast tumors identify patients likely to relapse despite tamoxifen therapy. Breast Cancer Res. 2006;8(6):R73.

[48] Lamorte L, Royal I, Naujokas M, Park M. Crk adapter proteins promote an epithelial-mesenchymal-like transition and are required for HGF-mediated cell spreading and breakdown of epithelial adherens junctions. Mol Biol Cell. 2002 May;13(5):1449–1461.

[49] Koda M, Sulkowska M, Kanczuga-Koda L, Sulkowski S. Expression of insulin receptor substrate 1 in primary breast cancer and lymph node metastases. J Clin Pathol. 2005 Jun;58(6):645–649.

[50] Koda M, Sulkowska M, Kanczuga-Koda L, Golaszewska J, Kisielewski W, Baltaziak M, et al. Expression of the Insulin Receptor Substrate 1 in primary tumors and lymph node metastases in breast cancer: correlations with Bcl-xL and Bax proteins. Neoplasma. 2005;52(5):361–363.

[51] Kagan BL, Henke RT, Cabal-Manzano R, Stoica GE, Nguyen Q, Wellstein A, et al. Complex regulation of the fibroblast growth factor-binding protein in MDA- MB-468 breast cancer cells by CCAAT/enhancer-binding protein beta. Cancer Res. 2003 Apr;63(7):1696–1705.

[52] Park Y, Shon SK, Kim A, Kim KI, Yang Y, Cho DH, et al. SOCS1 induced by NDRG2 expression negatively regulates STAT3 activation in breast cancer cells. Biochem Biophys Res Commun. 2007 Nov;363(2):361–367.

[53] Haffner MC, Petridou B, Peyrat JP, Rvillion F, Mller-Holzner E, Daxenbichler G, et al. Favorable prognostic value of SOCS2 and IGF-I in breast cancer. BMC Cancer. 2007;7:136.

[54] Zhang WJ, Li BH, Yang XZ, Li PD, Yuan Q, Liu XH, et al. IL-4-induced Stat6 activities affect apoptosis and gene expression in breast cancer cells. Cytokine. 2008 Apr;42(1):39–47.

[55] White KA, Yore MM, Deng D, Spinella MJ. Limiting effects of RIP140 in estrogen signaling: potential mediation of anti-estrogenic effects of retinoic acid. J Biol Chem. 2005 Mar;280(9):7829–7835.

[56] Rey JM, Pujol P, Callier P, Cavailles V, Freiss G, Maudelonde T, et al. Semi-quantitative reverse transcription-polymerase chain reaction to evaluate the expression patterns of genes involved in the oestrogen pathway. J Mol Endocrinol. 2000 Jun;24(3):433–440.

[57] Kerley JS, Olsen SL, Freemantle SJ, Spinella MJ. Transcriptional activation of the nuclear receptor corepressor RIP140 by retinoic acid: a potential negative-feedback regulatory mechanism. Biochem Biophys Res Commun. 2001 Jul;285(4):969–975.

[58] Dittmer J, Vetter M, Blumenthal SG, Lindemann RK, Klbl H. Importance of ets1 proto-oncogene for breast cancer progression. Zentralbl Gynakol. 2004 Aug;126(4):269–271.

[59] Buggy Y, Maguire TM, McGreal G, McDermott E, Hill ADK, O'Higgins N, et al. Overexpression of the Ets-1 transcription factor in human breast cancer. Br J Cancer. 2004 Oct;91(7):1308–1315.

[60] Tang W, Li Y, Yu D, Thomas-Tikhonenko A, Spiegelman VS, Fuchs SY. Targeting beta-transducin repeat-containing protein E3 ubiquitin ligase augments the effects of antitumor drugs on breast cancer cells. Cancer Res. 2005 Mar;65(5):1904–1908.

[61] Li Y, Clevenger CV, Minkovsky N, Kumar KGS, Raghunath PN, Tomaszewski JE, et al. Stabilization of prolactin receptor in breast cancer cells. Oncogene. 2006 Mar;25(13):1896–1902.

[62] Maeda T, Alexander CM, Friedl A. Induction of syndecan-1 expression in stromal fibroblasts promotes proliferation of human breast cancer cells. Cancer Res. 2004 Jan;64(2):612–621.

[63] Sun JM, Chen HY, Moniwa M, Litchfield DW, Seto E, Davie JR. The transcriptional repressor Sp3 is associated with CK2-phosphorylated histone deacetylase 2. J Biol Chem. 2002 Sep;277(39):35783–35786.

[64] Williams AO, Isaacs RJ, Stowell KM. Down-regulation of human topoisomerase IIalpha expression correlates with relative amounts of specificity factors Sp1 and Sp3 bound at proximal and distal promoter regions. BMC Mol Biol. 2007;8:36.

[65] Mertens-Talcott SU, Chintharlapalli S, Li X, Safe S. The oncogenic microRNA-27a targets genes that regulate specificity protein transcription factors and the G2-M checkpoint in MDA-MB-231 breast cancer cells. Cancer Res. 2007 Nov;67(22):11001–11011.

[66] Wang XB, Peng WQ, Yi ZJ, Zhu SL, Gan QH. [Expression and prognostic value of transcriptional factor sp1 in breast cancer]. Ai Zheng. 2007 Sep;26(9):996–1000.

[67] Zannetti A, Vecchio SD, Romanelli A, Scala S, Saviano M, Cali' G, et al. Inhibition of Sp1 activity by a decoy PNA-DNA chimera prevents urokinase receptor expression and migration of breast cancer cells. Biochem Pharmacol. 2005 Nov;70(9):1277–1287.

[68] Wang L, Li D, Fu Z, Li H, Jiang W, Li D. Association of CTLA-4 gene polymorphisms with sporadic breast cancer in Chinese Han population. BMC Cancer. 2007;7:173.

[69] Erfani N, Razmkhah M, Talei AR, Pezeshki AM, Doroudchi M, Monabati A, et al. Cytotoxic T lymphocyte antigen-4 promoter variants in breast cancer. Cancer Genet Cytogenet. 2006 Mar;165(2):114–120.

[70] Ghaderi A, Yeganeh F, Kalantari T, Talei AR, Pezeshki AM, Doroudchi M, et al. Cytotoxic T lymphocyte antigen-4 gene in breast cancer. Breast Cancer Res Treat. 2004 Jul;86(1):1–7.

[71] Jeong H, Mason SP, Barabsi AL, Oltvai ZN. Lethality and centrality in protein networks. Nature. 2001 May;411(6833):41–42.

[72] Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics. 2007 Jan;23(2):257–258.

[73] Josefson D. High insulin levels linked to deaths from breast cancer. BMJ. 2000 Jun;320(7248):1496.

[74] Saeed R, Deane CM. Protein protein interactions, evolutionary rate, abundance and age. BMC Bioinformatics. 2006;7:128.

[75] Chiromatzo AO, Oliveira TYK, Pereira G, Costa AY, Montesco CAE, Gras DE, et al. miRNApath: a database of miRNAs, target genes and metabolic pathways. Genet Mol Res. 2007;6(4):859–865.