12-2016

A New System for Human MicroRNA functional Evaluation and Network

Jiachun Han
jayceecsapp@gmail.com

A NEW COMPUTATIONAL SYSTEM FOR HUMAN MICRORNA FUNCTIONAL EVALUATION AND NETWORK CONSTRUCTION

by

Jiachun Han

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfilment of Requirements

For the Degree of Master of Science

Major: Computer Science

Under the Supervision of Professor Juan Cui

Lincoln, Nebraska

December, 2016

A NEW COMPUTATIONAL SYSTEM FOR HUMAN MICRORNA FUNCTIONAL EVALUATION AND NETWORK CONSTRUCTION

Jiachun Han, M.S.

University of Nebraska, 2016

Adviser: Juan Cui

MicroRNAs are functionally important endogenous non-coding RNAs that silence host genes in animal and plant via destabilizing the mRNAs or preventing the translation. Given the far-reaching implication of microRNA regulation in human health, novel bioinformatics tools are desired to facilitate the mechanistic understanding of microRNA mediated gene regulation, their roles in biological processes, and the functional relevance among microRNAs. However, most state-of-the-art computational methods still focus on the functional study of microRNA targets and there is no effective strategy to infer the functional similarity among microRNAs. In this study, we developed a new method to quantitatively measure the functional similarity among microRNAs based on the integrated functional annotation data from Gene Ontology, human pathways, and PFam databases. Through analyzing human microRNAs, we further demonstrated the use of the derived microRNA pairwise similarities to discover the cooperative microRNA modules and to construct the genome-scale microRNA-mediated gene network in human. The complete results and the similarity assessment system can be freely accessed at (http://sbbi.unl.edu/microRNASim)

ACKNOWLEDGMENTS

I would like to express my appreciation and sincere thanks to my advisor Dr. Juan Cui for her help and support. I have been inspired by her work ethic, patience and kindness. This research experience with her has taught me more than academic knowledge, but also how to be a responsible researcher in a lab environment.

I would also like to thank the committee memebers, Dr. Jitender Deogun and Peter Revesz for taking their precious time to be on my master committee.

I also want to thank all of my colleagues, Jiang Shu, Tian Gao, Bruno Vieira, Milad Rad, in Systems Biology and Biomedical Informatics Laboratory (SBBI) for their support and help in this study.

At last, I want to thank Nebraska Center for the Prevention of Obesity Diseases (NPOD). This work was funded through their COBRE grant of NIH (P20GM104320).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

MicroRNAs(miRNAs) are a class of small non-coding RNAs that can regulate gene expression post-transcriptionally by binding to the 3'- UTR of target genes and triggering the mRNA degradation or translation inhibition [4]. miRNAs are one of the most significant components in the cell. They participate in a vast of array of fundamental cellular processes and disease development [17], [45], [53], [56], such as cancer and obesity.

1.1 Motivation

Over the last decade, functional study of miRNAs has been largely dependent on the analysis of the target-associated pathways. Most early efforts were focused on the target prediction [34], [46] that was followed by pathway enrichment analysis [44]. In order to improve the prediction performance through statistical modeling and more sophisticated machine learning strategies, features of sequence and structure that can characterize miRNA-mRNA interactions have been thoroughly studied [13], [47], [48]. Meanwhile, as opposed to assessment on individual miRNA, there are attempts to evaluate miRNA similarity based on Gene Ontology (GO) semantic similarities or through target-involved pathways [31]. However, those studies

have been hampered by either showering small coverage of miRNAs or lacking comprehensive annotations on function.

1.2 Contribution

Our work has two main contributions as described below:

- We proposed a new computational framework for the assessment of functional similarities among miRNAs using several different functional annotation systems. Evaluating functional similarities among miRNAs from three different perspectives gave us a comprehensive understandings and more confidence in the outputs.
- Based on the miRNA similarity scores generated by our system, we demonstrated the identifications of miRNA functional modules and visualize them through involved pathways, which leads to the downstream application of our modules.

1.3 Outline

This study is organized as follows. Chapter 2 provides the background of miRNA, miRNA-mRNA interactions and miRNA functional similarity. Chapter 3 presents our three evaluation similarity systems in details. After the outputs of similarity scores are generated through our measurement system, we use Chapter 4 to explain the mechanism to integrate the three systems. This leads us to identify the clusters among miRNAs and the construction of miRNA mediated gene network in human, which presents in Chapter 5. At last, Chapter 6 illustrates our conclusions and future works.

Chapter 2

Background

2.1 MicroRNA

MicroRNAs (miRNAs), approximately 22 nucleotides in length, are non-coding RNAs molecules that play important roles in post-transcriptional regulation. The first pair of miRNA found in human cell was lin-14 and let-7 which was identified in *C. elegans* [5]. Experimental evidence shows that a single miRNA species can reduce the stability of hundreds of unique messenger RNAs and may repress the production of hundreds of proteins. In humans, 2,588 known endogenous miRNAs (according to miRBase v21 [30]) regulate over 60% of human genes and participate in a vast of array of fundamental cellular processes and disease development [17], [53], [56], [45]. Recent studies have reported regulated miRNAs in diverse cancer types, such as breast cancer [24], lung cancer [52], prostate cancer [38], colon cancer, ovarian cancer [54] and head and neck cancer [57]. miRNAs are also implicated in a number of neurological disorders including Alzheimers disease , multiple sclerosis [2] and schizophrenia [7]. miRNAs regulate diverse aspects of development and physiology, thus understanding its biological role is proving more and more important.

2.2 MiRNA Function Inference Based On MiRNA Target Prediction

Over the last decade, functional study of miRNAs has been largely dependent on the analysis of the miRNA-mRNA interaction. Most early efforts were focused on target prediction that are followed by pathway enrichment analysis, which is a method to identify function that are enriched among genes that are over-represented in a large set of genes or proteins, and may have an association with disease phenotypes [42]. The basis of this method is the assumption that function of miRNAs can be reflected by the function of their target genes. Thus miRNA functional annotation heavily relies on the miRNA-target prediction. Due to the lack of high-throughput biological methods to identify miRNA-mRNA bindings, many computational approaches have been developed to identify miRNA target genes. These include, TargetScan [6], miRDB [48], miRanda [27] and PITA [28]. The miRNA regulation exhibits a dramatic complexity, since a transcript can have many target sites for one miRNA and a transcript can also have target sites for several miRNAs. The many-to-many relations between miRNAs and mRNAs lead to the very complex miRNA regulatory mechanisms and pose great challenges on the computational approaches for miRNA-target prediction [32]. Some of the most popular techniques nowadays to predict miRNA-target prediction are base pairing and evolutionary conservation of target site [19], [51]. Base pairing is an algorithm based on sequence complementarity. This kind of algorithm tends to have low accuracy and high false positive result [6]. Conservation analysis was introduced to reduce the false positive prediction. Conservation refers to the maintenance of a sequence across species. Conservation Analysis means to compare sequence analysis to check conservation among the sequences. Although great efforts have been made, those miRNA-target prediction methods still suffer from large numbers of false positive prediction in general. Besides computational methods, experimental methods were also developed for miRNA-target prediction. Although

having high accuracy, these types of methods tend to provide limited information of miRNA-prediction due to time consumption and complexity of experiments.

2.3 MiRNA Functional Similarity

MiRNA function inference based on target prediction provides us with the information of individual miRNA functions. However, knowledge of pairwise miRNA functional similarity can give us a deeper insight into the miRNA functions in many applications. In addition, most miRNAs have multiple gene targets while many genes can be regulated by multiple miRNAs [31], [6], [39], [39]. Competition among different miRNAs takes place when they can potentially target the same genes at the same or adjacent binding sites, while collaboration of miRNAs exists when they bind to the different, non-overlapping regions of the same target genes or different genes involved in the same functional process [31], [6], [39], [12]. Given these facts, a cooperative module may form when more than one miRNA regulates the same or related pathways through regulating the same or different genes [44]. Most of miRNA similarity methods can be categorized into three groups. One is through miRNA-disease association to infer miRNA similarities, another is through measuring miRNA sequence and expression similarities. The third one is to indirectly infer miRNA similarity through their target genes. Each method has their own advantages and defects. Here, we focus on the third method. We infer miRNA functional similarity through the functional similarity of their protein-coding target genes. The functional similarity of protein-coding genes can be obtained through Gene Ontology (GO) database [3], [11], which will be discussed in detail in the next chapter.

Chapter 3

MiRNA Functional Similarity Calculation

In this chapter, we are going to propose three approaches to compute miRNA functional similarities based on GO and two other functional annotation systems, PFam (Protein family database) and biological pathways.

The miRNA-mRNA interaction data were downloaded from different resources, including 1) experimentally validated entries from miRTarBase (582 miRNAs included) [23] and 2) predicted interactions from both TargetScan (686 miRNAs included) [6] and miRDB (2,588 miRNAs included) [46].

The whole GO dataset was downloaded from the Gene Ontology Consortium [11], which is comprised of three domains (cellular component, molecular function and biological process). Obsolete terms were excluded, which results in a total of 42,144 terms. In addition, we compiled the functional annotations based on the PFam-A set from Uniprot [43] and a collection of 1447 pathways (1330 from GSEA database [42] and 117 from NPO Bioinformatics Japan database[9].

3.1 MiRNA Similarity Evaluation Based on GO

Gene ontology(GO) is a controlled vocabulary used to describe the biology of a gene product in any organism [11]. It has developed three separate ontologies: molecular function, biological process and cellular component to describe the attributes of gene products. Molecular function defines what a gene product does at the biochemical level without specifying where or when the event actually occurs or its broader context; biological process describes the contribute of a gene product to a biological objective; and cellular component refers to where in the cell a gene product functions. Each of the three GO domains is structured as a directed acyclic graph (DAG), with GO terms represented as nodes to describe the gene product attributes and three type of semantic relations, 'is-a', 'part-of', and 'regulate' as edges to annotate the relationships between GO terms. 'is-a' represents a simple class-subclass relationship, 'part-of' indicates a component relationship, and 'regulate' implies the relationship of direct control. Each node represents a gene term in the process of gene product description. Each ontology has the following properties:

- The bottom most level of the graph is the term itself, and at the upper levels are its ancestors GO terms, at the topmost level is the root of the GO tree.
- Each term is a child of one or multiple parents, and child terms are instances, components of, or regulate parent terms.
- The parent would be a broader GO term, and the child would be a more specific term with regard to describing a gene product.

Figure 3.1 shows an example of GO structures for an gene product YAB4. Each color represents a different ontology tree used to describe the gene product. Blue tree refers to biological process ontology; green tree refers to molecular function ontology; purple refers to cellular component ontology. Each node represents as a gene term. Different type of edges

represents as different relations among the terms. Solid edge represents 'is-a relation', coarse dash line represents 'part-of' and fine dash line presents 'regulate' relation.

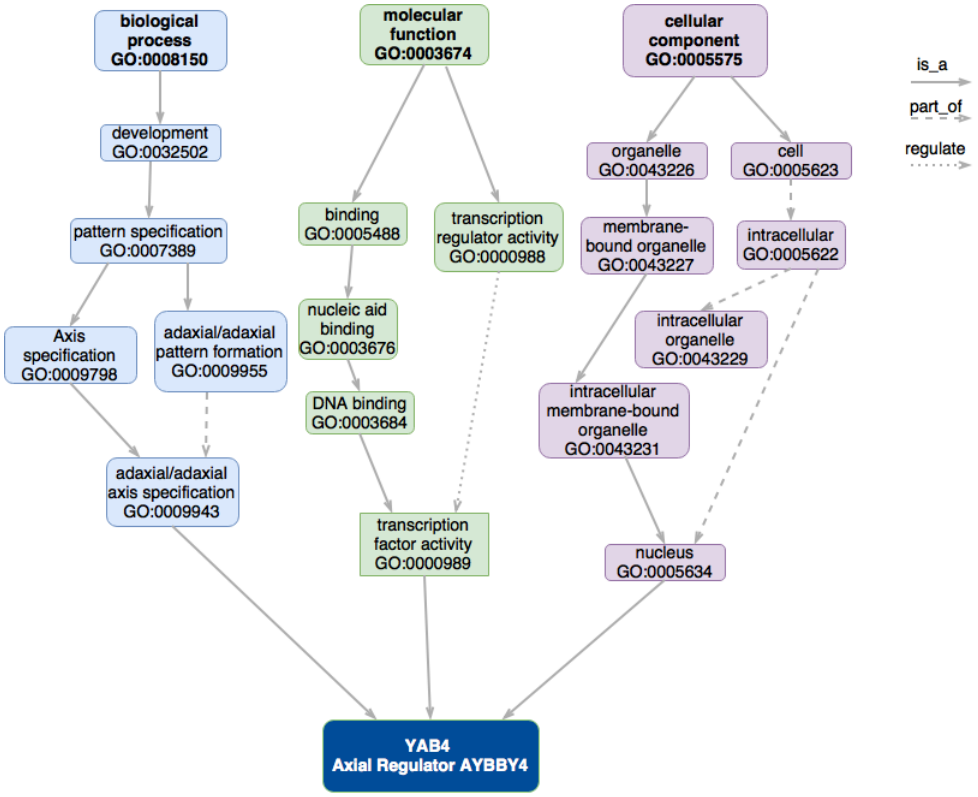


Figure 3.1: Three gene ontologies involved to describe gene product YAB4

As we mentioned earlier, each gene term is used to describe an attribute of a gene product. When all the genes are assigned to their description gene terms, we get a full ontology tree, through which we can calculate gene pair similarities. The calculation of pairwise gene similarities can be categorized into two groups, edge-based computational methods and node-based computational methods. Edge-based algorithms mainly depend on counting the number of edges along the paths linking the interested GO terms. Node-based approaches reply on comparing the properties of the GO terms involved, which can be related to the term nodes themselves, their ancestors and their descendants. The most commonly used

concept here is Information Content(IC), which can measure how specific and informative a term is [33].This concept will be discussed in the next section. Gene pair similarities provide the foundation for the evaluation of miRNA pair similarities through GO.

Figure 3.2 A) illustrates the structure of gene ontology with gene annotations. and B) miRNA relationships based on their target genes.

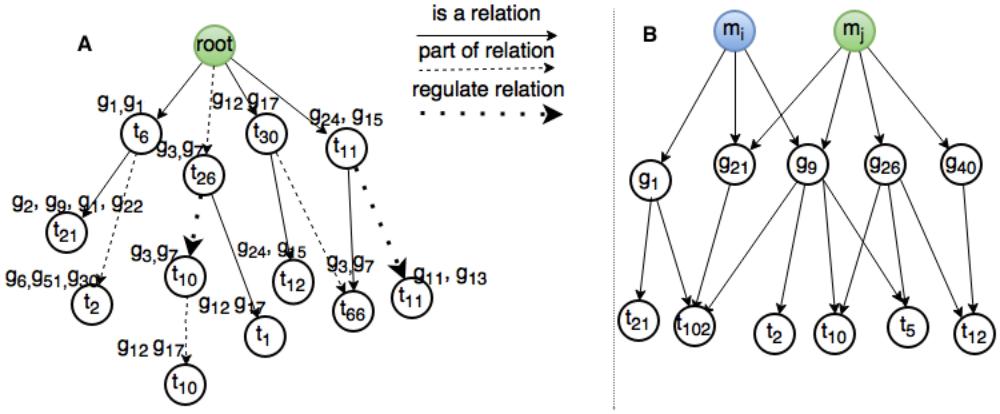


Figure 3.2: Illustrition of A) An Gene Ontology Structure. t_k refers to the gene terms and g_k refer to the genes that t_k describes and B) miRNA-gene iteration structure. m_i and m_j represent two miRNAs, g_k are the genes and t_k are the gene terms.

As discussed above, we took three steps to evaluate miRNA functional similarities using GO structure as shown in Figure 3.3. Firstly, we calculated the similarities of pairwise GO terms, which describes the attributes of miRNA target genes. Then we computed the similarities between gene pairs. At last, through the similarities of target gene pairs, we generated the pairwise miRNA similarities.

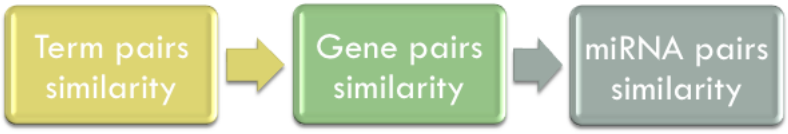


Figure 3.3: Workflow of miRNA similarity compuation

We implemented two methods to compute pairwise miRNA similarity as explained in [49], [50]. One is the edge-based similarity evaluation mechanism and the other one is hybrid that combines edge-based concept with the node-based Information Content (IC). Both methods consist of three components,α,β, and γ and α_{ic} ,β_{ic} , and γ_{ic} respectively, which are introduced in detail in the next section.

3.1.1 GO Term Similarity Calculation

3.1.1.1 Edge-Based GO Term Similarity

The first approach is comprised of three components α, β, and γ as shown in Figure 3.4.

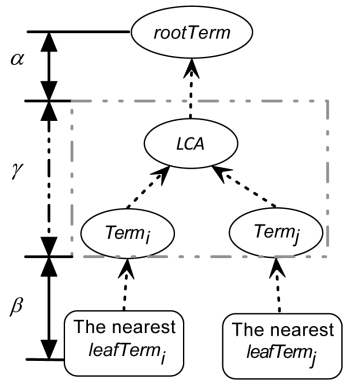


Figure 3.4: Edge-based structure of miRNA similarity algorithm. LCA refers to the lowest common ancestor. $Term_i$, $Term_j$ are the two terms concerned. α, β and γ represents one components of the edge-based algorithm

α represents the the depth of the Lowest Common Ancestors (LCA) of the two terms. LCA is the common ancestor which is the deepest in the tree. It is calculated as follows:

$$\alpha = \max_{\substack{p_m \in Paths(t_i) \\ p_n \in Paths(t_j)}} \{|P_m \cap P_n|\} - 1$$

Where Paths(t) is the collection of paths from the concerned term t to its root nodes. The more specific a term is, the more in detail the term can express a gene product. LCA returns

the maximum number of common terms to the root. The β value measures the relative generality of the two terms. It is computed as the summation of the two minimum distances between the concerned terms with all their descending leaf terms respectively; It is defined as:

$$\beta = \max\{\min_{u \in U}\{dist(t_i, u)\}, \min_{v \in V}\{dist(t_j, v)\}\}$$

Where U,V are the collections of leaf terms descending from terms t_i and t_j, respectively; $dist(s,t)$ is defined as the number of edges along the shortest path between terms s and t. The component γ denotes the local distance between the two terms with relative to their LCA. It is defines as:

$$\gamma = dist(LCA, t_i) + dist(LCA, t_j)$$

Finally the term similarity between any two terms based on the first method is defined as:

$$s_{term\_edge}(t_i, t_j) = \frac{\kappa}{\kappa + \gamma} * \frac{\alpha}{\alpha + \beta} \quad (1)$$

Where κ is the number of edges along the longest path in GO. We used 19 as κ in our study.

3.1.1.2 Hybrid-Based GO Term Similarity

In the second method, Information Content (IC) concept is added to represent how specific and informative a term describe a gene product. The IC of a term t is defined as $IC(t) = -\log p(t)$ where $p(t)$ is the percentage of cumulative genes of each term among the total number of human genes. Cumulative genes are the genes assigned to the interest term and all the unique genes belong to its descendants. The distance of any two terms in the second method is defined as the difference between their IC values. This method also consists of three components, α_{ic} , β_{ic} ,and γ_{ic}. Figure 3.5 shows the structure of the hybrid-based term similarity algorithm:

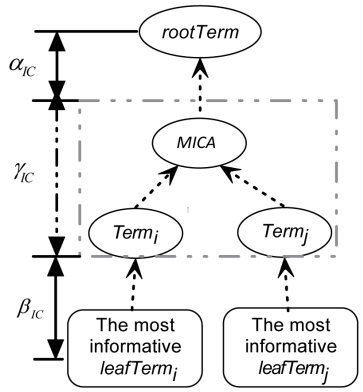


Figure 3.5: The hybrid-based algorithm structure. $MICA$ is the most informative ancestor, $Term_i$, $Term_j$ are the two concerned terms.

α_{ic} denotes the IC-based specificity of the Most Informative Common Ancestor ($MICA$) of any two terms,$t_i$and t_j. It is defined as

$$\alpha_{ic} = dist(root, MICA) = -\log p(MICA)$$

Where

$$dist(t_i, t_j) = IC(t_i) - IC(t_j) = logp(t_j) - logp(t_i)$$

The β_{ic} represents the IC-based generality of the two concerned terms. It is computed as the average of the two IC-based distances between the two concerned terms and their descending leaf terms. It is defined as:

$$\beta_{ic} = \frac{dist_{ic}(t_i, MIL_i) + dist_{ic}(t_j, MIL_j)}{2}$$

Where MIL is the Most Informative Leaf, the leaf with the highest IC. γ_{ic} is calculated by adding the distance between and the concerned terms.

$$\gamma_{ic} = dist(MICA, t_i) + dist(MICA, t_j)$$

At last, the second similarity method is defined as:

$$s_{term\_hybrid}(t_i, t_j) = \frac{1}{1+\gamma_{ic}} * \frac{\alpha_{ic}}{\alpha_{ic}+\beta_{ic}} \quad (2)$$

3.1.2 Gene Similarity Evaluation

Measurement of term-pair similarities provides us with the necessary information for the calculation of gene-pair similarities. For each pair of genes, g_i and g_j, two term sets $T(g_i)$ and $T(g_j)$ were used to represent the collection of the terms for the corresponding genes respectively. We built a gene-term relation table (as shown in Table 3.2), where each term in $T(g_i)$ represented a row and each term in $T(g_j)$ represented a column. Then the table was filled with the pairwise term similarities calculated by formula (1) or (2). The formula for gene similarity is defined as:

$$s(g_i, g_j) = \frac{\sum_{t_k \in T(g_i)} \max_{t_n \in T(g_j)}\{s_{term}(t_k, t_n)\} + \sum_{t_n \in T(g_j)} \max_{t_k \in T(g_i)}\{s_{term}(t_k, t_n)\}}{|T(g_i)| + |T(g_j)|} \quad (3)$$

Now let us use a gene pair,'HIST2H2BC' and 'H2BFM', as an example. Table 3.1 demonstrates the information of the terms belong to the two genes.

Table 3.1: Information on GO terms associated to genes HIST2H2BC and H2BFM

HIST2H2BC	
GO:0006334	nucleosome assembly
GO:0046982	protein heterodimerization activity
GO:0000788	nuclear nucleosome
GO:0003677	DNA binding
H2BFM	
GO:0006334	nucleosome assembly
GO:0046982	protein heterodimerization activity
GO:0000786	nucleosome
GO:0000788	nuclear nucleosome
GO:0003677	DNA binding

Looking into the detail of gene 'HIST2H2BC' and gene 'H2BFM' shown in Table 3.1, we can observe that these two genes have many terms in common. The only different between these two genes is GO term, GO:0000786 'nucleosome', that only belongs to gene 'H2BFM', and is one of the immediate direct parent of 'nuclear nucleosome' which describes both genes 'HIST2H2BC' and 'H2BFM'. Based on these facts, we can assume that genes 'H2BFM' and 'HIST2H2BC' be very similar. Table 3.2 gene-term relation table shows the term similarities associated with the two genes.

Table 3.2: Similarities between the terms that annotate genes H2BFM and HIST2H2BC respectively

HIST2H2BC / **H2BFM**	**GO:0006334**	**GO:0046982**	**GO:0000788**	**GO:0003677**
GO:0006334	0.89	0	0	
GO:0046982	0	1	0	0.38
GO:0000786	0	0	0.86	0
GO:0000788	0	0	1	0
GO:0003677	0	0.38	0	0.8

Applying the term similarities in the Table 3.2 to Equation (3), we get $s(HIST2H2BC, H2BFM) \approx 0.92$, which proves our assumption above that these two genes are very similar.

Next, to further analyze the performance of these two methods, we extracted two sets of similarity measures (edge-based gene similarity and hybrid-based gene similarity) on the pairs of genes belong to the same Protein Families (Intra-PFam genes) and the pairs of genes from the different Protein families (Inter-PFam genes). Intuitively, the distance among Intra-PFam target genes should be closer compared to that among the Inter-PFam target genes, which is observed in Fig. 3.6 with both methods. Specifically, the average similarity in the intra-group (blue) is much higher than the inter-group (pink) similarity, indicating a higher functional relevance within the intra-group.

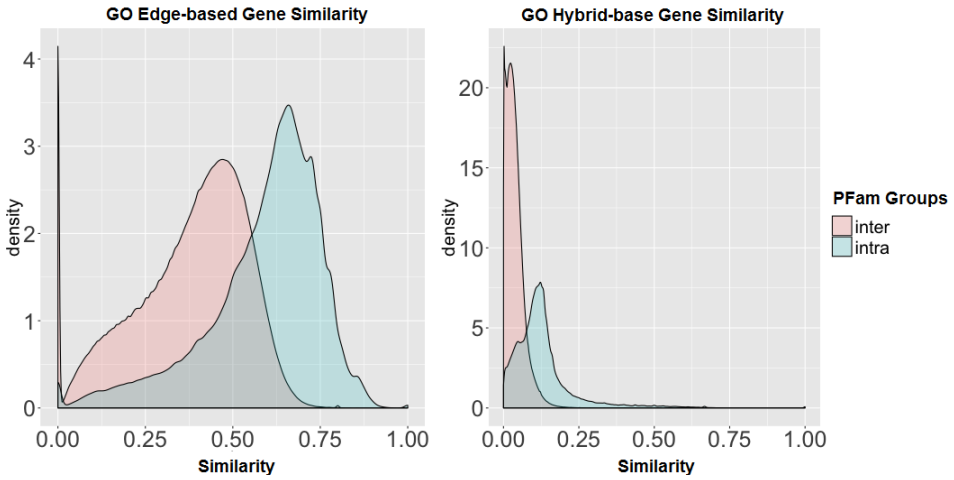


Figure 3.6: The distance distribution between inter-PFam gene similarity(pink) and intra-PFam gene similarity(blue)

In addition, we compared the distance distributions between Intra-PFam and Inter-PFam groups, where a Wilcoxon test is performed to evaluate the statistical significance. Through the Wilcoxon-test on these two sets of pairwise similarities, we observed that both scoring systems demonstrate similar discerning power on these two groups. Table 3.3 gives detailed statistics, where for both methods, their p-values are less than 2.2E-16 and the Wilcoxon scores are close. In general, this analysis result illustrates both methods can make reasonable calculation on gene distance, which therefore provides a solid base for further evaluation on miRNA. We also observed that the similarity scores of intra-PFam gene pairs are ranked within the top 18% and 19% among all gene pair similarities generated by the edge-based method and the hybrid-based method, respectively. In the rest of the analysis, we focus on the edge-based method mainly because it gives more spread-out similarity values (with respective to the mean), making it more efficient to distinguish various similarity levels.

Table 3.3: Wilcoxon analysis on both intra- and inter- set comparison using edge-based method and hybrid-based methods.

Statistics	**Edge-based method**	**Hybrid-based method**
W-score	$6.4e+13$	$7.53e+13$
P-value	$< 2.2e-16$	$< 2.2e-16$
Mean +Std. on (Intra-PFam-gene-set)	$0.41+0.13$	$0.13+0.10$
Mean +Std. on (Inter-PFam-gene-set)	$0.30+0.13$	$0.04+0.03$

3.1.3 MiRNA Similarity Computation

MiRNA similarity computation is similar to gene pair similarity computation. We first built a miRA-gene table,and filled it with pairwise gene similarity scores calculated from the former section. MiRNA similarity computation is defined as follows:

$$s^*(m_i, m_j) = \frac{\sum_{g_k \in G(m_i)} \max_{g_n \in G(m_j)}\{s(g_k, g_n)\} + \sum_{g_n \in G(m_j)} \max_{g_k \in G(m_i)}\{s(g_k, g_n)\}}{|G(m_i)| + |G(m_j)|} \quad (4)$$

Where $G(m_i)$ and $G(m_j)$ are the target gene sets regulated by miRNAs, m_i, m_j respectively.

At last, we calculated the pairwise miRNA similarities among all 2,588 human miRNAs reported in miRBase [23] using edge-based scoring approach. Three different sets of gene targets include the experimental validated targets from miRTarBase and the predicted targets from TargetScan and miRDB were applied. Figure 3.7 shows similar distributions of all miRNA pairwise similarity based on the different target sets, where the scores are within the similar range from 0 to 1. It is obvious that with more predicted miRNAs (some are noises) included (Figure 3.7B), more diverse functions show up, which leads to relatively more dissimilar miRNA pairs.

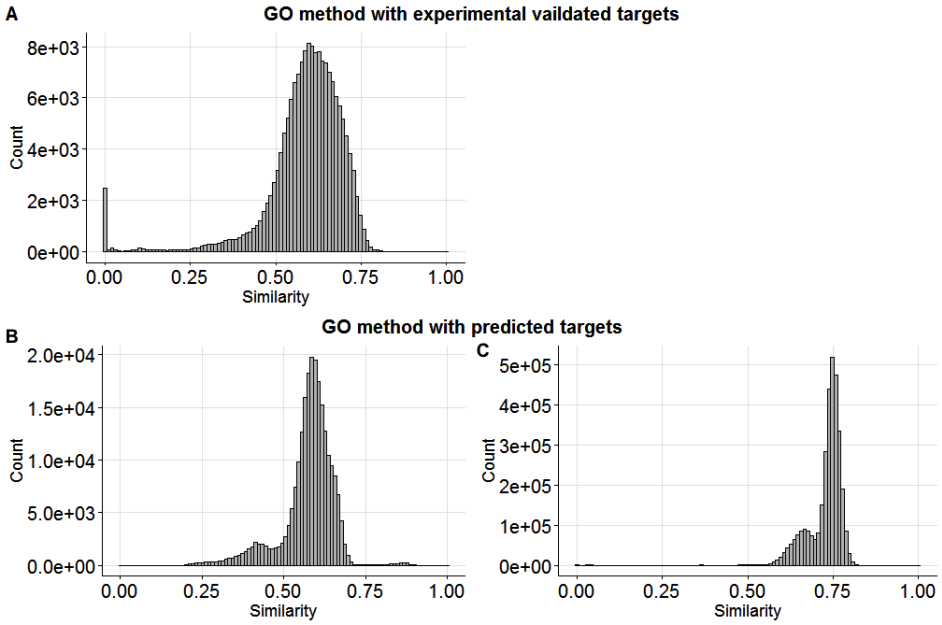


Figure 3.7: The edge-based similarity distributions on all pairs among 2,588 human miRNAs. The miRNA associated genes include both experimental targets from miRTarBase (A) and predicted targets from TargetScan (B) and miRDB (C)

3.2 Alternative Approaches for MiRNA Similarity Computation

In addition to GO-based similarity measurement, we designed another similarity measure system by utilizing functional information annotated from two different resources, PFam protein families dataset and GSEA biological pathway dataset. After this step, we can obtain a comprehensive understanding of a miRNA pair similarity from the perspectives of GO, protein family and biological pathway. At last, a density graph was generated for each similarity measure system to analyze their performance.

This miRNA similarity computation approach consists of two steps. At first, we built an interaction gene rate table for each miRNA. The table is populated with the interaction gene rate defined in Formula 5. It refers to the percentage of common genes regulated by

one miRNA and involved in a specific pathway or Protein family.

$$P(m_i) = \{\frac{G_{m_i} \cap G_{a_1}}{|G_{m_i}|}, \frac{G_{m_i} \cap G_{a_2}}{|G_{m_i}|}, ..., \frac{G_{m_i} \cap G_{a_n}}{|G_{m_i}|}\} (5)$$

Where G_{m_i} represents the set of gene targets of a miRNA, m_i, G_{a_i} is the set of genes belong to the same functional family or pathway. For instance, below is an example of interaction gene table for miRNA *hsa-miR-186-3p* based on pathway annotation system. The header row displays the example of biological pathways and for the fraction number. Denominator is the size of target gene set regulated by *miRNA hsa-miR-186-3p*. The numerator is the common genes regulated by *hsa-miR-186-3p* and also involved in the specific pathways.

Table 3.4: Interaction gene table for miRNA *hsa-miR-186-3p* based on pathway annotation

	$Metabolism\_Of\_RNA$	$MRNA\_Processing$	...	$MRNA\_Splicing$
$hsa-miR-186-3p$	26/88	5/188	...	12/88

After the interaction table was built, distance measurement was applied to miRNA pairs to get the distance between them. Euclidean Distance is applied here, which can be defined as:

$$s(p_{m_i}, p_{m_j}) = \sqrt[2]{(p_{m_i\_1}, p_{m_j\_1})^2 + (p_{m_i\_2}, p_{m_j\_2})^2 + ... + (p_{m_i\_n}, p_{m_j\_n})^2} = \sqrt[2]{\sum_{k=1}^{n} (p_{m_i\_k} - p_{m_j\_k})^2}$$

3.3 Performance Evaluation of the Three MiRNA Similarity Systems

Finally, we compared these three similarity systems using the density graphs shown in Figure 3.8. As mentioned at the beginning of this chapter, we have three miRNA-target datasets from miRTarbase, TargetScan and miRDB respectively. miRTarbase is an experimental

dataset and the other two are computational datasets. When applied the three miRNA similarity evaluations on the three datasets we used. We observed very similar patterns between the PFam- and pathway-based measurements (Figure 3.8 A and 3.8 B) where the predicted target sets from miRDB (pink) and TargetScan (blue) render condensed similarity measure due to the relative large numbers of targets compared to the experimental validated set from miRTarBase (green). In contrast, the GO-based system (Figure 3.8 C) shows a complementary measurement that smooths the distance distribution based on the validated targets(green). In general these graphs illustrate similar patterns, indicating high consistency among the three functional annotation systems.

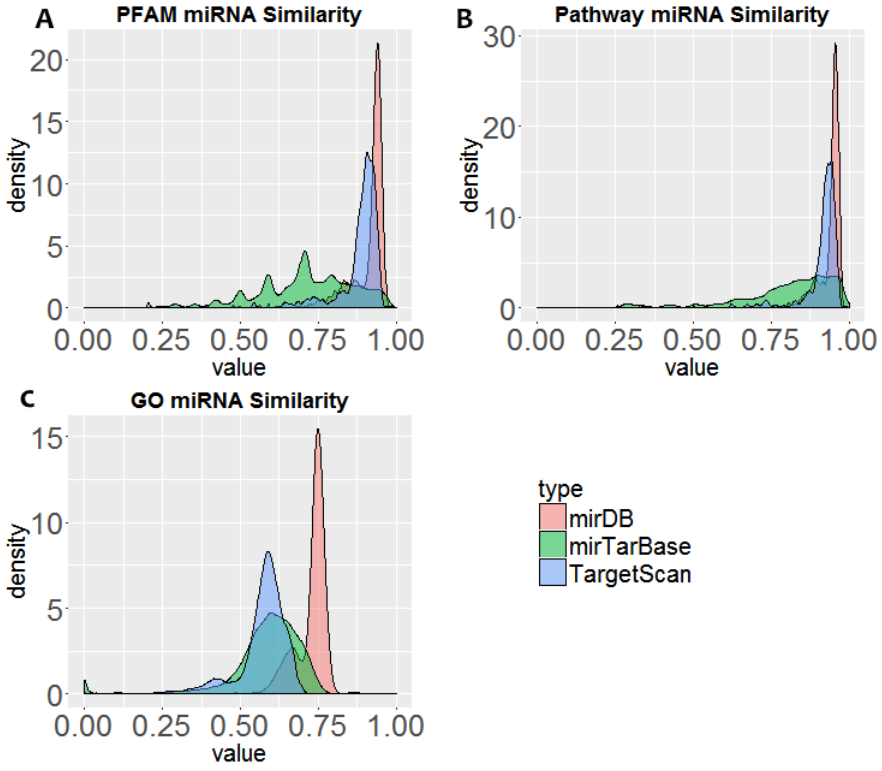


Figure 3.8: The distance distributions on pairwise miRNA based on GO, Pfam, and pathway annotaton

Chapter 4

Integration of Three Similarity Measures

In order to integrate the similarity calculation from all three methods, we first ranked all the miRNA pairs in each of the miRNA similarity system from the most similar to least similar to get a ranking list. Then we performed rank aggregation on three ranking lists. The Rank Aggregation algorithm is to combine many different rank orderings on the same set of candidates or alternatives, in order to obtain a consensus ordering [15].

The RobustRankAggreg (RRA) package in R [29] was used to get the consensus ranking among GO ranking list, PFam ranking list and pathway ranking list.

4.1 Rank Aggregation Algorithm

RRA was proposed by Kolde et al achieved the ranking result using order statistics with binomial probability [29]. It is largely used in bioinformatic domain due to its high noise tolerance and efficiency. The aggregation goal can be obtained through three steps. At the first step, a normalized rank vector is generated based on the ranks of an item in each list.

It is defined as:

$$r = \{\frac{r_i}{m} | i \in n\}$$

Where m is the number of items in the list and n is total number of lists. Then vector r is sorted in ascending order. Take a miRNA pair, (miR-639 and miR-208b-3p) as an example. This miRNA pair ranks top 22 in GO ranking list, top 15 in PFam ranking list and top 11 in pathway ranking list. Meanwhile miRTarbase dataset contain 169071 miRNA pairs in total. Therefore the ranking vector for this miRNA pair is (11/169071, 15/169071, 22/169071)

At the next step, a binomial probability is applied based on the comparison with the uniformly distributed values.

In most biological studies, some items in an experiment are noises and not reliable. Therefore it is also highly likely that only a subset of relevant items are informative in a list. To solve this problem, Kolde et al used a null model, which describes distribution of ranks when all studies produce irrelevant results, and estimates statistical significance. The simplest possible null model assumes that all studies are non-informative and produce randomly ordered item lists [29]. The goal here is to find the items highly ranked in many lists and ignore the small portion of non-informative cases. The author use binomial probability on the ordered vectors r. By evaluating the probability that $\hat{r}_{(k)} \leq r_{(k)}$ where $\hat{r}$ is the rank vector generated by the null model, each item of which is uniformly distributed. Therefore the probability that $\hat{r}_{(k)} \leq r_{(k)}$ is defined as:

$$\beta_{k,n}(r) = \sum_{l=k}^{n} \binom{n}{l} r_k^l (1 - r_k)^{n-l}$$

At last, the final rank for r is defined as :

$$\rho(r) = \min_{k=1,\ldots,n} \beta_{k,n}(r)$$

Since $\hat{r}$ is a uniformly generated one from null module, $min(\beta)$ means r_k is least possible to be randomly generated. Finally $\rho(r)$ is converted to $p-value$ through Bonferroni correction [8].

4.2 Consensus Similarity System Computation

When applied with the three similarity ranking lists, RRA generates a consensus ranking with $p-value$ for each item. Since we only have three ranking lists, and a large number of miRNA pairs even in our smallest dataset (miRTarbase $\approx 170k$). Therefore there is not enough information for RRA to rank all items with significant p value. However we only care about the pairs that are functionally close enough to form miRNA modules. Therefore we picked all the miRNA pairs with $p-value \leq 5\%$; Those pairs cover 11% of all the miRNA pairs. Next to estimate the contribution of each ranking to the consensus ranking, we applied Poisson Linear Regression[36] to calculate the coefficients for each of the three similarity. The formula we used for the Poisson linear regression calculation is defined as

$$C_{rank} = \alpha * R_{GO} + \beta * R_{PFam} + \gamma * R_{pathway} \quad (6)$$

where R_{GO} , R_{PFam} and $R_{pathway}$ refer to the 3 ranking lists of GO similarity, PFam similarity, pathway similarity and y is the consensus ranking.

Table 4.1: The coefficients obtained from the linear regression systems with $5\% p-value$ threshold

α **(GO)**	β **(PFam)**	γ **(pathway)**
$3.126e-05$	$1.899e-05$	$6.130e-05$

At last, to get our consensus similarity scores, we applied the coefficients generated above

into formula 7 below:

$$C_{sim} = \alpha * S_{GO} + \beta * S_{PFam} + \gamma * S_{pathway} \quad (7)$$

where S_{GO} , S_{PFam} and $S_{pathway}$ refer to the GO similarity score, PFam similarity score and pathway similarity score respectively.C_{sim} refers to the consensus similarity score for all miRNA pairs. The new consensus similarity system will be used downstream for miRNA functional modules detection.

Figure 4.1 shows the transition from the separate similarity systems to the consensus similarity system. A) is similarity measure based on GO annotation system. B) is similarity measure based on PFam annotation system. C) is the similarity measure based on pathway annotation system. D) is the consensus similarity system

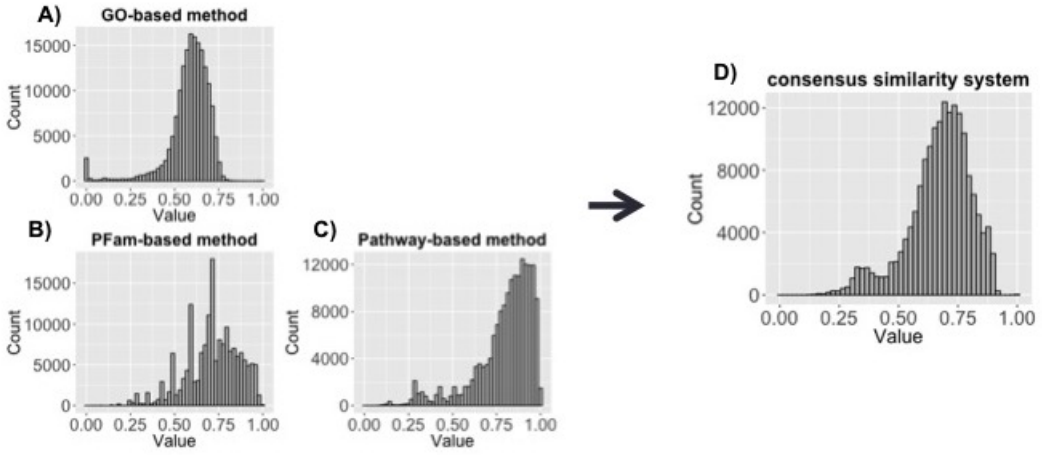


Figure 4.1: Transition graphs from miRNA similarity measurements based on GO, PFam and pathway to consensus similarity measure. x axis represents the similarity of miRNA pairs. y axis is the number of miRNA pairs

Chapter 5

Detection of MiRNA Functional Modules

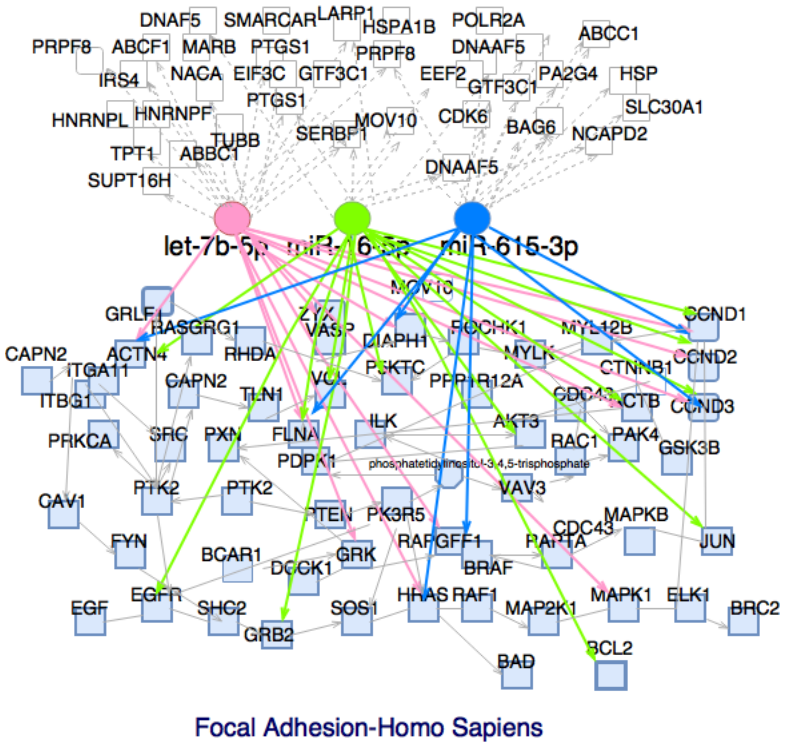
5.1 Functionally Related MiRNA Modules

5.1.1 Statistical MiRNA Module Analysis

In order to display the usefulness of similarity measure in identifying miRNA modules that can cooperatively regulate human genes, we collected 181 miRNA modules that have been predicted by a published statistical models [12]. 56 unique miRNAs are involved, which form into miRNA modules of different sizes (from 2 to 4). We conducted the similarity measure among these miRNAs and found all 221 pairs from these 181 modules are consistently ranked top 13% out of the 169, 071 miRNA pairs among all three ranking lists based on GO, PFam, and pathway. From this analysis, we confirmed several modules that are most functional relevant such as (miR-484, miR-615-3p, and let-7b-5p), (miR-16-5p and miR-92a-3p), (miR-455-3p and miR-652), (miR-877, hsa-miR-92a and miR-615-3p) and (miR-93, let-7b, miR-488). Table 5.1 shows us five of the highly possible cooperative modules among the 181

modules.

In addition, to display the usefulness of our miRNA similarity system, we used an miRNA modules (hsa-let-7b-5p/-miR-615-3p/-16-5p) with high functional similarity to create an example of miRNA regulation network.



Focal Adhesion-Homo Sapiens

Figure 5.1: Illustration of hsa-let-7b-5p/-miR-615-3p/-16-5p involved in Focal Adhesion pathway

In Figure 5.1, the blue nodes represents the genes involved in the biological pathway Focal Adhesion. Each miRNA is highlighted with different color circle node. The blue square nodes represents genes consisting of the focal adhesion pathway. The colored edges from their corresponding miRNAs to refer to the regulation relation between the miRNAs and their target genes involved in the pathway.

Table 5.1: A list of highly possible cooperative module among miRNAs

Module Index	miRNAs	Similarity: GO PFam pathway	3 Most Enriched Pathway	Enriched GO Terms
1	miR-30a miR-615-3p	0.78 0.96 0.97	1)KEGG metabolic pathway 2)Reactome metabolism of proteins 3)Reactome immune system	1)GO:0006487 protein N-linked glycosylation 2)GO:0035335 peptidyl-tyrosine dephosphorylation 3)GO:0016021 integral component of membrane
44	miR-455-3p miR-652	0.79 0.98 0.99	1)Reactome translation 2)Reactome metabolism of proteins 3)Reactome metabolism of mRNA	1)GO:0005840 ribosome 2)GO:0006415 translational termination 3)GO:0071934 thiamine transmembrane transport
78	miR-877 miR-92a miR-615-3p	$0.78 - 0.80$ $0.97 - 0.98$ $0.97 - 0.99$	1)Reactome immune system 2)Reactome cell cycle 3)KEGG metabolic pathway	1)GO:0005515 protein binding 2)GO:0005813 centrosome 3)GO:0005737 cytoplasm
85	miR-93 let-7b miR-488	$0.78 - 0.80$ $0.97 - 0.98$ $0.98 - 0.99$	1)Reactome immune system 2)KEGG metabolic pathway 3)Reactome adaptive immune system	1)GO:0048664 neuron fate determination 2)GO:0000209 protein polyubiquitination 3)GO:0005200 structural constituent of cytoskeleton
103	miR-324-3p miR-18a*	0.77 0.97 0.98	1)Reactome translation 2)Reactome metabolism of proteins 3)KEGG metabolic pathway	1)GO:0005761 mitochondrial ribosome 2)GO:0060491 regulation of cell projection assembly 3)GO:0000922 spindle pole

5.1.2 Cluster Detection

To find the miRNA clusters, we used Markov Chain Cluster Algorithm(MCL) [14]. It is an unsupervised cluster algorithm based on the concept of random walk in graphs. MCL has been approved to be one of the most efficient algorithms in detecting clusters in biochemical filed such as Protein Protein Interactions (PPI) [16]. It is also widely used in many other non-biochemical areas. The rationale behind MCL is that if you start at a node, and then randomly travel to a connected node, you are more likely to stay within a cluster than travel between. MCL algorithm simulates random walks within a graph by repeating the operations of expansion and inflation. Expansion refers to matrix multiplication to expanse the length of the paths to promote the dense region. However, power of matrix can be used to find higher-length path but the effect will diminish as the flow goes on [14]. The solution for this is inflation; raise all the entries in a given column to a certain power greater than One (e.g. squaring) and rescaling the column to have the sum One again. MCL repeats these two steps, expansion and inflation until it reaches a steady state(convergence).

Through MCL, we identified fifteen clusters ranging from eighteen miRNA nodes to three miRNA nodes, among which two clusters we have found in two papers were claimed to work as modules. Figure 5.2 shows a four-node miRNA cluster claimed by Hasser et al in their paper [21], that co-regulate gene CCND2 and TNRC6B, which are two genes highly involved in breast cancer and prostate cancer.

Figure 5.3 shows a three-node miRNA cluster claimed by Hajarnis et al in their paper [36], that co-regulate genes PKD1, MIR17HG that are highly involved in kidney disease.

Figure 5.3 shows the six miRNA pairs we picked from the fifteen clusters we identified with high similarities. The range of the similarity score among these fifteen clusters is from

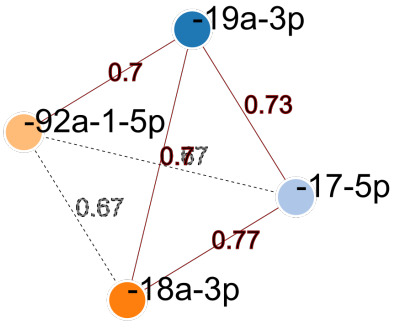


Figure 5.2: Illustration of 4 nodes miRNA clusters. Nodes represent miRNAs. Edges are their similarities

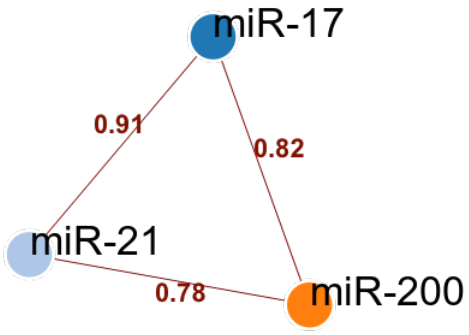


Figure 5.3: 3 nodes miRNA clusters. Nodes represent miRNAs. Edges are their similarities

0.7 to 0.9, which again confirms that miRNAs pairs from the same cluster do have high similarities.

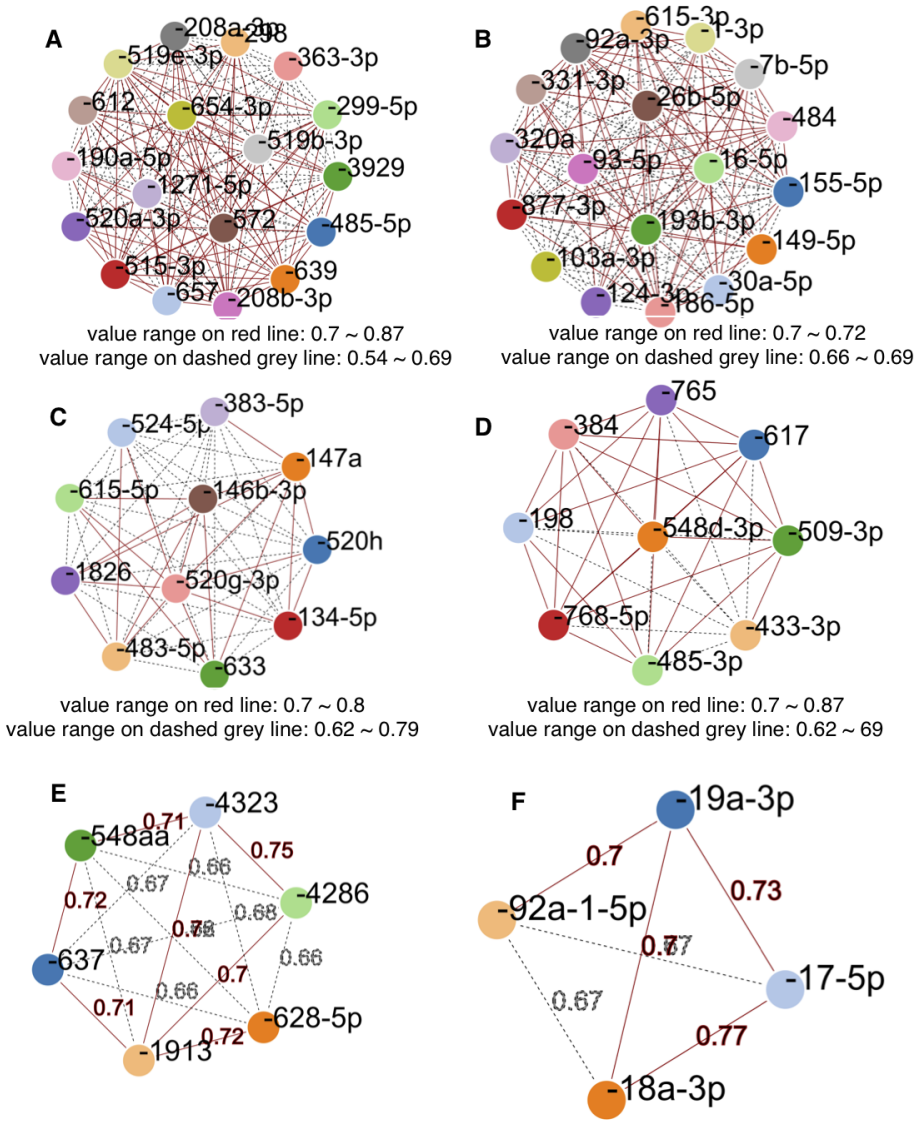


Figure 5.4: Clusters generated by MCL. Nodes represent miRNAs. Edges are the similarities among them

5.2 Validation Based on Experimental Data

We collected the gene expression profiles from two miRNA knockout/transfection experiments to validate our similarity estimation. In the first experiment, hsa-miR-141-3p and hsa-miR-200c-3p were knocked out, respectively, in human SK-BR-3 cell line [37], while the second set of data was collected based on the transfection studies of hsa-miR-1-3p and hsa-miR-155-5p [20]. Between these two pairs of miRNAs, miR-141-3p and miR200c-3p consistently shows higher similarity than miR-1-3p and miR-155-5p (0.72 versus 0.67). Based on the comparative analysis among the deferentially expressed genes (more than 1.5 fold change) identified in each experiment, we observed more common genes altered by miR-141 and -200c (18.6%), compared to those altered by miR-1 and -155 (10.1%), which is highly consistent with the similarity assessment.

5.3 MiRNA Similarity Website Implementation

To allow the open access to the results of our study, we created an online database. It provides the complete results of our study and also the information of Protein-Protein Interaction Graph (PPI) and biological pathway enrichment table, which assist users to better understand the gene miRNA functions from other perspectives. It can be freely accessed at (http://sbbi.unl.edu/microRNASim).

Chapter 6

Conclusions and Future Work

Implication of miRNAs in human health has attracted increasing number of studies [17], [18], [35], [55],[1], [26], [25] to elucidate regulatory roles in all major cellular processes that are involved in the disease development. These efforts have been focused on the identification of associated pathways through examining miRNA targets. Considering the fact that current algorithms for miRNA target prediction suffer greatly from large numbers of false positive prediction, it is advisable to first focus on miRNAs that have reliable functional annotation on the experimentally verified targets. Large number of miRNA-mRNA interactions has been discovered using high-throughput sequencing technologies. For examples, 18,514 interactions were detected by crosslinking, ligation, and sequencing of hybrids (CLASH) [22] and 72,311 were reported by covalent ligation of endogenous Argonaute-bound RNAs (CLEAR)-CLIP experiment [41].At the beginning of this analysis, the experimental validated target dataset we downloaded from miRTarBase has covered all such large-scale interaction information.

In this study, we demonstrated the pairwise similarities obtained using different methods (GO edge-based and hybrid-based) based on different annotation data and target sets with different confidence levels are promising to become a new measure for evaluation of functional relevance among miRNAs. Three annotation systems (GO, PFam and pathway) render

the functional relevance of miRNAs from different perspectives; therefore they have been combined for the inference. In addition, the approach that integrated the information content of each GO term did not provide much advantage and therefore has been removed from the downstream analysis.

Compelling evidence shows that miRNAs can regulate genes in a cooperative manner, e.g miR-17, -18a, 19a, and -92a-1 co-regulate 44 functionally related genes, such as CCND2, TNRC6B, and PHF12 [21], which we have identified as a miRNA module (clusters shown in Figure 5.2). Note that no evidence shows physical interactions between miRNAs, therefore the co-regulation through a complex may not be the case. There is an increased appreciation of examining miRNA co-regulations while our existing knowledge is extremely limited. Our analysis has shown that several miRNAs are involved in the same pathway. For example, hsa-miR-92a, -399-5p and -423-3p regulate different targets that are involved in several pathways such as Tricarboxylic acid (TCA) cycle, Heparan sulfate biosynthesis and Fc gamma R-mediated phagocytosis in kidney cancer. Meanwhile, the same gene GICLG1 is co-regulated by the same set of miRNAs under different subtypes of kidney cancer. Meanwhile, the same gene SUCLG1 is co-regulated by the same set of miRNAs under different subtypes of kidney cancers (manuscript under preparation). In addition, we also uncovered miRNA regulation of the same pathways under different conditions. For example, miR-92a, -193b and -186 co-regulated ErbB and WNT signaling pathways during the tumor development of kidney, lung, and stomach cancers. To facilitate the study along this line, our system can be used for the identification of miRNA cooperative modules as described in the previous sections. The module hsa-miR-769-3p/-193b and hsa-miR-197/-149 identified through the clustering illustrates the use of such a property.

Out of these studies, there is a lack of miRNA visualization within large biological networks and most existing tools for network construction are focused on the network of the predicted target genes. We have demonstrated in this study that by integrating the interac-

tions between miRNAs represented by the functional similarity, one will be able to include miRNA into the functional network (pathways) through existing tools such as VANESA [10] and Cytoscape [40].

Lastly, there are some technical issues involved in this study, which can be further explored, for example, the ranking aggregation for a long list of pairs (up to 234955 in our case) based on very few ranking opinions (only three) represents a challenges. Score-based approach aggregation can be also investigated.

We proposed a new system for the assessment of functional relevance of human miRNAs by integrating heterogeneous annotation data and different-level target information available in public. As demonstrated in this paper, the similarity information derived from such system can facilitate the reliable identification of miRNA co-regulatory modules and the construction of the miRNA-mediated gene regulation network. Stemming from this work, our next focus will be the integration of conditional dependent genomic data on both miRNA and their targets into this system that can capture the quantitative and dynamic properties of miRNA regulation system and better facilitate the automatic detection of miRNA functional modules. In addition, with the increased appreciation of dietary miRNA research, particularly on its bioavailable and biological roles in human health, we are motivated to integrate this system into the ongoing development of an exogenous miRNA discovery pipeline and a dynamic model on miRNA regulation under development in our group, with the hope of providing the whole miRNA community an integrated platform with much more comprehensive analytical functions on both endogenous and exogenous miRNA.

Bibliography

[1] Ali Sobhi Afshar, Joseph Xu, and John Goutsias. Integrative identification of deregulated mirna/tf-mediated gene regulatory loops and networks in prostate cancer. *PLoS One*, 9(6):e100806, Jun 2014. PONE-D-14-02584[PII].

[2] Yukihiro Akao, Yoshihito Nakagawa, and Tomoki Naoe. MicroRNA-143 and -145 in colon cancer. *DNA and Cell Biology*, 26(5):311–320, may 2007.

[3] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nat Genet*, 25(1):25–29, May 2000. Commentary.

[4] Shveta Bagga, John Bracht, Shaun Hunter, Katlin Massirer, Janette Holtz, Rachel Eachus, and Amy E. Pasquinelli. Regulation by let-7 and lin-4 mirnas results in target mrna degradation. *Cell*, 122(4):553 – 563, 2005.

[5] Shveta Bagga, John Bracht, Shaun Hunter, Katlin Massirer, Janette Holtz, Rachel Eachus, and Amy E. Pasquinelli. Regulation by ¡em¿let-7¡/em¿ and ¡em¿lin-4¡/em¿ mirnas results in target mrna degradation. *Cell*, 122(4):553–563, 2016/11/08 XXXX.

[6] David P. Bartel. Microrna: Target recognition and regulatory functions. *Cell*, 136(2):215 – 233, 2009.

[7] N. J. Beveridge, E. Gardiner, A. P. Carroll, P. A. Tooney, and M. J. Cairns. Schizophrenia is associated with an increase in cortical microrna biogenesis. *Mol Psychiatry*, 15(12):1176–1189, Dec 2010. 19721432[pmid].

[8] C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.

[9] Michelle D. Brazas, Joseph T. Yamada, and B. F. Francis Ouellette. Providing web servers and training in bioinformatics: 2010 update on the bioinformatics links directory. *Nucleic Acids Research*, 38(suppl 2):W3–W6, 2010.

[10] Christoph Brinkrolf, Sebastian Janowski, Benjamin Kormeier, Martin Lewinski, Klaus Hippe, Daniela Borck, and Ralf Hofestädt. Vanesa - a software application for the visualization and analysis of networks in system biology applications. *J. Integrative Bioinformatics*, 11, 2014.

[11] The Gene Ontology Consortium. Gene ontology consortium: going forward. *Nucleic Acids Research*, 43(D1):D1049–D1056, 28 January 2015.

[12] Jun Ding, Xiaoman Li, and Haiyan Hu. Microrna modules prefer to bind weak and unconventional target sites. *Bioinformatics*, 31(9):1366–1374, 2015.

[13] Jun Ding, Xiaoman Li, and Haiyan Hu. Tarpmir: a new approach for microrna target site prediction. *Bioinformatics*, 2016.

[14] Stijn Van Dongen. Graph clustering via a discrete uncoupling process. *SIAM Journal on Matrix Analysis and Applications*, 30(1):121–141, 2008.

[15] Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, pages 613–622, New York, NY, USA, 2001. ACM.

[16] A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7):1575–1584, 2002.

[17] Maoxiao Feng, Xiaochuang Luo, Chunming Gu, Yumin Li, Xuejiao Zhu, and Jia Fei. Systematic analysis of berberine-induced signaling pathway between mirna clusters and mrnas and identification of mir-99a125b cluster function by seed-targeting inhibitors in multiple myeloma cells. *RNA Biology*, 12(1):82–91, 2015. PMID: 25826415.

[18] Miki Fuse, Satoko Kojima, Hideki Enokida, Takeshi Chiyomaru, Hirofumi Yoshino, Nijiro Nohata, Takashi Kinoshita, Shinichi Sakamoto, Yukio Naya, Masayuki Nakagawa, Tomohiko Ichikawa, and Naohiko Seki. Tumor suppressive microrna s (mir-222 and mir-31) regulate molecular pathways based on microrna expression signature in prostate cancer. *J Hum Genet*, 57(11):691–699, Nov 2012.

[19] Dimos Gaidatzis, Erik van Nimwegen, Jean Hausser, and Mihaela Zavolan. Inference of mirna targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*, 8(1):69, 2007.

[20] H Guo, NT Ingolia, JS Weissman, and DP Bartel. Mammalian microrna s predominantly act to decrease target mrna levels. *Nature*, 466(5):835–40, 2010-08-12 00:00:00.0.

[21] Jean Hausser and Mihaela Zavolan. Identification and consequences of mirna-target interactions [mdash] beyond repression of gene expression. *Nat Rev Genet*, 15(9):599–612, Sep 2014. Review.

[22] Aleksandra Helwak, Grzegorz Kudla, Tatiana Dudnakova, and David Tollervey. Mapping the human mirna interactome by clash reveals frequent noncanonical binding. *Cell*, 153(3):654–665, 2016/11/01 XXXX.

[23] Sheng-Da Hsu, Yu-Ting Tseng, Sirjana Shrestha, Yu-Ling Lin, Anas Khaleel, Chih-Hung Chou, Chao-Fang Chu, Hsi-Yuan Huang, Ching-Min Lin, Shu-Yi Ho, Ting-Yan Jian, Feng-Mao Lin, Tzu-Hao Chang, Shun-Long Weng, Kuang-Wen Liao, I-En Liao, Chun-Chi Liu, and Hsien-Da Huang. mirtarbase update 2014: an information resource for experimentally validated mirna-target interactions. *Nucleic Acids Research*, 42(D1):D78–D85, 2014.

[24] Marilena V. Iorio, Manuela Ferracin, Chang-Gong Liu, Angelo Veronese, Riccardo Spizzo, Silvia Sabbioni, Eros Magri, Massimo Pedriali, Muller Fabbri, Manuela Campiglio, Sylvie Ménard, Juan P. Palazzo, Anne Rosenberg, Piero Musiani, Stefano Volinia, Italo Nenci, George A. Calin, Patrizia Querzoli, Massimo Negrini, and Carlo M. Croce. Microrna gene expression deregulation in human breast cancer. *Cancer Research*, 65(16):7065–7070, 2005.

[25] M. D. Jansson, N. D. Damas, M. Lees, A. Jacobsen, and A. H. Lund. mir-339-5p regulates the p53 tumor-suppressor pathway by targeting mdm2. *Oncogene*, 34(15):1908–1918, Apr 2015. Original Article.

[26] Shuai Jiang, Hong-Wei Zhang, Ming-Hua Lu, Xiao-Hong He, Yong Li, Hua Gu, Mo-Fang Liu, and En-Duo Wang. Microrna-155 functions as an oncomir in breast cancer by targeting the suppressor of cytokine signaling 1 gene. *Cancer Research*, 70(8):3119–3127, 2010.

[27] Bino John, Anton J. Enright, Alexei Aravin, Thomas Tuschl, Chris Sander, and Debora S. Marks. Human MicroRNA Targets. *PLoS Biol*, 2(11):e363+, October 2004.

[28] Michael Kertesz, Nicola Iovino, Ulrich Unnerstall, Ulrike Gaul, and Eran Segal. The role of site accessibility in microrna target recognition. *Nat Genet*, 39(10):1278–1284, Oct 2007.

[29] Raivo Kolde, Sven Laur, Priit Adler, and Jaak Vilo. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, 2012.

[30] Ana Kozomara and Sam Griffiths-Jones. mirbase: integrating microrna annotation and deep-sequencing data. *Nucleic Acids Research*, 39(suppl 1):D152–D157, 2011.

[31] Azra Krek, Dominic Grun, Matthew N. Poy, Rachel Wolf, Lauren Rosenberg, Eric J. Epstein, Philip MacMenamin, Isabelle da Piedade, Kristin C. Gunsalus, Markus Stoffel, and Nikolaus Rajewsky. Combinatorial microrna target predictions. *Nat Genet*, 37(5):495–500, May 2005.

[32] Bing Liu, Jiuyong Li, and Murray J. Cairns. Identifying mirnas, targets and functions. *Brief Bioinform*, 15(1):1–19, Jan 2014. bbs075[PII].

[33] Gaston K. Mazandu and Nicola J. Mulder. Information content-based gene ontology semantic similarity approaches: Toward a unified framework theory. *Biomed Res Int*, 2013:292063, Sep 2013. 24078912[pmid].

[34] Mariana R. Mendoza, Guilherme C. da Fonseca, Guilherme Loss-Morais, Ronnie Alves, Rogerio Margis, and Ana L. C. Bazzan. Rfmirtarget: Predicting human microrna target genes with a random forest classifier. *PLoS One*, 8(7):e70153, Jul 2013. PONE-D-13-09878[PII].

[35] Avaniyapuram Kannan Murugan, Arasambattu Kannan Munirajan, and Ali S. Alzahrani. Micrornas: Modulators of the ras oncogenes in oral cancer. *Journal of Cellular Physiology*, 231(7):1424–1431, 2016.

[36] Lama Noureddine, Sachin Hajarnis, and Vishal Patel. Micrornas and polycystic kidney disease. *Drug Discovery Today: Disease Models*, 10(3):e137 – e143, 2013. MicroRNAs involved in disease with emphasis on fibrosis.

[37] Y. oung-K. ook Kim, G. abbine Wee, J. oha Park, J. ongkyu Kim, D. aehyun Baek, J. in-S. oo Kim, and V. N. arry Kim. - TALEN-based knockout library for human micrornas. - 20(- 12):– – 1464, - 2013/12//print.

[38] Kati P. Porkka, Minja J. Pfeiffer, Kati K. Waltering, Robert L. Vessella, Teuvo L.J. Tammela, and Tapio Visakorpi. Microrna expression profiling in prostate cancer. *Cancer Research*, 67(13):6130–6135, 2007.

[39] Herv Seitz. Redefining microrna targets. *Current Biology*, 19(10):870 – 873, 2009.

[40] Michael E. Smoot, Keiichiro Ono, Johannes Ruscheinski, Peng-Liang Wang, and Trey Ideker. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, 27(3):431–432, 2011.

[41] Prashant K. Srivastava, Taraka Ramji Moturu, Priyanka Pandey, Ian T. Baldwin, and Shree P. Pandey. A comparison of performance of plant mirna target prediction tools and the characterization of features for genome-wide target prediction. *BMC Genomics*, 15(1):348, 2014.

[42] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–15550, Oct 2005. 010215545[PII].

[43] J. O. U. R. Ty. A1 - The UniProt ConsortiumT1 - UniProt: a hub for protein informationY1 - 2014/10/27JF - Nucleic Acids ResearchJO - Nucleic Acids ResearchN1 - 10.1093/nar/gku989UR - http://nar.oxfordjournals.org/content/early/2014/10/27/nar.gku989.abstractN2 - UniProt is an important collection of protein sequences and their annotations, which has doubled in size to 80 million sequences during the past year. This growth in sequences has prompted an extension of UniProt accession number space from 6 to 10 characters. An increasing fraction of new sequences are identical to a sequence that already exists in the database with the majority of sequences coming from genome sequencing projects. We have created a new proteome identifier that uniquely identifies a particular assembly of a species and strain or subspecies to help users track the provenance of sequences. We present a new website that has been designed using a user-experience design process. We have introduced an annotation score for all entries in UniProt to represent the relative amount of knowledge known about each protein. These scores will be helpful in identifying which proteins are the best characterized and most informative for comparative analysis. All UniProt data is provided freely and is available on the web at http://www.uniprot.org/.ER -.

[44] Ioannis S. Vlachos, Konstantinos Zagganas, Maria D. Paraskevopoulou, Georgios Georgakilas, Dimitra Karagkouni, Thanasis Vergoulis, Theodore Dalamagas, and Artemis G. Hatzigeorgiou. Diana-mirpath v3.0: deciphering microrna function with experimental support. *Nucleic Acids Res*, 43(Web Server issue):W460–W466, Jul 2015. 25977294[pmid].

[45] Song Wang, Yanxun Zhao, Dongsheng Li, Liangchen Zhu, and Zugang Shen. Identification of biomarkers for the prognosis of pancreatic ductal adenocarcinoma with mirna microarray data. 30(2):0.

[46] Xiaowei Wang. mirdb: A microrna target prediction and functional annotation database with a wiki interface. *RNA*, 14(6):1012–1017, Jun 2008. RA[PII].

[47] Xiaowei Wang. Improving microrna target prediction by modeling with unambiguously identified microrna-target pairs from clip-ligation studies. *Bioinformatics*, 32(9):1316–1322, 2016.

[48] Nathan Wong and Xiaowei Wang. mirdb: an online resource for microrna target prediction and functional annotations. *Nucleic Acids Research*, 2014.

[49] Xiaomei Wu, Erli Pang, Kui Lin, and Zhen-Ming Pei. Improving the measurement of semantic similarity between gene ontology terms and gene products: Insights from an edge- and ic-based hybrid method. *PLoS One*, 8(5):e66745, May 2013. PONE-D-13-06360[PII].

[50] Xiaomei Wu, Lei Zhu, Jie Guo, Da-Yong Zhang, and Kui Lin. Prediction of yeast proteinprotein interaction network: insights from the gene ontology and annotations. *Nucleic Acids Research*, 34(7):2137–2150, 2006.

[51] Peter Yakovchuk, Ekaterina Protozanova, and Maxim D. Frank-Kamenetskii. Base-stacking and base-pairing contributions into thermal stability of the dna double helix. *Nucleic Acids Research*, 34(2):564, 2006.

[52] Nozomu Yanaihara, Natasha Caplen, Elise Bowman, Masahiro Seike, Kensuke Kumamoto, Ming Yi, Robert M. Stephens, Aikou Okamoto, Jun Yokota, Tadao Tanaka, George Adrian Calin, Chang-Gong Liu, Carlo M. Croce, and Curtis C. Harris. Unique microrna molecular profiles in lung cancer diagnosis and prognosis. *Cancer Cell*, 9(3):189–198, 2016/11/08 XXXX.

[53] Da Yang, Yan Sun, Limei Hu, Hong Zheng, Ping Ji, ChadV. Pecot, Yanrui Zhao, Sheila Reynolds, Hanyin Cheng, Rajesha Rupaimoole, David Cogdell, Matti Nykter, Russell Broaddus, Cristian Rodriguez-Aguayo, Gabriel Lopez-Berestein, Jinsong Liu, Ilya Shmulevich, AnilK. Sood, Kexin Chen, and Wei Zhang. Integrated analyses identify a master microrna regulatory network for the mesenchymal subtype in serous ovarian cancer. *Cancer Cell*, 23(2):186 – 199, 2013.

[54] Hua Yang, William Kong, Lili He, Jian-Jun Zhao, Joshua D. O'Donnell, Jiawang Wang, Robert M. Wenham, Domenico Coppola, Patricia A. Kruk, Santo V. Nicosia, and Jin Q. Cheng. Microrna expression profiling in human ovarian cancer: mir-214 induces cell survival and cisplatin resistance by targeting pten. *Cancer Research*, 68(2):425–433, 2008.

[55] Wenyu Zhang, Jin Zang, Xinhua Jing, Zhandong Sun, Wenying Yan, Dongrong Yang, Feng Guo, and Bairong Shen. Identification of candidate mirna biomarkers from mirna regulatory network with application to prostate cancer. *Journal of Translational Medicine*, 12(1):66, 2014.

[56] Xi-Mei Zhang, Lin Guo, Mei-Hua Chi, Hong-Mei Sun, and Xiao-Wen Chen. Identification of active mirna and transcription factor regulatory pathways in human obesity-related inflammation. *BMC Bioinformatics*, 16(1):76, 2015.

[57] Xiaoying Zhang, Murray Cairns, Barbara Rose, Christopher O'Brien, Kerwin Shannon, Jonathan Clark, Jennifer Gamble, and Nham Tran. Alterations in mirna processing and expression in pleomorphic adenomas of the salivary gland. *International Journal of Cancer*, 124(12):2855–2863, 2009.