Machine Learning Approaches for Identifying microRNA Targets and Conserved Protein Complexes

Hanaa Aboelenen Abdelgiad Torkey

Dissertation submitted to the Faculty of
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science and Application

Lenwood S. Heath, Chair
Ruth Grene
Xinwei Deng
Liqing Zhang
Mahmoud M. ElHefnawi

17^{th} April, 2017
Blacksburg, Virginia

Keywords: microRNA target, machine learning, algorithms, optimization, graph mining, network alignment, protein complex.
Copyright 2017, Hanaa Torkey

Machine Learning Approaches for Identifying microRNA Targets and Conserved Protein Complexes

Hanaa Aboelenen Abdelgiad Torkey

ABSTRACT

Much research has been directed toward understanding the roles of essential components in the cell, such as proteins, microRNAs, and genes. This dissertation focuses on two interesting problems in bioinformatics research: microRNA-target prediction and the identification of conserved protein complexes across species. We define the two problems and develop novel approaches for solving them. MicroRNAs are short non-coding RNAs that mediate gene expression. The goal is to predict microRNA targets. Existing methods rely on sequence features to predict targets. These features are neither sufficient nor necessary to identify functional target sites and ignore the cellular conditions in which microRNA and mRNA interact. We developed MicroTarget to predict microRNA-mRNA interactions using heterogeneous data sources. MicroTarget uses expression data to learn candidate target set for each microRNA. Then, sequence data is used to provide evidence of direct interactions and ranking the predicted targets. The predicted targets overlap with many of the experimentally validated ones. The results indicate that using expression data helps in predicting microRNA targets accurately.

Protein complexes conserved across species specify processes that are core to cell machinery. Methods that have been devised to identify conserved complexes are severely limited by noise in PPI data. Behind PPIs, there are domains interacting physically to perform the necessary functions. Therefore, employing domains and domain interactions gives a better view of the protein interactions and functions. We developed novel strategy for local network alignment, DONA. DONA maps proteins into their domains and uses DDIs to improve the network alignment. We developed novel strategy for constructing an alignment graph and then uses this graph to discover the conserved subnetworks. DONA shows better performance in terms of the overlap with known protein complexes with higher precision and recall rates than existing methods. The result shows better semantic similarity computed with respect to both the biological process and the molecular function of the aligned subnetworks.

Machine Learning Approaches for Identifying microRNA Targets and Conserved Protein Complexes

Hanaa Aboelenen Abdelgiad Torkey

GENERAL AUDIENCE ABSTRACT

Much research has been directed toward understanding the roles of essential components in the cell, such as proteins, microRNAs, and genes. The processes within the cell include a mixture of small molecules. It is of great interest to utilize different information sources to discover the interactions among these molecules. This dissertation focuses on two interesting problems: microRNA-target prediction and the identification of conserved protein complexes across species. We define the two problems and develop novel approaches for solving them. MicroRNAs are a recently discovered class of non-coding RNAs. They play key roles in the regulation of gene expression of as much as 30% of all mammalian protein encoding genes. MicroRNAs regulation activity has been implicated in a number of diseases including cancer, heart disease and neurological diseases. We developed MicroTarget to predict microRNA-gene interactions using heterogeneous data sources. The predicted target genes overlap with many of the experimentally validated ones.

Proteins carry out their tasks in the cell by interacting with each other. Protein complexes conserved among species specify the cell core processes. We identify conserved complexes by constructing an alignment graph leveraging on the conservation of PPIs between species through domain conservation and domain-domain interactions (DDI) in addition to PPI networks. Better integration of domain conservation and interactions in our developed conserved protein complexes identification system helps biologists benefit from verified data to predict more reliable similarity relationships among species. All the test data sets and source code for this dissertation are available at:
`https://bioinformatics.cs.vt.edu/~htorkey/Software.`

Dedication

I would like to dedicate this thesis to my loving parents.

Acknowledgments

I would like to thank the Almighty God. I would like also to express my gratitude and thanks to my advisor Prof. Heath, for his time, guidance, continuous encouragement, and valuable discussions on my dissertation work through the past four years. He been a great support to me and without you, I would not have been able to stay focused and finish my PhD work. It would take more than few words to express my gratitude to you.

I thank my committee members, Prof. Grene, Prof. Dong, Prof. Zhang, and Prof. ElHefnawi for their support, cooperation and comments to improve my work all along the way. Special thanks to Prof. Grene who always found a time for me to meet and discuss. She always supported me and provided me with valuable ideas to verify my computational methods from biological perspective. Special thanks for VT-MENA program director, prof. Sedki Riad.

I am eternally in debt to my parents, without them I could not be able to complete my PhD. Special thanks to my dear mother and Father for their love, and caring after me when I really needed him. Thanks to my beloved sisters and brother Abdo for continuous support and encouragement.

My beloved brother, Mohammed Torkey who I can't find words for his support, sacrifices and trying to make it work for me. I'm very grateful for having him in my life. My sincere gratitude to all my friends, specially Sherin Gannam, who I met here in the United States for their unlimited support, love, and help whenever I needed.

Contents

1 Introduction **1**

1.1 MicroRNA Target Prediction 1

1.1.1 Motivations and contributions 2

1.2 Identifying Conserved Protein Complexes 4

1.2.1 Motivations and contributions 6

1.3 Dissertation Organization 7

2 MicroRNA Target Prediction: Biological Background **9**

2.1 MicroRNA 9

2.1.1 MicroRNA Biogenesis 10

2.1.2 microRNA Mechanism of Action 10

2.2 Experimental Identification of microRNA Targets 11

3 MicroRNA Target Prediction: Literature Review **15**

3.1 Principles of microRNA target recognition 15

3.1.1 Sequence complementary of seed binding site 15

3.1.2 Site accessibility 16

3.1.3 Conservation 17

3.1.4 Thermodynamic stability 17

3.2 Computational target prediction methods 17
3.2.1 Rule-based methods 18
3.2.2 Machine Learning Methods 20
3.2.3 Model-Based Methods 21

4 MicroTarget: microRNA Target Prediction Approach 23

4.1 Preliminaries and Problem Definition 24
4.2 The Proposed Approach 26
4.2.1 MiRLasso for graph structure learning 26
4.2.2 Learning microRNA Direct Targets 33
4.2.3 Scoring microRNA targets 34
4.2.4 Target ranking 37
4.3 MicroTarget Results 38
4.3.1 Data sources 38
4.3.2 Performance comparison with existing methods 39
4.3.3 Studying the tissue-specificity of the prediction 44
4.3.4 Analysis of the scoring features 45
4.3.5 Evaluating SVR model for the ranking 46
4.4 Discussion 48

5 Conserved Protein Complexes: Biological Background 51

5.1 Protein-protein interaction 51
5.1.1 Identifying Protein Interactions 52
5.2 Protein Structure 55
5.2.1 Structural domains 57

5.2.2 Domain-Domain Interactions 57
5.3 Protein complex 58

6 Conserved Protein Complexes: Literature Review 59

6.1 PPI Comparative Analysis 59
6.2 Existing LNA methods 61
6.2.1 Alignment graph based methods 61
6.2.2 Information Fusion Methods 63
6.2.3 Other Methods 65

7 DONA: Identifying Conserved Protein Complexes 67

7.1 Problem Definition 67
7.2 The proposed approach 68
7.2.1 DONA framework 69
7.2.2 Alignment graph Construction 69
7.2.3 Scoring the alignment graph 73
7.2.4 Alignment graph Search 76
7.3 DONA Results 80
7.3.1 Data sets 80
7.3.2 Case study 82
7.3.3 Comparison with other methods 82
7.3.4 Biological relevance of conserved subnetworks 87
7.3.5 The effect of MCL parameter on the performance 90
7.4 Discussion 91

8 Conclusions and Future Directions 96

8.1 MicroRNA target prediction 96
8.1.1 Future direction 98
8.2 Identifying conserved complexes 99
8.2.1 Future direction 100

Bibliography **101**

List of Figures

2.1 microRNA biogenesis and mechanism of action. It go under several processing steps before maturation to its active form. After processing, the mature microRNA incorporates into the RNA-induced silencing complex, then binds to the complementary sites in the 3′-UTR of their target genes. microRNA down-regulates the protein synthesis via translation repression or mRNA degradation [22]. 14

4.1 The conceptual view of MicroTarget includes using microRNA and mRNA expression data to infer the candidate targets for each microRNA, using sequence data to get the direct microRNA-targets interactions, and finally scoring and validate results. 27

4.2 An example of the precision matrix and its corresponding graph structure . . 28

4.3 Comparison with the existing methods with the percentage of the overall validated targets that have been predicted by each method. 40

4.4 Small network for mir-96 and mir-141 and their predicted targets from our approach. 41

4.5 Z-score comparison with the existing methods for the top scored targets. . . 42

4.6 The ROC curves of MicroTarget, targetScan, MirWalk and GenMiR++. . . 43

4.7 Venn diagram for the miR-200 family predicted targets versus experimentally validated targets. Numbers in the yellow circle are the experimentally validated targets from MirTarBase and MirWalk. 45

4.8 ROC analysis for the SVR model with different data sets 47

4.9 Total ranking score for the top 100, 200, and 300 scored target with different kernel functions for the SVR model. 49

5.1 PPI identification methods; A) The yeast-two-hybrid system: If protein X and protein Y interact, then their DNA-binding domain (DBD) and activation domain (AD) will combine to form a functional transcriptional activator, UAS refers to upstream activator sequence of the promoter [20]. B) affinity purification coupled to mass spectrometry; first, tagged protein is pulled down via its tag together with the associated proteins and other non-specific interacting proteins. Then the protein samples collected are broken down into peptides and analyzed by mass-spectrometry. Finally, the list of peptide is sequenced and the proteins from each sample are reported as the interaction ones [141]. 53

5.2 (A) type of protein structure [129]. (B) An example of domain organization tertiary structure of protein ZPR1 as in Pfam database; the schematic illustration of the modular architecture, and ribbon representation of the tertiary structure [39]. 56

6.1 Evaluation analysis between the current methods on curated PPI that we know the real alignment in them between mouse and rat species, nodes with green colored name are the known conserved nodes. 66

7.1 The general framework for DONA. Given two input PPI networks; (i) mapping the network proteins into their domain using Pfam database is performed, (ii) the alignment graph is built, (iii) scores are assigned to its nodes and edges, (iv) and the alignment graph is clustered. 70

7.2 The types of edges in DONA alignment graph. 72

7.3 Comparing our approach DONA with the existing approach in a case study. 82

7.4 Methods comparison based on the change of the predicted complexes with F-score. 88

7.5 Precision and recall for the detected complexes in human-yeast alignment. . 89

7.6 Precision and recall for the detected conserved complexes in Mouse-Rat alignment. 89

7.7 Number of complexes detected with different inflation level in different alignment, refer to table 7.3 for the name of the alignment. 92

7.8 Number of complexes detected with different inflation level in different alignment. 93

7.9 Some examples of conserved modules found in human-mouse alignment by our approach. The original PPI networks in these modules regions include several noisy interactions, thereby reducing their topological significant when identified only by PPIs data, adding DDI improve the performance. 95

List of Tables

4.1 Breast cancer related-genes and the number of predicted microRNAs and the validated microRNAs 44

4.2 Correlation among features that are used for scoring the predicted targets. Number of matches refers to the number of seed binding sites between the microRNA and the mRNA. Matching length refers to the maximum sequence complementarity between the microRNA and the gene. Seed ΔG and total match ΔG refer the site accessibility estimated based on the seed region and the maximum sequence complementarity, respectively. P_{value} points to the P_{value} of the seed binding site prediction 46

4.3 Positive and negative data sets for SVR analysis 48

7.1 Statistics of PPI networks used. 81

7.2 The number of complexes available in databases for evaluating DONA. . . . 81

7.3 Each cell shows the symbol used to represent the different alignment throughout the chapter. 83

7.4 The number of solutions produced for each alignment in the different methods. 84

7.5 The number of known complexes hit with F-score 0.3 in the different methods, and standard error over 20 runs for DONA and AlignMCL, the number in parentheses. 85

7.6 The number of known complexes hit with F-score 0.5 in the different methods, and the standard error over 20 runs for DONA and AlignMCL, the number in parentheses. 86

7.7 The number of known complexes hit with F-score 0.7 in the different methods, and the standard error over 20 runs for DONA and AlignMCL, the number in parentheses. 87

7.8 Purity and GO enrichment analysis for mouse-rat and human-mouse alignments. 90

7.9 Purity and GO enrichment analysis human rat alignment. 91

7.10 Comparing the best matching solutions for Exocyst, and F0F1 ATP synthase complexes in mouse-rat alignment. 94

7.11 Comparing the best matching solutions for Arp 2/3, TFIID, and 20S proteasome complexes in human-fly alignment. 94

List of Abbreviations

3D Three-dimension

ADMM Alternating direction method of multipliers

AP Alignable protein pair

CE Composite edge

ceRNA Competitive endogenous RNA

DDIs Domain-domain interactions

DIOPT DRSC Integrative Ortholog Prediction Tool

GGM Gaussian graphical model

GNA Global network alignment

LNA Local network alignment

MCL Markov cluster algorithm

mRNA messenger RNA

PDB Protein Data Bank

PPIs Protein-protein interactions

ROC Receiver Operator Characteristic

SDE Simple direct edge

SIE Simple indiect edge

SVR Support vector regression

UTR Untranslated region

Y2H Yeast-two-hybrid

Chapter 1

Introduction

In this chapter, we introduce the two computational problems in bioinformatics, along with the motivations for working on these problems and contributions for the developed approaches in solving them. Then, we give a brief overview of how the dissertation is organized.

1.1 MicroRNA Target Prediction

Understanding the relationship between genes and their regulators has recently received considerable attention. Many studies have demonstrated that microRNAs are primary gene regulators at the post-transcriptional level [27]. These microRNAs are short (19-24 nucleotides in length) non-coding RNAs. They regulate genes by binding to the complementary sequences on the target messenger RNA (mRNA) transcripts. This binding activity usually results in translation repression or mRNA degradation [159]. By regulating target genes, microRNAs are involved in most biological processes, including developmental timing, cell proliferation, metabolism, differentiation, and cellular signaling [4]. Identifying microRNA target genes will give new insights into biological processes. There are many potential target sites for any given microRNA. The process of validating a microRNA target in the laboratory is time consuming and costly [74]. Computational prediction of microRNA targets will facilitate the process of narrowing down the potential targets for experimental validation.

The mechanism by which microRNA sequence complementarity conveys functional binding to mRNAs provides the rules for microRNA target prediction. Nucleotides 2 through 8

of microRNAs are called the seed region. Seed region matching has been described as a key feature for identifying microRNA targets [25]. Target prediction methods use sequence mapping along the genome for the seed region to find potential seed binding sites. A perfect match for the seed region of a microRNA occurs on average every 4 kb in a genome [118]. Therefore, the seed binding sites must be filtered to reduce the number of false positive targets [126]. Computational target prediction identifies relevant features that characterize microRNA targeting. Multiple features that are relevant to microRNA target recognition have been proposed, such as conservation of the seed region, accessibility of the seed binding site, and the stability of the binding process [154].

Current computational methods have difficulties in identifying target genes. Methods that rely on the conservation of binding sites cannot predict non-conserved targets [91]. Relying on site accessibility to filter the seed binding sites can remove true positive targets. Most prediction methods use a combination of features to compensate for the limitations of each feature alone. These methods are reviewed in Chapter 3.

Effective regulation of a target requires that the microRNA and the target be located in the same cellular compartment. Among the identified microRNAs, some exhibit tissue-specific expression patterns and play potential roles in maintaining tissue function [106]. Therefore, the study of the microRNA regulatory network using expression profiles is necessary to understand their regulation and function.

1.1.1 Motivations and contributions

Identifying microRNA targets experimentally is a costly and time consuming process; thus most researchers depend on computational tools to first predict a set of favorable targets for further experimental validation [96]. However, there are problems with the current computational methods that are used to identify microRNA targets. Most computational methods rely on using sequence data. They search for binding sites between the microRNAs and the genes, then filter out these binding sites. One way to filter those binding sites is using the conservation of seed binding sites between different species. However, recently studies show that there are microRNAs that have a large number of non-conserved target seed binding sites [56]. Xu et al. [160] shows that the identification of mRNAs and proteins that are upregulated upon inhibition or the removal of an endogenous miRNA demonstrate that non-conserved targeting is even more widespread than conserved targeting. Another way of

filtering the predicted seed binding sites is relying on site accessibility. Site accessibility is a measure of the ease and stability with which a microRNA can locate and bind with its target [67]. If the binding of a microRNA to a seed binding site is stable, the gene that contains this binding site is considered more likely to be a true target. Free energy is used as a measure of the stability of a biological system. However, free energy estimation relies on empirical measurements that may not be complete or accurate [68]. Computational methods that do not take these issues into account may produce biased results. There is a need for new methods that can detect microRNA targets and take into consideration all the factors that affect microRNA target regulation.

In the course of this dissertation, a new machine learning approach has been developed to predict microRNA-mRNA regulatory interactions with high confidence. Expression data has been employed to infer the candidate target set for each microRNA. Using only expression data will enable use to differentiate between direct and indirect interactions. Therefore, sequence data is used. Using sequence data, microRNA candidate targets are filtered with seed binding site matching. Then, the predicted targets are scored by a set of microRNA targeting features. The developed system is called MicroTarget. First, it takes mRNA and microRNA expression profiles and infers the candidate target set for each microRNA. We formulate the problem of inferring the regulation between microRNAs and mRNAs as a network structure learning problem. The problem input is a matrix of microRNA and mRNA expression values. MicroTarget predicts an undirected graph structure corresponding to the conditional dependence among the microRNAs and mRNAs. A Gaussian graphical model (GGM) [165] has been employed as the underlying model, and a convex optimization estimator is used for graph structure inference. The resulting edges in the inferred graph represent the candidate interactions. The second stage of MicroTarget is identifying direct interactions. We identify the microRNA direct targets by searching for matches to the seed region on all 3′-UTRs of the candidate targets returned by the first stage. The third stage is scoring and ranking the results with a set of features. These features are: site accessibility, conservation in related species, multiple binding sites per target mRNA, and context matching. Context matching is the sequence matching surrounding the seed region. We use the support vector regression (SVR) model to rank the predicted targets using this feature set.

MicroTarget have been applied to breast cancer expression data sets. The 3′-UTRs of the candidate targets are downloaded from the Ensembl database for human for prediction and

for other species for conservation scoring. To validate the results, the inferred targets are compared with the validated targets at the three largest experimentally confirmed target databases: miRTarBase v4.5 [56], MirWalK [31], and OncomiRdbB [69]. Also, we compare the result with other existing methods. Spearman rank correlation coefficient is computed between the scoring features to test their dependence. MicroTarget shows better performance than the existing methods. The main contributions of our research in this problem can be summarized as:

- We take advantage of expression profiles for microRNAs and mRNAs, as microRNA and its target have to be expressed in the same tissue to interact. We formulate the problem as regulatory network prediction problem from the expression data, which have not been proposed by any other method.
- Instead of filtering out the predicted targets with the targeting features as the current methods do, we estimate several individual scores with these features to rank microRNAs targets. We also add new features, that have not been considered by existing methods, based on the properties and overall complementary between microRNA and its target.
- A composite score was estimated for each target by SVR ranking model from the individual scores described above. The prediction of experimentally validated targets as the top ranked targets proves that scoring the targets with a combined features set plays an important role in identifying potential miRNA target genes.
- We evaluate the importance and correlation among microRNA targeting features. Spearman rank correlation coefficient is computed between the scoring features to evaluate their dependence.
- Our approach can provide a set of promising targets in specific tissue, based on the experssion data used, for each microRNA for farther experimental validation.

1.2 Identifying Conserved Protein Complexes

The second problem that was addressed is predicting conserved protein complexes across different species. An important reason behind the searching for conserved protein complexes

between species is that conservation implies functional significance. Sequence conserved proteins form the basis of comparative genomics. However, it is also critical to consider the conserved patterns of interactions among proteins themselves, which helps to transfer biological knowledge and function annotation at a higher level than comparing only protein sequences [26]. Identifying conserved protein complexes can aid in our understanding of evolutionary mechanisms of protein and protein interaction networks among species. Moreover, it is a fundamental step towards identifying the conserved mechanisms from model organisms to higher level organisms, such as cell cycle, DNA transcription, and protein translation. These mechanisms are considered the backbones for the living system [78].

Over the last decade, high-throughput experimental techniques have supported collection of a large number of protein-protein interactions (PPIs) for many species [50]. A popular representation of this data is a network. A node of the network represents a protein and an edge between two nodes represents an interaction between the two corresponding proteins. PPI network analysis across species provides awareness of similarities, differences, and the conserved components between species [135]. A central approach for this analysis relies on network alignment. PPI network alignment is a methodology that maps proteins and interactions in one organism with their counterparts in another organism. The thousands of interactions within each network as well as the complex homology relations among the species poses significant challenges for network alignment methods [116].

Network alignment is related to the subgraph isomorphism problem. This problem works on identifying the common subgraphs between two networks. The subgraph isomorphism problem is known to belong to the class of NP-hard problems [65]. For this reason, the techniques for solving this problem rely on heuristics and sometimes the use of additional data to guide the alignment process. The alignment may consist of one-to-one mapping between proteins of two networks (pairwise alignment), or many-to-many mapping among proteins of more than two species. Likewise, network alignment can be global or local alignment. Global network alignment (GNA) aims to find the best overall alignment between the input networks. The mapping in the global alignment should cover all of the input nodes. In local network alignment (LNA), the goal is to find local regions of isomorphism between the input networks. Each region is representing a mapping that is independent of others [111].

An important and difficult problem associated with GNA is their validation and the biological interpretation of the results. This difficulty arises from the noisy and incomplete nature of PPI network data [150]. LNA aims to find small but highly conserved subgraphs, irrespective

of the overall similarity among the networks. It outperforms GNA in learning novel protein functional knowledge and the biological quality of alignment. Another advantage supporting LNA is that it helps focus more on the reliable parts of the networks despite the noisy data. LNA is often used to detect conserved subnetworks, such as protein complexes, modules, and pathways from a set of species [36]. An overview of LNA methods is provided in Chapter 6.

1.2.1 Motivations and contributions

Despite the progress made by the research community in devising local network alignment strategies, these network alignment methods suffer from key drawbacks. They depend on protein sequence similarity to facilitate network alignment. Sequence similarity is only relevant to a subset of highly conserved proteins, which leave significant network regions poorly specified by sequence homology. Furthermore, with the high level of PPI data noise, the presence of several false negatives in PPIs leads to sparse alignment graphs if we consider only the direct connected pairs in both aligned networks. These issues cause approaches looking for highly connected subgraphs to fail to detect conserved complexes. Moreover, protein interactions occur through physical binding of small segments of proteins called domains, mostly these segments are conserved. Therefore, looking into protein interactions at the domain level can trim the limitations of the PPI data. In addition, Faisal et al. [36] showed that species co-evolution is more evident if we focus on the interacting domains that are responsible for PPIs.

In this dissertation, a new approach, called DONA (Domain-Oriented Network Aligner), is developed that addresses these issues by providing a general and effective framework for local network alignment. The proposed approach provides a way to account for both topological and homology information of the aligned networks, as well as employing DDIs data instead of just using the PPIs data. Our approach starts by constructing an alignment graph based on the protein-domain mapping, interactions found in the input networks and the known domains interactions for these proteins. Then using the Markov cluster algorithm (MCL) [34], it extracts the conserved sub-networks that form protein complexes or functional modules.

In a case study, we tested our approach in predicting a known conserved sub-network between a mouse and a rat PPI networks. DONA is able to identify this known conserved sub-network with more efficiency than other methods with precision and recall higher than the existing

methods. In a large data set of PPI networks for five different species, DONA performance has been compared to other methods in terms of its output overlapping with the known protein complexes and semantic similarity of the identified sub-networks, which computed with respect to the molecular function coherence of the aligned sub-networks. Our main contributions in this research can be summarized as:

- Rather than explicitly restrict its attention to align homologous proteins, DONA decomposes PPI networks in terms of their component domains and DDIs, and employs their conservation into a new strategy for building an alignment graph. Our results demonstrate that integrating domain interaction data significantly enhances the quality of the alignment.
- We propose a new scoring scheme to measure the conservation level between proteins and their interactions in the alignment graph.
- DONA uses a more scalable algorithm for searching the alignment graph, based on Markov clustering, comparing to the existing methods that mostly use seed-and-extend algorithm which proved to be inefficient for large PPI networks.
- We built an extensive testing data sets for identifying the conserved protein complexes between five different species. A collection set of conserved sub-networks among these species is identified. As currently there is no benchmark data set for conserved protein complexes in the literature, we hopes that this data set could be useful.

1.3 Dissertation Organization

The dissertation is organized as follow. Chapter 2 presents the biological background for microRNA biogenesis, mechanisms of gene regulation, and experimental method for identifying microRNA targets. Chapter 3 explains the principles of microRNA target prediction computationally and reviews the existing methods for microRNA target prediction. Chapter 4 represents the developed approach, MicroTarget, for predicting microRNA targets and its results.

The second problem in this dissertation, identifying conserved protein complexes, is represented in the next chapters. Chapter 5 shows the biological background for protein com-

plexes, protein-protein interactions, as well as domain-domain interactions. The computational methods for identifying conserved protein complexes using PPI network alignment are reviewed in Chapter 6. And chapter 7 shows the proposed method (DONA) for local network alignment to identify conserved proteins complexes among species and its results. Finally, the conclusion and future work are presented in Chapter 8.

Chapter 2

MicroRNA Target Prediction: Biological Background

The process by which DNA is transcribed into messenger RNA (mRNA) and an mRNA is translated into a protein represents the central dogma in molecular biology. The first step of gene expression is DNA transcription into RNA. The resulting RNA can be mRNA if the expressed gene is a protein coding gene. Otherwise, it is a non-coding RNA [132]. The second step is the translation of mRNA into a sequence of amino acids that composes a protein [125]. This chapter presents the biological background about both microRNA biogenesis, mechanism of action, and experimental identification of microRNA targets.

2.1 MicroRNA

Recent insight into molecular biology has revealed that about 80% of the human genome is transcribed into RNA, and out of the transcribed RNA about 2% is translated into protein [2]. This results in a large number of non-coding RNAs, called ncRNAs. A microRNA is a 19 to 24 nucleotidies single stranded RNA. The first identification of microRNA was the discovery of the let-7 microRNA in *C. elegans* [125]. A few years later, let-7 microRNA was also detected in humans, Drosophila, and other species [8]. The human genome encodes thousands of microRNA genes. There are two classes of microRNA genes: those that are generated from overlapping introns of protein coding transcripts and others that are encoded in the exons [47]. It is thought that microRNAs can have hundreds of targets. Most microRNAs in

plants show near perfect complementarity to their targets. This feature facilitates identifying microRNA-target interactions [47]. For microRNAs in animals, the target recognition is more complex because very few microRNA nucleotidies are perfectly complementary to the target. In the following only animal microRNAs are considered.

2.1.1 MicroRNA Biogenesis

MicroRNAs are transcribed as long hairpin RNA substrates of the DNA strand in the nucleus by RNA polymerase II. This process generates the primary RNA, which is called pri-microRNA. Then in the nucleus, a microprocessor complex recognizes the pri-microRNA double-stranded stem and the RNase III endonuclease, Drosha cleaves the pri-microRNA to create the precursor RNA stem-loop structure (pre-microRNA). Pre-microRNA is about 65 nucleotidies long and contains the microRNA sequence. Pre-microRNA is exported out of the nucleus (into the cytoplasm) by exportin-5 [51].

Once in the cytoplasm, a second RNase III enzyme, Dicer, recognizes and processes pre-microRNA to generate mature microRNA sequences. Mature microRNA is loaded into the RISC (RNA-induced silencing complex) to bind to its target [97]. After the microRNA binds to the target, the interaction with the mRNA is triggered. Figure 1 shows the biogenesis of microRNA and the binding to the target mRNA.

The transcription process for some microRNAs residing in introns (sometimes called intronic microRNAs) is slightly different. These intronic microRNAs are processed from the spliced introns of their host genes. In this case, introns are folded and make either long or short hairpin structures which, in the latter case, directly form the precursor microRNAs and prevent Drosha incorporation [130].

2.1.2 microRNA Mechanism of Action

The initial clues to microRNA regulation came from the observation that the lin-4 microRNA has some sequence complementary to conserved sites within the lin-14 mRNA, within a region of the 3′-UTR. A molecular genetic analysis had shown that these sites are required for the repression of lin-14 [155].

In animals, microRNAs bind to the RISC (RNA-induced silencing complex) and guide it

to cause either translational repression of mRNAs or site-specific endonucleolyitc cleavage in microRNA-mRNA pairs [63]. Whether the mRNA is cleaved or mRNA translation is inhibited depends on the complementarity of the microRNA and the mRNA. If there is a high degree of complementarity, the target mRNA is sequence-specifically cleaved by the RISC complex [8]. This case is more frequent in plants than in animals and induces direct mRNA degradation and cleavage. Usually after mRNA cleavage, the mcroiRNA remains whole and can regulate another target.

When microRNA-mRNA complementarity is not enough for cleavage mRNA translation will be repressed. The RISC complex contains at least one Argonaute protein (called Ago). The Argonaute protein family has several members. Whether the microRNAs guide mRNA cleavage or translation repression also depends on which specific Ago protein the microRNA is incorporated with [79]. Several studies suggested that microRNAs uses multiple mechanisms to cause translation repression of the target mRNA.

An mRNA can contain multiple sites (called target sites) for the same or different microRNAs. Accordingly, several different microRNAs can act together to repress the same gene. It seems that these multiple target sites work independently. The response to multiple microRNAs increases nearly the same as if the responses to the single microRNAs for their own were multiplied [126]. These microRNAs predominantly bind to sites in the 3′-untranslated region (3′-UTR) of their target mRNA. Nevertheless targeting can also occur in 5′-UTRs. Although a significant number of target sites have been found in 5′-UTRs, they seem to be less effective and are still less frequent than 3′-UTRs target sites. 5′-UTRs targeting is even rarer [22].

2.2 Experimental Identification of microRNA Targets

During the past decade, numerous efforts have been made to improve microRNA target identification and numerous mRNA targets have been experimentally validated.

Reporter assay

Reporter assay is one of the methods used for experimentally validating putative microRNA-mRNA interactions. It starts with cloning 3′-UTRs of genes of interest or 3′-UTR segments

containing the microRNA binding site into expression vectors that bear a reporter gene. Constructs that carry 3′-UTRs with the mutated target sites, to enable microRNA binding, are used as the negative control [102]. Finally, the transient transfection of the cells with reporters followed by measuring the reporter activity is performed. It has been observed that the expression of microRNAs in diseased tissues are different compared to that in normal ones. Luciferase reporters are costly and lack reproducibility between samples, which makes this approach unlikely to be scalable to genome-wide determination of microRNA-target sites [106].

Over-expression experiments

In these experiments, first microRNAs are transfected into the cell. Then the change of the expression level of transcripts is measured using mRNA expression profiling. The transcripts whose expressions significantly decrease after microRNA transfection are declared targets. This method has been extensively used to evaluate the sequence features proposed for target identification and validate the functional targets predicted by computational methods [25]. However when microRNA is over-expressed, it can saturate RISC complexes and displace other endogenous microRNAs, which in turn causes low affinity target sites to appear important.

Knock-down experiments

In these experiments, the expression of microRNA is inhibited using different strategies and the significantly up-regulated transcripts are treated as targets of the inhibited microRNA. One approach to inhibit the microRNA is to use synthetic microRNA targets. These synthetic targets are chemically modified, single stranded nucleic acids designed to specifically bind to the microRNA under the experiment [151].

MicroRNA Biotin-tagging

In this technique, cells are transfected with biotinylated microRNA duplexes and microRNA-mRNA complexes are captured from cell lysates using streptavidin beads [110]. The advantage of this technique is that it can specifically pull down mRNA targets of a single microRNA.

Proteome analysis

Another high throughput microRNA target identification method is proteome analysis. It relies on measuring the change of protein level in response to microRNA introduction. Proteome analysis employs stable isotope labeling with amino acids in cell culture followed by quantitative mass spectrometry. The limitations of this method is that some changes detected in protein levels result from an indirect microRNA regulation instead of a direct binding to the targeted transcripts. Comparing cell transcriptomes after microRNA overexpression or knockdown reference to the transcriptome of untreated cells also identifies the microRNA targets [86].

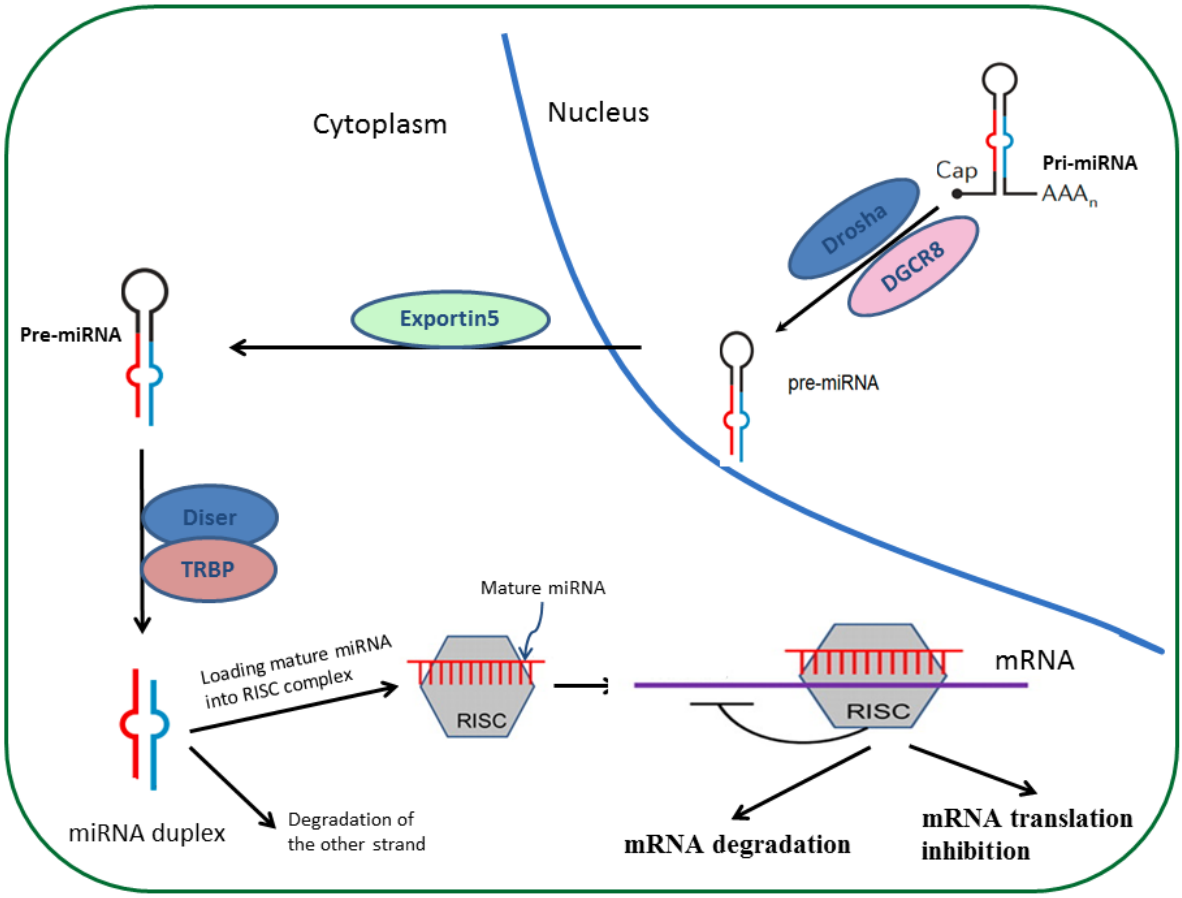

Figure 2.1: microRNA biogenesis and mechanism of action. It go under several processing steps before maturation to its active form. After processing, the mature microRNA incorporates into the RNA-induced silencing complex, then binds to the complementary sites in the 3′-UTR of their target genes. microRNA down-regulates the protein synthesis via translation repression or mRNA degradation [22].

Chapter 3

MicroRNA Target Prediction: Literature Review

Experimental identification of microRNA targets is difficult; therefore several computational tools have been proposed to predict microRNA targets. This chapter presents the principles of target prediction and existing computational prediction methods.

3.1 Principles of microRNA target recognition

The microRNA target prediction methods mostly exploit the principles identified using experimental methods to provide a genome wide prediction of the targets of all known microRNAs. These principles are microRNA seed pairing with the target site, conservation of mRNA target sites, the accessibility of the target site, and thermodynamic stability of the microRNA-target duplex. The next sections explain in detail these features.

3.1.1 Sequence complementary of seed binding site

At the 5′-end of the microRNA there is a region called the seed. It is centered on nucleotides 2 to 8. Watson-Crick pairing of the mRNA target site to this seed region is the most important factor for microRNA target prediction. The seed region of microRNAs is important because of the way the microRNA is bound by the silencing complex. For efficient pairing to be ideal,

RISC presents nucleotides 2 to 8 of the microRNA pre-organized in the shape of an A-form helix to the mRNA, while other configurations appear to result in lower affinity [118]. Most microRNA targets have a 7 nucleotides match. Some methods require perfect 8 nucleotide pairing to increase the specificity, where others search for 6 nucleotides seed pairing, yielding greater sensitivity. Strictly requiring seed pairing improves the performance of microRNA target prediction tools.

In addition to seed pairing, sequence complementary to the 3′-end of microRNAs also plays a role in target recognition [68]. It can supplement seed pairing and consequently improves binding specificity and affinity. Such 3′-end pairing mostly take place at microRNA nucleotides 13 to 17 with a length of 3 or 4. The pairing between the mRNA and 3′-end region of microRNAs can compensate for a mismatch in the seed region. However, 3′-end pairing sites are rare and only emerge when a specific member of a microRNA family is required for regulation. That is because most microRNAs within a family have the same seed region but differ in their remaining sequence [109].

Not only the sequence complementary of the target site defines whether an mRNA is a target of the microRNA; other factors also can have an effect. For instance, the position of the site influences the efficacy of targeting. In long UTRs, the binding sites should not fall in the middle of the 3′-UTR, because at this location the site might be less accessible to the silencing complex. Moreover, high local AU content seems to increase the site accessibility because of the weaker mRNA secondary structure [48]. Additionally, the proximity to binding sites of co-expressed microRNAs can also enhance site efficacy.

3.1.2 Site accessibility

For binding to the microRNA, the target site has to be accessible, which means it has to be opened and must not interact with other sites within the mRNA, at least in the region corresponding to the seed. Often, it is the accessibility of the 3′-UTR that must be assessed. When microRNA is assembled into the RNA-induced silencing complex (RISC) and the mRNA seed binding sites are in the active state, the microRNA-mRNA pairing is likely. However, it is more favorable when short regions with a length of approximately 15 nucleotides upstream and downstream of the target site that are opened as well [92]. Two factors have to be considered when assessing site accessibility: first, this opening energy cost estimated as $\triangle G_{open}$, and second, the free energy of the microRNA-target duplex $\triangle G_{duplex}$.

The total free energy change equals the difference between $\triangle G_{duplex}$ and $\triangle G_{open}$ and represents a score for the accessibility of the target site and the probability for a microRNA-target interaction [127].

3.1.3 Conservation

The mRNA binding sites that are conserved across species are more likely to be biologically functional and have more potential for being microRNA target sites. The use of conserved site sequences can significantly reduce the false positive rate of a prediction tool. Sites are regarded as conserved if they are retained at orthologous locations in multiple genomes, which means they have to appear exactly at the same position in the alignment of the 3′-UTR sequences [44]. Also, sites can be regarded as conserved if they just can be found somewhere in the sequences but not in the same aligned positions. When the site is missing or has changed in only one of the multiple species that are considered, the sites can be regarded as poorly conserved [48].

3.1.4 Thermodynamic stability

Another way to identify microRNA targets is the consideration of thermodynamic stability of the microRNA-target duplex. It is an energetically more favorable state when two RNA complementary strands are hybridized. The lower the free energy of two strands, the more energy is needed to disrupt this duplex formation. Therefore, an RNA duplex is in a thermodynamic stable state (means the binding of the microRNA to the mRNA is stronger) when the free energy is low [152]. In other words, a microRNA has a higher affinity to bind to an mRNA when the following duplex has a low free energy.

3.2 Computational target prediction methods

Computational methods for microRNA-targets prediction can be divided into three categories: rule-based, machine learning, and model-based methods. This section outlines the popular microRNA target prediction methods in each category.

3.2.1 Rule-based methods

Rule-based methods rely on a set of rules to be satisfied by the 3′-UTR for its gene to be a target. They are testing the rules according to a particular order, and the testing rules are essentially filtering steps. Therefore, the order of testing the set of rules affects the performance.

TargetScan [82] is among the most popular target prediction methods. First, microRNAs conserved in multiple organisms and a set of candidate 3′-UTR sequences from these organisms are prepared. Then, it searches the 3′-UTR for a seed match. It sets $match = 1$ if there is a perfect seed match or disqualifies the 3′-UTR ($match = 0$) otherwise. Then a score is computed based on the seed match and the site accessibility. A 3′-UTR is predicted to be a target if its score is higher than a threshold. The threshold is chosen based on the organism. Its false positive rate was estimated as 30% for mammalian microRNA targets. TargetScan also provides a wide range of information about microRNA and target transcript sequences and has been frequently updated. TargetScan was updated to TargetScanS [45], which requires a shorter seed match (6 nucleotides instead of 7) and does not consider site accessibility. Results show that the false positive rate is reduced to 22% compared to TargetScan.

Rehmsmeier et al. [124] proposed RNAhybrid to utilize seed match (also supporting user defined seed matches), free energy, and p-value of the estimated free energy as the prediction features. The method starts with finding all possible seed binding sites as candidate targets. Then, a 3′-UTR is predicted as a target if both the minimum free energy and its p-value are less than user defined cutoffs. RNAhybrid modified the RNA secondary structure prediction tool RNAfold [90] for estimating cite accessibility.

John et al. [63] proposed miRanda, which uses three steps to identify the target. First, the microRNA sequences are scanned against the 3′-UTRs sequence. It considers matching along the entire microRNA sequence. Next, the free energy of each microRNA target pair score is calculated. Targets that have a free energy score below the threshold are then passed to the conservation step. A predicted target can be ranked high in the results by either obtaining a high individual score from the match and free energy or by having multiple predicted sites. The authors appy miRanda to predict human microRNA targets. 2000 putative human microRNA targets were identified, suggesting that fewer than 10% of the human genes are regulated by microRNAs.

Dweep et al. [31] proposed MiRWalk, which relies on identifying multiple binding sites between the microRNA and the 3′-UTR. It searches the complete sequence of the 3′-UTRs starting with a 7 nucleotide seed from positions 1 and 2 of the microRNA sequences. As soon as it identifies a perfect match, it extends the length of the microRNA seed until a mismatch arises. It returns all possible hits with 7 or longer matches. Then the probability distribution of the longest binding sites is calculated using a Poisson distribution. Afterwards, miRWalk compares the identified microRNA binding sites with the results obtained from 8 different target prediction programs. It also performs an automated text mining search in the titles and the abstracts of PubMed articles, using curated dictionaries, to find experimentally validated targets. A total of 1360 unique PubMed article identifiers (PMID) were found have at least one miRNA name present in their titles and/or abstracts. This algorithm discovers 1870 positive miRNA-target and 61 negative miRNA-target pairs. Finally, predicted and validated information is stored in a relational database.

Kertesz et al. [67] proposed a target prediction method called PITA that incorporates the role of target site accessibility. PITA is based on the experimental observation that a strong secondary structure formed by 3′-UTR will prevent the binding of miRNA. It defines a thermodynamic model for microRNA target interaction and calls it the accessibility energy. First, the seed binding sites are searched. Then a score for each candidate site is estimated. If $\triangle E_{duplex}$ is the free energy gained by binding the microRNA to the target, and $\triangle E_{open}$ is the free energy lost by unpairing the target site nucleotides, then a score is defined as the energy gained by transitioning from the state in which the the target strands are unbound and the state in which the microRNA binds the target as:

$$\triangle E = \triangle E_{duplex} - \triangle E_{open}.$$

The total score for all the binding sites n for each microRNAtarget pair is estimated as:

$$score = \log(\sum_{i=1}^{n} e^{\triangle E_i}).$$

Kiriakidou et al. [74] modified PITA into DIANA-microT to predict human microRNA targets. First, DIANA-microT retrieves orthologous human and mouse 3′-UTRs from human mRNA and 94 conserved microRNAs in human and mouse. Then, it filters the seed binding sites by a free energy threshold.

3.2.2 Machine Learning Methods

Instead of using a set of rules to filter the targets, Kim et al. [70] proposed MiTarget, which collects biologically relevant information from the literature and designs features that imply the manner of microRNA targeting. To build the training data set, 152 positive targets and 83 negative targets are collected from the literature. It trains a support vector machine (SVM) model based on the training data and the feature vector. It predicted significant functions of some human microRNA, such as miR-1, miR-124a, and miR-373, using Gene Ontology analysis.

Lui et al. [89] proposed SVMicro, another SVM based target prediction method. SVMicro uses two stages. First, a data set for the SVM is constructed, which consists of the 3′-UTR of targets and the microRNA sequences of 314 experimentally validated positive target and 186 negative target sequences. Second, 46 features are designed, based on the data and existing knowledge of microRNA binding to the target. Then, it uses SVM to predict the targets.

Betel et al. [9] proposed MirSVR, which uses miRanda to identify candidate target sites and support vector regression (SVR) to score the candidate target. It computes a score that represents the strength of microRNA-target pairing and trains the SVR on nine microRNA experiments performed on HeLa cells and a number of other features, such as the position of the target site within the 3′-UTR. MiRSVR analysis shows that some targets with non-conserved, imperfect complementary seed match have significantly high scores. It also shows that approximately 7% of the target sites are non-canonical. Its results show that the area under the curve of ROC analysis (AUC) equal 0.63. Although MiRSVR claims that it achieved its strength from the SVR classifier, it did not gain any performance improvement when replaying their regression classifier with an SVM type classifier.

Ding et al. [29] proposed TarPmiR, which applied a machine learning approach to the CLASH (crosslinking ligation and sequencing of hybrids) data to identify seven new features of microRNA target sites. They identified seven new features together with six conventional features of microRNA target sites from tha CLASH data set. Then, they apply a random forest based algorithm to integrate these features to predict microRNA target sites.

3.2.3 Model-Based Methods

Krek et al. [77] presented a hidden Markov model to predict microRNA targets, called PicTar. PicTar searches for the seed matches of each microRNA in the 3′-UTRs. Then, it checks whether perfect seed matches are conserved or not in the species under consideration. If perfect matches are conserved, PicTar further checks whether optimal microRNA target binding free energy is below a cutoff value. Perfect matches that pass these steps are called anchors. The 3′-UTRs containing multiple anchors are used for the training data set. To perform the prediction, a hidden Markov model is built to model the fact that several microRNAs can act together to repress the same target. PicTar experimentally validated 7 out of 13 predicted targets and 8 out of 9 previously known targets, but still its false positive rate was estimated to be around 30%.

Huang et. al. [59] proposed GenMiR++, which uses a Bayesian model to infer a probability for each candidate mRNA of being a real target. First, it uses TargetScanS prediction on the human genome to predict the set of all possible targets. Second, it uses microRNA and mRNA expression profiles to score the targets. The GenMiR++ calculates scores by attempting to reproduce the mRNA profile by a weighted combination of the genome wide average normalized expression profile and the negatively weighted profiles of a subset of the microRNAs. the GenMiR++ model is very complex and computationally expensive. It performed an experimental validation for the predicted high scoring targets of let-7b. A list of 34 targets predicted by TargetScanS was considered as candidates, among which 12 were predicted by GenMiR++ to have the highest scores. The experiment results showed that 5 out of 12 targets were down-regulated.

Naifang et al. [105] modify GenMir++ to reduce the computing time. They define Bayesian prior probability and solve its posterior probability by Markov Chain Monte Carlo (MCMC) techniques. A major drawback of this method is that its posterior is not suitable for data where the number of variables are higher than the number of samples.

Khorshid et al. [68] proposed MIRZA. Using a set of mRNAs cross linked in Ago-CLIP (cross-linking immunoprecipitation) experiments and a set of microRNAs, MIRZA models the microRNA-mRNA hybrid structures. It infers the model parameters by maximizing the binding probability of mRNA sequences in Ago-CLIP data. Dongen et al. [146] proposed Sylamer. Let N denote the number of genes ranked based on their expression levels in a miRNA over-expression experiment. Let M_i denote the number of genes whose expression

levels is less than an incremental cut-off value. Sylamer computes a P-value using a hypergeometric test to identify if seed matches are significantly over-represented in a set of genes compared to seed matches presented in N genes. Then, it generates a curve using computed P-values and searches for the occurrence of a peak at the top of the rank gene list that implies down-regulated targets of the over-expression miRNA.

Despite the preceding methods, the existing methods using sequence data alone still have poor performance in term of specificity and sensitivity. Unlike sequence data, expression data are condition specific and dynamic and so provide useful clues about the set of active microRNAs and mRNAs. These facts motivated us to incorporate tissue expression data for mRNA and microRNA to improve the target prediction. Chapter 4 presents our proposed approach for microRNA target prediction using sequence and gene expression data.

Chapter 4

MicroTarget: microRNA Target Prediction Approach

MicroRNAs are known to play an essential role in gene regulation in plants and animals. The standard method for understanding microRNA-gene interactions is randomized controlled perturbation experiments. These experiments are costly and time consuming. Therefore, using computational methods is necessary. Currently, several computational methods have been developed to discover microRNA target genes. These methods are explained in Chapter 3. However, these methods have limitations based on the features that are used for prediction. The commonly used features are complementarity to the seed region of the microRNA, site accessibility, and evolutionary conservation. Unfortunately, not all microRNA target sites are conserved or adhere to exact seed complementary, and relying on site accessibility does not guarantee that the interaction exists. The study of regulatory interactions composed of the same tissue expression data for microRNAs and mRNAs is necessary to understand the specificity of regulation and function.

My proposed approach for microRNA targets prediction is a machine learning technique that addresses the question of whether there is an interaction between a microRNA and a particular mRNA or not and ranks each target mRNA. The approach emphasizes the sensitivity in searching for all potential targets and the specificity in assessing each predicted target. We developed the MicroTarget approach to predict a microRNA-gene regulatory network using heterogeneous data sources, especially gene and microRNA expression data. First, MicroTarget uses expression data to learn a candidate target set for each microRNA.

Then, it uses sequence data to provide evidence of direct interactions. MicroTarget scores and ranks the predicted targets based on a set of features. To systematically explain my approach for predicting microRNA targets, we first provide the formulation of the prediction problem. This chapter explains the proposed approach and its results.

4.1 Preliminaries and Problem Definition

To predict microRNA targets computationally, various data are required, including nucleotide sequences of microRNAs, mRNA 3′-UTR sequences, sequence conservation, and expression data. For a given microRNA sequence of length m, let $W = w_1, w_2, \ldots, w_m$ represents the nucleotide sequence of the microRNA, where $w_i \in S$ denotes the nucleotide at the i^{th} position, and $S = \{A, C, G, U\}$. For testing whether the 3′-UTR of an mRNA is a potential target, the 3′-UTR sequence of the mRNA is retrieved and denoted as $R = r_1, r_2, ..., r_n$, where $r_k \in S$ represents the nucleotide at the k^{th} position of the 3′-UTR. The seed sequence of a microRNA is defined as the first 2 through 8 nucleotides starting at the 5′-end and counting toward the 3′-end.

Let V represent a feature vector derived from R and W, with v_l denoting the value of l^{th} feature. One way for target prediction is to decide whether mRNA is a target or not based on the feature vector V. However, relying on sequence features to predict the targets is not sufficient since effective regulation of a target requires that the microRNA and the target be located in the same cellular compartment [107]. Therefore, adding expression data is necessary to understand microRNA target regulation.

The proposed approach takes mRNA and microRNA expression profiles and infers the candidate target set for each microRNA. The problem of inferring the regulation between microRNAs and mRNAs using expression data is formulated as a network structure learning problem. Several concepts and notations are used throughout the dissertation for adding the expression data for the prediction.

Let X be a t-dimensional vector and $X_1, X_2, \ldots, X_t$ denote the t variables, where t is the number of microRNA and mRNA, and let X_k be the vector of expression levels (samples) for the k^{th} variable, $k = 1, 2, 3, \ldots, n$, where n is the number of samples. Two variables X_1 and X_2 are conditionally independent given X_3 if $f(X_1|X_2, X_3) = f(X_1|X_3)$, where $f(X_1|X_3)$ is the conditional density of X_1 given X_3 and $f(X_1|X_2, X_3)$ is the conditional density of X_1

given X_2 and X_3. Conditional independence is a fundamental property in Gaussian graphical models.

A Gaussian graphical model (GSM) is a graph representation of the random variables. The GGM was introduced by Dempster [165] under the name of covariance selection models. It is a graphical interaction model for the multivariate normal distribution; two nodes are connected by an edge if the corresponding variables are conditionally dependent. In other words, a GGM can be defined as a family of multivariate normal distributions for X that satisfy the conditional independence statements implied by the graph. It is determined by assuming conditional independence of selected pairs of variables given all the remaining variables. Precisely, if $G = (N, E)$ is a graph and X is a random vector taking values in R^N, then the GGM for X on G is given by assuming that X follows a multivariate normal distribution that satisfies the pairwise Markov property [7]. The GGM $t \times t$ covariance matrix is estimated as

$$S = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{\mu})(x_i - \bar{\mu})^T \tag{4.1}$$

where

$$\bar{\mu} = \frac{1}{n}\sum_{i=1}^{n}(x_i).$$

Banerjee et al. [7] prove that using the inverse covariance matrix (precision matrix) in inferring the graph structure is more efficient than using the covariance matrix if the underlying model is GGM. The variables conditional independence in GGM is reflected in the zero entries of the precision matrix [43]. If the number of samples is fewer than the number of variables, as it is in our data set, the covariance matrix will be singular and therefore cannot be inverted [163]. In this case, we need to find a method for estimating the precision matrix directly instead of inverting the covariance matrix. Each entry θ_{ij} in the precision matrix $\Theta = (\theta_{ij})_{1 \leq ij \leq t}$ corresponds to the relation between two variables i and j, where $\theta_{ij} = 0$ if and only if the x_i and x_j are conditionally independent.

Our goal for target prediction is equivalent to identifying the precision matrix from the expression data that can predict if a mRNA is a target or not. However, some regulation that predicted only using expression data can be indirect. Therefore, using sequence mapping between microRNA W and mRNA R is required to confirm the direct interaction.

4.2 The Proposed Approach

This section explains the proposed approach MicroTarget; its framework is shown in Figure 4.1. First, MicroTarget takes mRNA and microRNA expression profiles and infers the candidate target set for each microRNA. The problem of inferring the regulation between microRNAs and mRNAs is formulated as a network structure learning problem. The problem input is a matrix of microRNA and mRNA expression values. The proposed approach predicts an undirected graph structure corresponding to the conditional dependence among the microRNAs and mRNAs. It employs a Gaussian graphical model as the underlying model and a convex optimization estimator for graph structure inference. The resulting edges in the inferred graph represent the candidate interactions.

The second stage of MicroTarget is identifying direct interactions. We identify the microRNA direct targets by searching for matches to the seed region in all $3'$-UTRs of the candidate targets returned by the first stage. The third stage of MicroTarget is scoring and ranking the result targets from stage two with a set of features. These features are: site accessibility, conservation in related species, number of binding sites per target mRNA, and context matching. Context matching is sequence matching surrounding the seed region. Then the predicted target is ranked based on the scores estimated from these features. The support vector regression (SVR) model is used to rank the predicted targets from the feature set.

4.2.1 MiRLasso for graph structure learning

For the first stage of MicroTarget, we propose miRLasso algorithm, which takes the expression data samples as an input matrix and outputs a matrix that represents a graph structure. The graph encodes the conditional dependencies between the microRNAs and mRNAs. The algorithm assumes that the samples are normally distributed, and the GGM is used as the underlying model [43].

Let a graph $G = (V, E)$ represent the regulatory network between the microRNAs and mRNAs. The vertices of the graph represent the microRNAs and mRNAs (variables). Let $X = (X_1, ..., X_t)$ be a variable set, which can be represented by an undirected graph $G = (V, E)$. The vertex set is $V := X_1, ..., X_t$. The edge set E consists of vertex pairs (i, j) that are joined by an edge. If X_i is independent of X_j given the other variables, then $(i, j) \notin E$. For illustration, Figure 4.2 illustrates a precision matrix for 6 variables and its corresponding

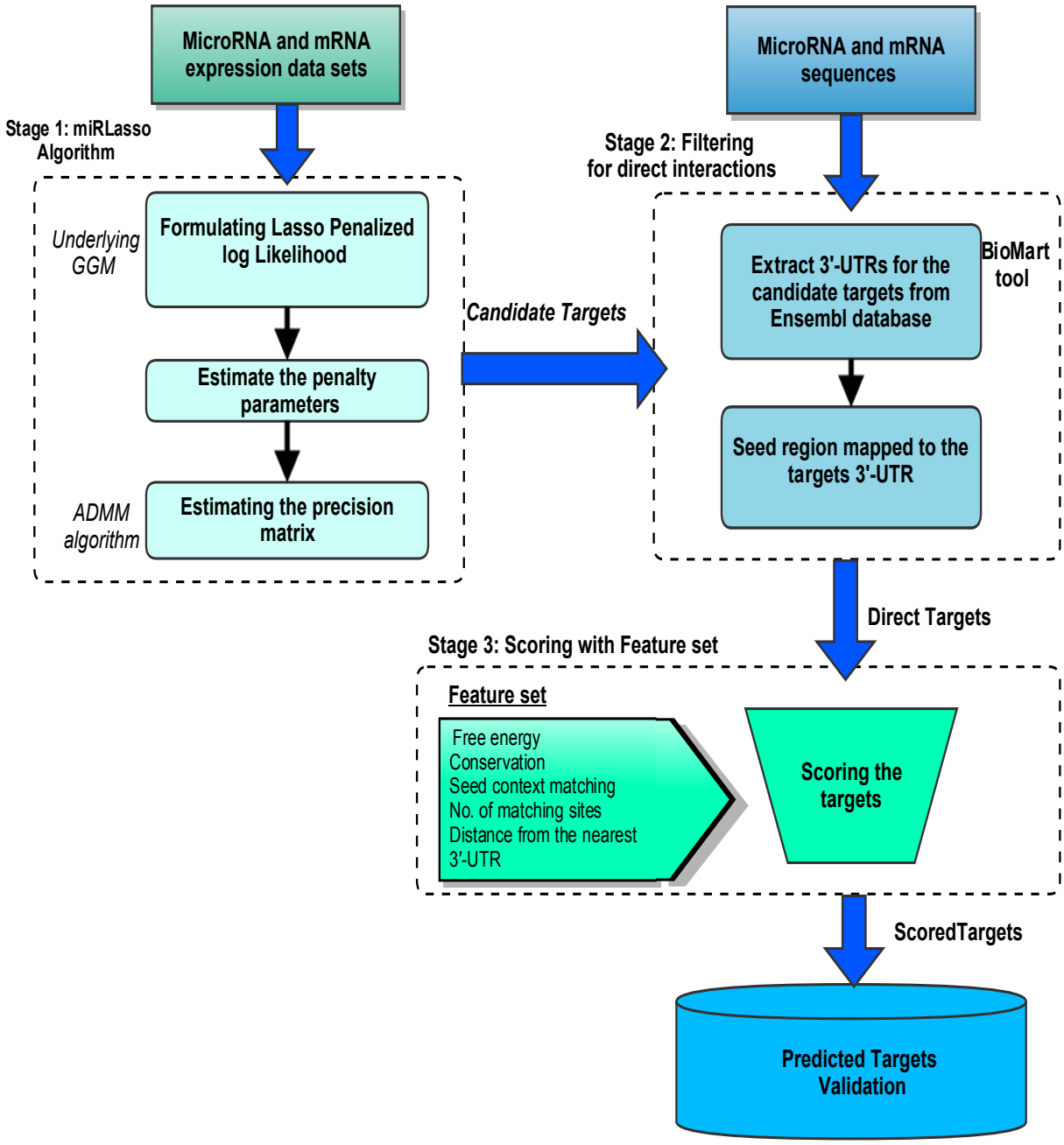

Figure 4.1: The conceptual view of MicroTarget includes using microRNA and mRNA expression data to infer the candidate targets for each microRNA, using sequence data to get the direct microRNA-targets interactions, and finally scoring and validate results.

$$\Theta = \begin{pmatrix} \theta_{1,1} & \theta_{1,2} & \theta_{1,3} & 0 & 0 & 0 \\ \theta_{2,1} & \theta_{2,2} & 0 & \theta_{2,4} & \theta_{2,5} & \theta_{2,6} \\ \theta_{3,1} & 0 & \theta_{3,3} & 0 & \theta_{3,5} & 0 \\ 0 & \theta_{4,2} & 0 & \theta_{4,4} & 0 & 0 \\ 0 & \theta_{5,2} & \theta_{5,3} & 0 & \theta_{5,5} & \theta_{5,6} \\ 0 & \theta_{6,2} & 0 & 0 & \theta_{6,5} & \theta_{6,6} \end{pmatrix}$$

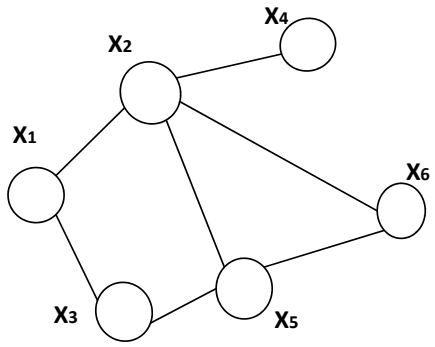

Figure 4.2: An example of the precision matrix and its corresponding graph structure

undirected graph structure. The GGM that describes the conditional dependence among the parameters is encoded by the sparsity of the precision matrix Θ.

Graph structure learning means estimating the zero and nonzero entries in the precision matrix. The precision matrix Θ is estimated by maximizing the log likelihood. The Gaussian log likelihood takes the form

$$l(\Theta) = \frac{n}{2}(\log\det(\Theta) - trace(S\Theta)). \tag{4.2}$$

Maximizing this equation with respect to Θ yields the maximum likelihood estimate for the precision matrix. If the number of variables exceeds the number of observations, all entries in the estimated precision matrix will be non-zero. This results in a dense graph. For the estimated precision matrix to be sparse, as there are few samples compared to the number of the parameters (microRNAs and mRNAs), the introduction of regularization is required. A penalty function $g(\Theta)$ is added to the maximization in Equation (4.2) to encourage sparsity of the graph, using the Lasso penalty [21]. Regularization with the l_1 norm seems to be pervasive throughout many fields of mathematics. In statistics, Lasso is an example of the application of an l_1 regularization in linear regression. The Lasso l_1 penalty comes from a Laplace prior [43].

MicroTarget utilizes a graphical Lasso penalty that is inspired by the joint graphical Lasso from [28]. If $\theta_{i,j}$ is the Θ matrix entry at the i^{th} row and the j^{th} column, and Z refers to a previously estimated Θ then, the penalty function $g(\Theta)$ is

$$g(\Theta) = \lambda_1 \sum_{i \neq j}^{t} |\theta_{i,j}| + \lambda_2 \sum_{i \neq j}^{t} |\theta_{i,j} - Z_{i,j}|. \tag{4.3}$$

The first penalty term, regularized by λ_1, assigns a cost to matrices with large absolute values, thus effectively enforcing the sparsity. The second penalty term, regularized by λ_2, encourages the accuracy of the resulting matrix by penalizing the difference between the current learned matrix and the previous one.

Estimating the precision matrix can be formulated as a convex optimization problem, which is solved by maximizing the penalized log likelihood with respect to Θ:

$$maximize_{\Theta} \left\{ \frac{n}{2} (\log \det(\Theta) - trace(S\Theta)) - g(\Theta) \right\}. \tag{4.4}$$

For computational implementation, the precision matrix is estimated by minimizing the negative penalized log likelihood. The optimization problem is solved using the alternating direction method of multipliers (ADMM) [15]. ADMM is a form of augmented Lagrangian algorithm that is well suited to dealing with structured problems. It decomposes the original problem into two subproblems, solves them sequentially, and updates its dual variables at each iteration. ADMM attracted renewed attention recently due to its applicability to various machine learning problems. In particular:

- ADMM takes advantage of the structure of the problems that involve optimizing sums of fairly simple but sometimes nonsmooth convex functions.
- In most cases, ADMM is computationally efficient overall. In particular, the total number of iterations of the ADMM is considerably fewer than the number of iterations of most optimization solver algorithms, like the dual coordinate descent algorithm.
- It is relatively easy to implement the ADMM in a distributed memory and parallel manner. This property is important for high dimensional data sets problems in which the entire data set may not fit readily into the memory of a single processor.

ADMM is similar to dual ascent. It consists of an x-minimization step, a z-minimization step, and a dual variable update step. The step size of the dual variable update is equal to the augmented Lagrangian parameter.

Precision Matrix Estimation with ADMM

ADMM introduces a set of auxiliary variables denoted as Z and U, where Z corresponds to the previous Θ and U is the dual variable. This allows us to minimize Equation (4.4) with respect to Θ and Z in an iterative fashion. Consequently, Equation (4.4) can be reformulated as the following constrained minimization problem:

$$\begin{aligned} &\underset{\Theta}{\text{minimize}} \quad -\frac{\mathrm{n}}{2}(\log\det(\Theta) - \mathrm{trace}(\mathrm{S}\Theta)) + \mathrm{g}(\mathrm{Z}), \\ &\text{subject to} \quad \Theta = Z. \end{aligned} \tag{4.5}$$

We replace Θ by Z in the penalty terms. As a result, Θ terms are involved only in the likelihood component of Equation (4.4), while Z terms are involved in the penalty components. The use of the ADMM algorithm requires the formulation of the augmented Lagrangian corresponding to the likelihood an d penalty equations as:

$$\mathcal{L}_\rho(\Theta, Z, U) = \left\{ -\frac{n}{2}(\log\det(\Theta) - trace(S\Theta)) + g(Z) - \frac{\rho}{2}||\Theta - Z + U||_F^2 \right\}. \tag{4.6}$$

The precision matrix estimator minimizes Equation (4.6) with respect to the variables, Θ, Z, and U. This allows us to decouple the Lagrangian in such a manner that the individual structure associated with variables Θ and Z can be exploited. For $k = 1, ..., R$ (R maximum number of iterations) iterations, Θ_k is the estimate of Θ in the k^{th} iteration. The same notation goes for Z_k and U_k.
The estimator initializes $\Theta_1 = I$ and $Z = U = 0$, where I is the $t \times t$ identity matrix.

At each iteration k the algorithm performs three steps, as follows
Step 1: Update Θ.
At this step, we treat Z_{k-1} and U_{k-1} as constants. As a result, minimizing Equation (4.6) with respect to Θ corresponds to

$$\Theta_k \leftarrow \underset{\Theta}{\text{argmin}} \left\{ -n/2(\log\det(\Theta) - trace(S\Theta)) - \rho/2||\Theta - Z_{k-1} + U_{k-1}||_F^2 \right\}. \tag{4.7}$$

If ρ is set to zero, only the log likelihood terms will be left in Equation (4.6). That results in an unsparse Θ. Setting ρ to be a positive constant implies that Θ will be a compromise between minimizing the log likelihood and remaining in the proximity of Z_{k-1}, the previous

Θ. Let VDV^T denote the singular value decomposition of $S - \rho/2Z_{k-1} + \rho/2U_{k-1}$, the solution is given at [156] by $V\check{D}V^T$, where $\check{D}$ is the diagonal matrix with diagonal entries

$$\check{D}_{ll} = \frac{n}{2\rho}(-D_{ll} + (D_{ll}^2 + 4\rho/n)^{1/2}).$$

Step 2: Update Z
Update Z by minimize the following equation with respect to Z:

$$Z_k \leftarrow argmin_Z\Big\{\frac{\rho}{2}||Z - (\Theta_k + U_{k-1})||_F^2 + g(Z)\Big\}. \tag{4.8}$$

Solving Equation (4.8) will depend on the form of the penalty. Let

$$A = \Theta_k + U_{k-1}. \tag{4.9}$$

By substituting Equation (4.9) into Equation (4.8), it can be written as

$$Z_k \leftarrow argmin_Z\Big\{\frac{\rho}{2}||Z - A_k||_F^2 + g(Z)\Big\}. \tag{4.10}$$

Given the penalty in Equation (4.3), then Equation (4.10) takes the form

$$Z_k \leftarrow argmin_Z\Big\{\frac{\rho}{2}||Z - A_k||_F^2 + \lambda_1 \sum_{i\neq j}^{t} |Z_{i,j}| + \lambda_2 \sum_{i\neq j}^{t} |Z_{i,j} - (Z_{i,j})_{-1}|\Big\}, \tag{4.11}$$

where $Z_{i,j}$ is an element in Z matrix at the k iteration, and $(Z_{i,j})_{-1}$ is the corresponding element at the $k-1$ iteration. This equation is separable with respect to each pair of the elements (i, j) in the matrix. Then Equation(4.11) can be rewritten as

$$Z_i \leftarrow argmin_Z\Big\{\frac{\rho}{2}\sum(Z_{ij} - A_{ij})^2 + \lambda_1 \sum_{i\neq j}^{t} |Z_{i,j}| + \lambda_2 \sum_{i\neq j}^{t} |Z_{i,j} - (Z_{i,j})_{-1}|\Big\}. \tag{4.12}$$

Step 3: Update U
This corresponds to an update of U_i as follows:

$$U_k = U_{k-1} + \Theta_k - Z_k$$

The final Θ that is estimated from this algorithm is the estimate of the precision matrix.

Algorithm 1 provides pseudocode for miRLasso optimization. The parameters λ_1, λ_2, and ρ are estimated using the same method as in [28]. The parameter ρ is estimated using cross-validation, and λ_1 and λ_2 are estimated using Akaike information criterion (AIC). The algorithm is guaranteed to converge to a global optimum. The global convergence of ADMM has been established by He et al [54]. The algorithm iterates until convergence is reached. To guarantee convergence, we require two constraints. First, the result Θ should satisfy the constraint $\Theta_k = Z_k$. The second constraint refers to the minimization of the augmented Lagrangian. For the first constraint, we check $||\Theta_k - Z_k||_2^2$ at each iteration. Step 3 of miRLasso ensures that the Z_k are always dual feasible. It checks $||Z_k - Z_{k-1}||_2^2$ to verify dual feasibility in Z_k variables. The algorithm converges when $||\Theta_k - Z_k||_2^2 \leq \tau_1$ and $||Z_k - Z_{k-1}||_2^2 \leq \tau_2$, where τ_1 and τ_2 are the convergence thresholds. Here, miRLasso uses a small threshold, as in [54], to ensure convergence.

Let Θ_e be the estimated precision matrix. Recall that we define the estimated graph $G = (V, E)$ where $(i, j) \in E$ if $\theta_{ij} = 0$. Theoretically, it is possible that miRLasso delivers some precision matrix estimates with very small nonzero values. To get the graph structure, the estimated precision matrix is threshold to get the final sparse precision matrix Θ_f.

For Θ_e estimated from miRLasso ADMM iterations such that the smallest nonzero element of Θ satisfies

$$\Theta := \min_{i,j \in p} |\Theta_{ij}| \leq ||\Theta||_1 \sqrt{\frac{\log p}{n}}.$$

For every element in Θ_e, to get Θ_f let:

$$\theta_{ij} = \begin{cases} \theta_{ij} & \text{if } |\Theta_{ij}| > ||\Theta||_1 \sqrt{\frac{\log p}{n}}; \\ 0 & \text{if } |\Theta_{ij}| \leq ||\Theta||_1 \sqrt{\frac{\log p}{n}}. \end{cases}$$

Under these conditions, there exists a constant such that the above threshold estimator achieves exact recovery. More discussions on this constant and its estimation can be found in [166]. Since the algorithm requires an eigen decomposition for every S update, and the Z and Θ updates are constant time operations, the run time complexity is $O(mn^3)$, where m is the number of iterations and n is the size of the data set observations.

4.2.2 Learning microRNA Direct Targets

The results from the miRLasso algorithm represent the candidate microRNA-target interactions. These results have been used as the input for Stage 2. The main idea of Stage 2 is to filter out the candidate interactions by deleting the indirect ones. The binding of a microRNA to an mRNA induces a direct regulation for the corresponding gene. A microRNA binds to a specific site within the 3′-UTR region of the mRNA sequence. It can bind to multiple sites in the same 3′-UTR. The binding of a microRNA to a gene is weak at the central region and strong at the seed region. Therefore, the seed region (positions from 2 through 8 from the 5′-end of the microRNA) is used for finding direct interactions. Genes that do not have seed binding sites will have zero probability of being direct targets. The matching between the seed region and the binding site at the 3′-end of the mRNA is necessary for defining the direct interactions. However, in some cases, an exact matching is not required for a functional interaction and a non-canonical pairing with G:U wobbles or mismatches may be acceptable [51]. Therefore, our algorithm allows for non-canonical base pairing.

The output of the miRLasso algorithm is taken as the input to the filtering stage. This stage starts with finding the microRNA seed region. Then, it search along the 3′-UTR sequence of each candidate target to find the segments with complementarity to the seed region. Such a segment is called a seed binding site. Given that more than one binding site can be found in the same 3′-UTR, we continue searching after finding the first binding site. The number of binding sites in the same 3′-UTR is denoted by B_{ij}, where i is the target gene and j is the microRNA. If $B_{ij} \geq 1$, then the target i is a direct target for the microRNA j. B_{ij} is also used later in the scoring. Picking $B_{ij} \geq 1$ is to ensure that there is a least one binding site between the candidate target and the microRNA. For each microRNA, the candidate targets with zero binding sites are removed from its target set. Removing these targets corresponds to removing edges from the inferred graph with the first stage of MicroTarget. The result graph after filtering the direct interactions is the predicted microRNA-gene regulatory network.

The resulting graph $H = (V_h, E_h)$ is the inferred microRNA-mRNA regulatory network. Next, MicroTarget scores and ranks each predicted microRNA-mRNA regulatory interaction.

Algorithm 1 My implementation of the ADMM algorithm to solve the precision matrix estimation problem. The final Θ that results from this algorithm is the miRLasso estimate for the precision matrix.

Input: Initialize: $\Theta = I$, $Z = 0$ and $U = 0$

Output: $p \times p$ precision matrix Θ over number of variables p

1: Select the parameters ρ, λ_1 and λ_2.

2: **for** $k = 1, 2, 3, ...$ until convergence **do**

3: i Update Θ as the minimization (with respect to Θ) of

$$\Theta_k \leftarrow \text{argmin}_\Theta \Big\{ -n/2(logdet(\Theta) - trace(S\Theta)) - \rho/2||\Theta - Z_{k-1} + U_{k-1}||_F^2 \Big\}$$

ii Update Z parameter as minimization of:

$$Z_k \leftarrow argmin_Z \Big\{ \tfrac{\rho}{2}||Z - (\Theta_k + U_{k-1})||_F^2 + g(Z) \Big\}$$

iii Update U as:

$$U = \Theta_k + Z_k$$

4: **end for**

5: **return** Θ

4.2.3 Scoring microRNA targets

In this stage, the predicted targets are scored, and each microRNA target is ranked based on the estimated scores. Each target gets a set of scores from a set of features. These features are conservation, site accessibility, context matching, and number of seed binding sites.

Conservation

Conservation refers to the evolution of a sequence across species. Target binding sites are functional sequences. This fact makes the target sites subject to evolutionary conservation across various organisms. Therefore, it can provide evidence that the predicted target site is

functional. The role of conservation in microRNA target prediction is broad and has been incorporated into prediction in various ways, based on the prediction method itself. The reference species used here are chimpanzee, mouse, and dog. To determine which binding sites are conserved in the reference species, we started with the binding site in the 3′-UTR that is complementary to a microRNA seed region and search the genomes of the reference species for matches. A seed binding site is considered to be conserved in a species if there exists at least one site in that species with the corresponding seed complementarity. Ensembl API [162] is used to compute the average seed match probability to be a conserved element, and we use this probability as the conservation score.

Site Accessibility

Site accessibility is a measure of how easily a microRNA can locate and hybridize with its target. When a microRNA binds to its target mRNA, it forms a duplex. The minimum folding energy for the duplex is used to measure the site accessibility. A minimum binding site length was proposed by [92]; it suggested that duplex formation requires a minimum of 7 nucleotides. However, the free energy has been computed for both the 7 nucleotides seed binding sites as well as the maximum matching region between the microRNA and the mRNA. The Vienna package [53] is used to compute the score for both the seed binding sites and the maximum matching region. Let $\triangle G_{bind}$ be the energy gained by binding of the microRNA to the mRNA, and $\triangle G_{open}$ be the estimated as the free energy of the 3′-UTR constrained to maintain the binding site single stranded subtracted from the free energy of the same unconstrained 3′-UTR. Then, the minimum free folding energy ($\triangle G_{duplex}$) of the microRNA-mRNA duplex estimated as:

$$\triangle G_{duplex} = \triangle G_{bind} - \triangle G_{open}.$$

If we have n binding sites in the 3′-UTR of a target, and $\triangle G_{duplex_i}$ is the the minimum free folding energy of the site i in the mRNA, then the score is calculated as in [157] for the site accessibility of the target as

$$Score = \log\left(\sum_{i=1}^{n} e^{\triangle G_{duplex_i}}\right).$$

Algorithm 2 Filtering out the indirect interactions algorithm that is applied for each microRNA

```
for target i ∈ microRNA j target set do
  if B_ij < 1 then
    Target(i) ← dropped
  else
    if B_ij ≥ 1 then
      Target i ← pass
    end if
  end if
end for
```

The cofold function of the Vienna RNA Secondary Structure library is used. This function is specifically designed to compute the duplex free energy. It takes into account the intra-molecular and the inter-molecular pairs, which make it more accurate than the duplexfold function that is used in PITA [67].

Context Matching

Context matching refers to the properties of the sequence mapping between the microRNA and its target. These include the mismatches, which include G:U wobble pairs or gaps in the seed region, the number of nucleotide matches around the seed region, and the distance between the seed binding site and the 3′-UTR start, which is computed as the number of nucleotides from the target site to the closest 3′-UTR end point. This distance is scaled by dividing by the length of the 3′-UTR. A vector A_{ij} is define for each predicted interaction between target i and microRNA j to encode this contextual information. A_{ij} contain 4 values. The first one (a_{ij1}) is the number of the seed binding sites. The second value (a_{ij2}) is the number of mismatches in the seed region. The third value (a_{ij3}) is the number of nucleotides matches around the seed region, and the last value (a_{ij4}) is the distance between the seed binding site and the 3′-UTR start estimated as explained earlier.

4.2.4 Target ranking

An integrated ranking score was developed by combining the information from the scoring features described above. For this propose, the support vector regression (SVR) algorithm [149] is employed to model the degree of microRNA regulation given the numerical values of the features set (binding site accessibility, conservation, and contextual information).

SVR is a nonlinear regression method and is a special class of kernel based regression. Sometimes, it is viewed as an alternative to neural networks, with the advantage that the problem is rewritten as a quadratic programming problem or as a least squares problem for least squares. SVR models are able to model nonlinear relationships between variables using the kernels. A typical use of the SVR involves two steps: first, training a data set to obtain a model and then using the model to predict information of a testing data set. SVR model outputs the probability estimates for each target. Then this probability is used to rank the targets.

The SVR model uses labeled training data to learn a function that estimates the output probability for a target from its feature vector. Suppose that the labeled training data (x_i, r_i) for $i = 1, 2, \ldots, m$ is used to learn a linear function f as:

$$f(x) = (w, x) + b.$$

$f(x)$ estimates the output valued r for a sample from its feature vector x, w is the weight vector, and b is the bias term. SVR uses an ϵ-insensitive loss function $l(f(x), r) = max(0, |f(x) - r| - \epsilon)$ that makes the model only penalize samples whose outputs fall outside ϵ and around the prediction function [149].

The feature vector for each target is a vector of the scores estimated in the scoring. The training data are obtained from miRTarBase v4.5 [56], MirWalk [31], and OncomiRdbB [69] and are input to the model as the feature vectors for the real targets from these data sets. Then, the inferred function is applied on the test data, the predicted targets from Stage 2, and estimates the score for each predicted target. The LIBSVM package [23] has been used. In its model, the RBF (Gaussian radial basis function) kernel function is used, and the parameters α (which control the peak of the Gaussian functions) and β (which control the cost for the regression errors) were adjusted using leave-one-out cross-validation method on the training data.

4.3 MicroTarget Results

4.3.1 Data sources

The sample microRNA and mRNA expression profiles from an earlier study [33] have been used. The expression data of 518 microRNAs from 105 breast cancer tissue samples in this publication have been deposited in NCBI Gene Expression Omnibus (GEO) and are accessible through GEO Series accession number GSE19536. The expression profile of 30,982 mRNAs from the same tissue samples are accessible through GEO Series accession number GSE19783.

Mature microRNA sequences were downloaded from miRBase database [76]. The miRBase database is a large database for published microRNA sequences and annotations. The current release (version 21) contains 28,645 entries of microRNAs sequences in 223 species. We downloaded microRNA sequences for human. Full length 3′-UTR sequences were downloaded from the Ensembl database [161] using the BioMart tool [73]. Ensembl BioPerl is used to generate the 3′-UTR sequences for all human mRNA transcripts. When multiple transcripts are available for a gene, the longest isoform is used. Ensembl has also been used for downloading species conservation information (human, chimpanzee, mouse, and dog).

Given the expression for microRNA and mRNA from the same samples, MicroTarget quantifies the regulatory effect for microRNA on mRNA. The expression-based identification considers both up- and down-regulations. The microRNAs have increasingly been linked to functions that are either tumor promoting or tumor suppressing. Changes in microRNA expression and their targets have been noted at various stages of cancer progression [80]. The changes in the expression of miR-200 family members have been documented in various types of cancer, including lung, ovary, stomach, and breast cancer [87]. The members of the miR-200 family are miR-200a, miR-200b, miR-200c, and miR-429. Also miR-146a, let-7, and their targets have been experimentally tested for their association with breast cancer [11]. Therefore we have used the miR-200 family, let-7, and miR-146a to emphasize how MicroTarget performs better in tissue specific prediction.

Ground truth for validation

Once microRNA targets are predicted, the next step is to validate the predicted microRNA-target interactions with the experimentally validated interactions. As the number of experimentally validated targets of microRNAs are still limited, we use the union of three regularly updated databases. These databases are miRTarBase v4.5 [56], MirWalk [31], and OncomiRdbB [69]. OncomiRdbB and miRTarBase include verified interactions that are manually curated from the literature, while miRWalk contains experimentally validated and predicted targets, only the validated targets have been used. There are 20,195 interactions with 348 microRNAs in OncomiRdbB, 25,810 interactions with 246 microRNAs in miRWalk, and 37,372 interactions with 576 microRNAs in miRTarBase. After removing the duplicates, the total number of unique interactions is 56,858; we refer to these as validated interactions.

4.3.2 Performance comparison with existing methods

The main idea of MicroTarget is to combine expression data of mRNAs and microRNAs from the same samples, with sequence data, to improve the specificity and sensitivity of the predictions. Our approach provides for each microRNA a group of mRNAs that are identified as its predicted targets in a particular experiment or condition, and a corresponding score for the significance of this prediction. An extensive evaluation of MicroTarget was carried out using the data set explained earlier. To investigate the performance of our approach over the commonly used microRNA target prediction methods, we apply TargetScan, MirWalk, and GenMir++ prediction methods to our data sets and compare their performance with MicroTarget. We limited our gene set to the genes for which we have their expression to compare our results with the other three methods. The validation results using experimentally confirmed databases show that the results of our approach perform better than other methods.

Figure 4.3 presents a comparison between MicroTarget and three other methods in terms of the number of validated interactions out of the predicted ones. It shows the percentage of the real interactions predicted by our approach and by the other three methods. Our approach has the largest number of confirmed predicted sites compared to the other tools. MicroTarget is able to predict 76.24% of the validated interactions, compared to 58.2%, 48.96%, and 63.46% for TargetScan, GenMir++, and MirWalk, respectively. MirWalk is

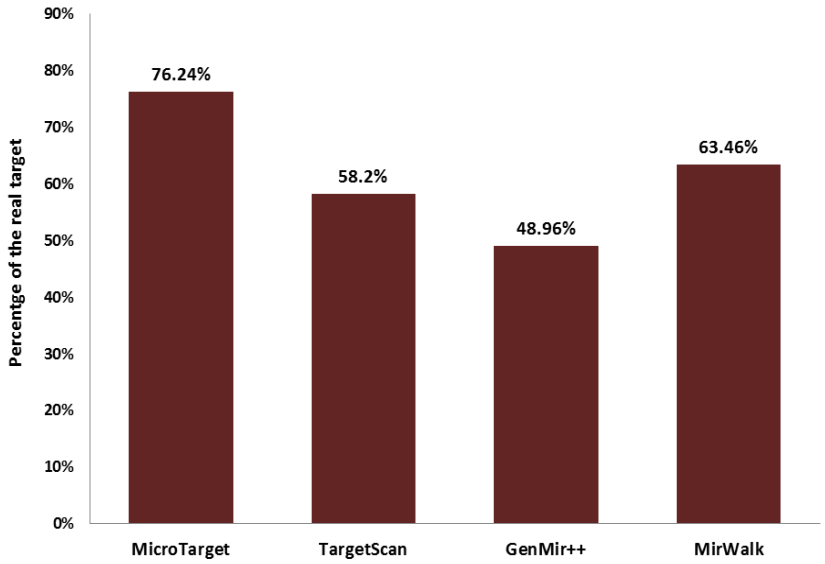

Figure 4.3: Comparison with the existing methods with the percentage of the overall validated targets that have been predicted by each method.

quite close in the percentage. This happens because MirWalk integrates result from more than one algorithm, each with different filtering features, and combining the results together. The above results demonstrated the successful performance of MicroTarget in the human data set in the same cell type.

Further analysis of the results of MicroTarget shows that it can obtain more targets that could not be found by the existing methods in the comparison, and the discovered targets are statistically significant and functionally enriched in the cell tissue under study. The results shows that MicroTarget outperforms existing methods by predicting microRNA-mRNA interactions that cannot be predicted by other methods. For instance, Figure 4.4 shows the interactions for mir-96 and mir-141 and their validate targets from our approach predicted when other methods fail. It was generally believed, until recently, that microRNAs exerted their repressive action on their targets via translation down-regulation. However, a study at [88] shows that microRNA can mediate target up-regulation. Using expression data for identifying targets considers both up- and down-regulations. In fact, there are 581 up-regulations in the data set [80]. MicroTarget is able to identify 485 (83.47%) of those regulations. On the other hand, MirWalk and GenMir++ were only able to predict 8 (1.3%) and 43 (7.40%)

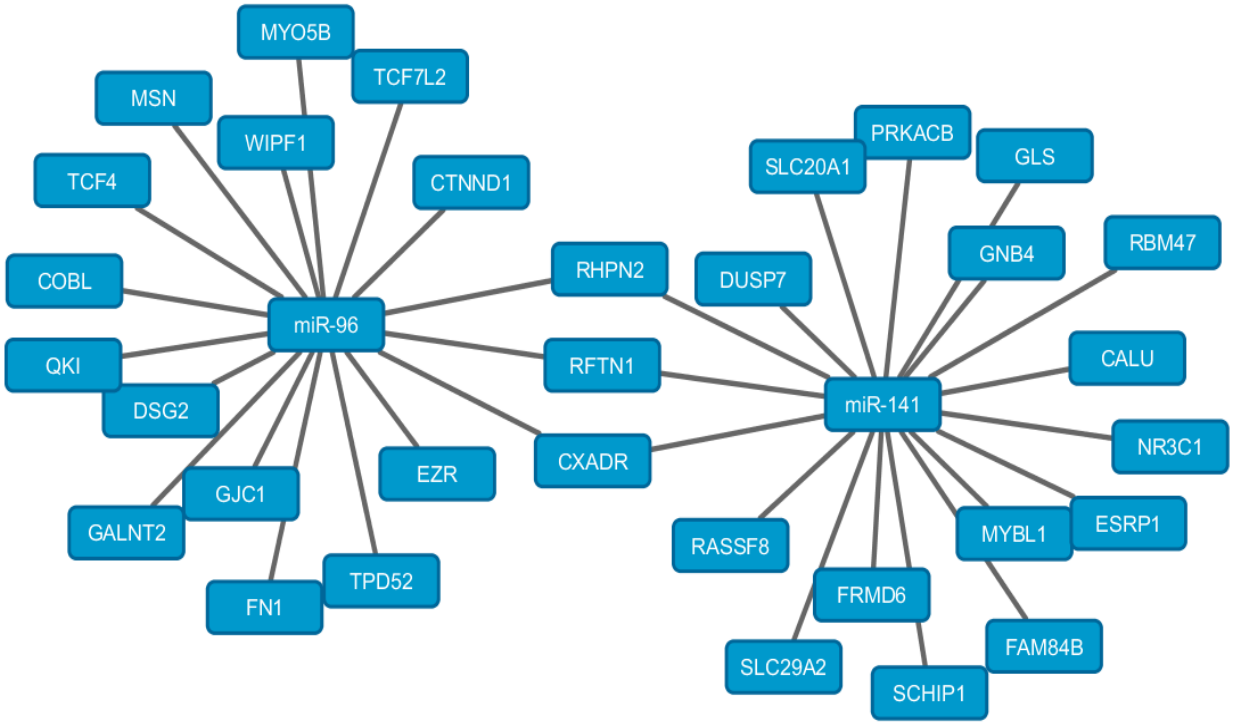

Figure 4.4: Small network for mir-96 and mir-141 and their predicted targets from our approach.

respectively, while TargetScan does not predict any of these regulations. This suggested that the traditional methods like TargetScan almost cannot reliably predict these interactions. Compared to sequence based predictions, our approach does not filter the prediction results like existing methods do, but provides probability for ranking each target, which helps in predicting novel targets for experimental verification. To our knowledge, this technique is novel for microRNA target prediction

Top scored predicted targets

We preform statistical analysis of the predictions by each method based on z-score. This z-score reflects the performance of a prediction method in finding validated targets comparing to the expected rate in the ground truth data set. The z-score can be defined as follows:

$$z - score = \frac{R - \mu}{\sigma * \sqrt{n}}$$

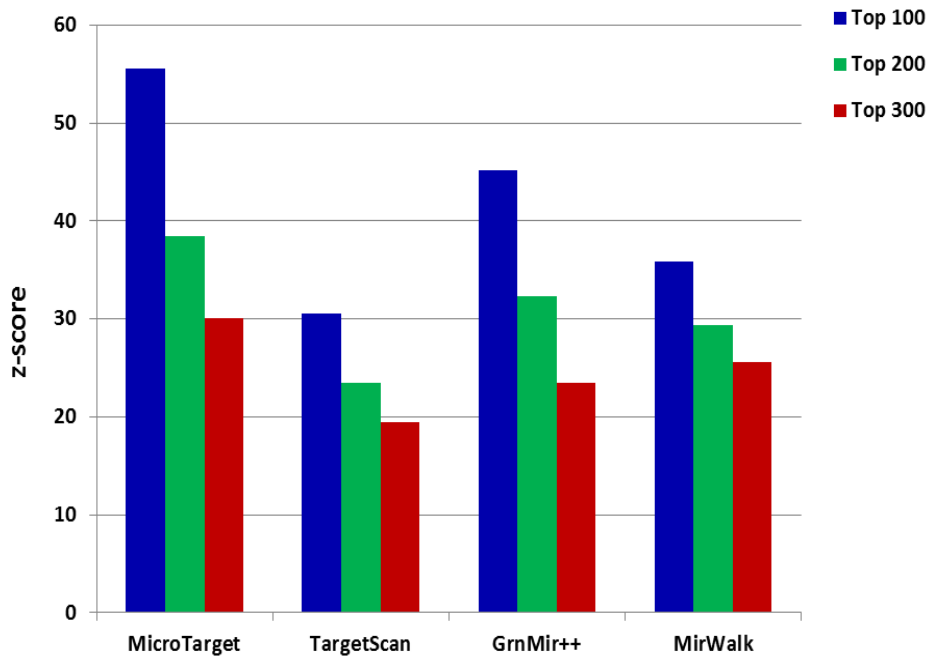

Figure 4.5: Z-score comparison with the existing methods for the top scored targets.

Here, R is the ratio of number of confirmed targets and number of all possible microRNA-mRNA interactions in a data set, μ is the ratio of confirmed targets in the expressively validate targets and all possible microRNA-mRNA interactions, and σ is the standard deviation and calculated using the Bernoulli distribution as $\sigma = \sqrt{\mu(1-\mu)}$. A higher z-score indicates more significant prediction results. Figure 4.5 presents z-score comparisons between our approach and the other three methods for the top scoring 100, 200, and 300 targets. MicroTarget shows a better z-score value for its top scored target that other algorithm. For the top 100 scored target, MicroTarget has z-score = 55.5 compared to 30.5, 45.2, 35.8 for TargetScan, GenMir++ and MirWalk respectively.

ROC analysis for MicroTarget

The performance of MicroTarget has been analyzed using Receiver Operator Characteristic (ROC), which is shown in Figure 4.6. ROC is a plot of the true positive rate (sensitivity)

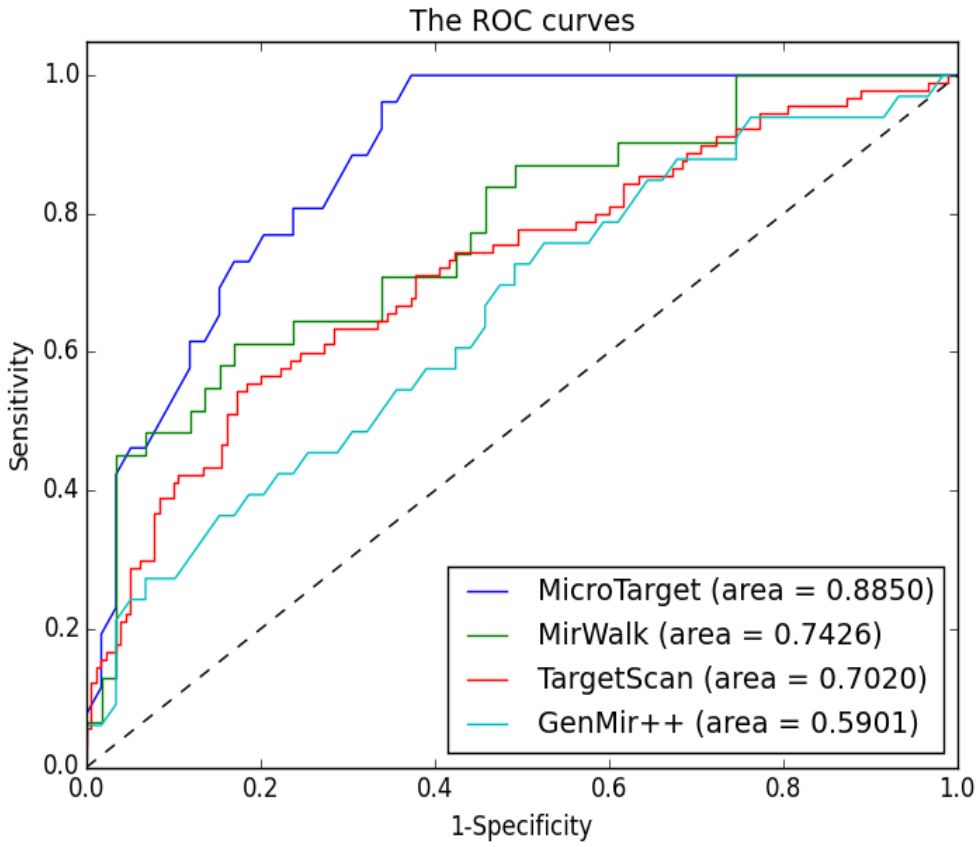

Figure 4.6: The ROC curves of MicroTarget, targetScan, MirWalk and GenMiR++.

against the false positive rate (1-specificity) for the different possible cutoffs of a diagnostic test, where

$$sensitivity = TP/(TP + TF)$$

$$specificity = TN/(TN + FP)$$

Here TP represent a true positive, TN stands for true negative, FN stands for false negative, and FP represents false positive. Sensitivity is also called true positive rate, specificity represents the false positive rate. The Area Under the Curve (AUC) of each method is calculated to measure the performance of the method. The higher the AUC, the better the prediction. We apply MicroTarget and GenMiR++ on the breast cancer expression data and run targetScan and MirWalk prediction. Then we compute their true positive rate and false positive rate under different overlap thresholds.

Table 4.1: Breast cancer related-genes and the number of predicted microRNAs and the validated microRNAs

Gene	MicroTarget Predicted	targetScan Predicted	GenMir++ Predicted	MirWalk Predicted	# of Validated microRNA
BRCA1	101	89	43	67	107
BRCA2	34	20	17	20	37
CDH1/FZR1	21	20	15	19	21
FOXO1	28	25	17	17	30
EZH2	43	30	29	30	47
HIF1A	51	47	49	41	51

The figure shows the ROC curves and AUC values. As can be seen, MicroTarget has the better performance in term of AUC, 0.8850, which should be expected since it considers a variety of features in prediction, while MirWalk, TargetScan and GenMir++ get 0.7426, 0.7020, and 0.5901 respectively. TargetScan has relatively good sensitively but produces high false positives. For a small false positive rate, MirWalk can achieve relatively higher sensitivity than GenMir++.

4.3.3 Studying the tissue-specificity of the prediction

It has been shown that many microRNAs exhibit tissue-specific expression patterns and lead to tissue-specific profiles for their targets [38]. Changes in microRNA expression and their targets have been noted at various stages of cancer progression [80]. The OncomiRdbB [69] database has microRNAs and their targets that have been frequently shown to be deregulated in cancer. Table 4.1 represents some of the cancer-related genes and the number of their regulatory microRNAs from the different methods [133]. For instance, MicroTarget was able to predict 101 regulators for BRCA1 out of 107 validated regulators. Using expression data in the prediction enables our approach to identify the targets that are strongly associated with the biological condition of interest.

There are four microRNAs, miR-200a, miR-200b, miR-200c, and miR-141, all of which are part of the miR-200 family. These microRNAs are known to have a role in breast cancer. Figure 4.7 shows a Venn diagram for the miR-200 family predicted targets versus experimentally validated targets. The numbers in the yellow circle are the number of validated targets

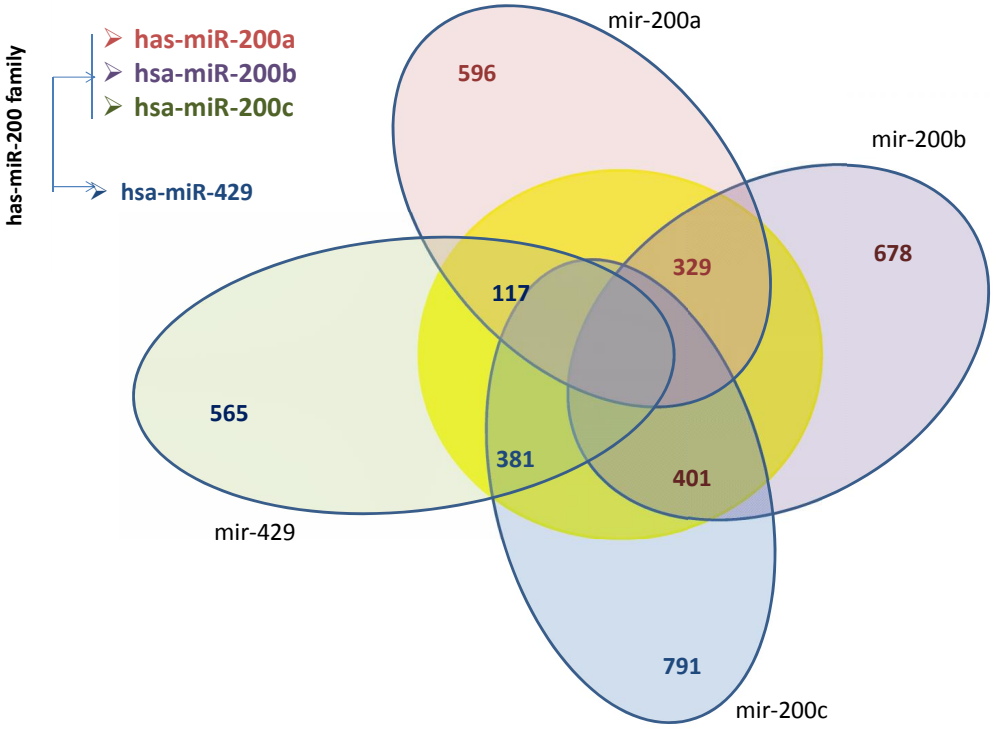

Figure 4.7: Venn diagram for the miR-200 family predicted targets versus experimentally validated targets. Numbers in the yellow circle are the experimentally validated targets from MirTarBase and MirWalk.

that MicroTarget predicted, while the numbers outside of the yellow circle are the novel predicted targets. In total, 1,228 true targets were predicted out of 1,371 for the miR-200 family. For instance, 329 miR-200a targets out of 358 validated targets were predicted.

4.3.4 Analysis of the scoring features

To understand the mutual relationship between the predicted target scores and the set of features, Spearman rank correlation [104] between the feature pairs has been performed. Spearman rank correlation is a non-parametric test that is used to measure the strength of association between two variables. The coefficient $r = 1$ means a perfect positive correlation, and $r = -1$ means a perfect negative correlation. For a correlation between features x and

Table 4.2: Correlation among features that are used for scoring the predicted targets. Number of matches refers to the number of seed binding sites between the microRNA and the mRNA. Matching length refers to the maximum sequence complementarity between the microRNA and the gene. Seed ΔG and total match ΔG refer the site accessibility estimated based on the seed region and the maximum sequence complementarity, respectively. P_{value} points to the P_{value} of the seed binding site prediction

		Matching Length	No.of Matches	Seed ΔG	Total Match ΔG	Conservation	Matching P_{value}
Matching Length	r	1.00	-0.069435	0.709736	0.608855	0.038358	0. 82109
	p	0.00	0.0072552	0.000001	0.000008	0.548400	0.000000
No.of Matches	r	-0.069435	1.00	0.642026	0.500000	0.608855	0.98000
	p	0.0072552	0.00	0.00031	0.0001	0.00421	0.00580
Seed ΔG	r	0.709736	0.642026	1.00	0.56939	0.214750	0.642026
	p	0.000001	0.00031	0.00	0.00067	0.000800	0.00658
Total Match ΔG	r	0.608855	0.500000	0.56939	1.00	0.038358	0.500000
	p	0.000008	0.0001	0.00067	0.00	0.005484	0.001054
Conservation	r	0.000008	0.0001	0.214750	0.038358	1.00	0.608855
	p	0.038358	0.608855	0.000800	0.005484	0.00	0.000320
Matching P_{value}	r	0. 8210	0.98000	0.642026	0.500000	0.608855	1.00
	p	0.00	0.00580	0.00658	0.001054	0.000320	0.00

y, the formula for calculating the coefficient is

$$r = 1 - \left(6 * \sum_{i=1}^{n} (d_i^2)/(n^3 - n) \right).$$

where d_i is the difference in score from x to y and n is the number of data points. Spearman correlation coefficients between the pairs of the features and the p-value of the correlation are shown in Table 4.2. Each cell contains the Spearman rank correlation coefficient r and the p-value of the correlation. Let the matching length be the number of nucleotides complementary between the microRNA and the mRNA. The positive correlation between the matching length and the matching p-value indicates that a high level of sequence matching is associated with high scoring for the target.

4.3.5 Evaluating SVR model for the ranking

Performance comparison of MicroTarget target ranking has been preformed by an ROC analysis with different SVR training data sets. Training data sets are retrieved from the

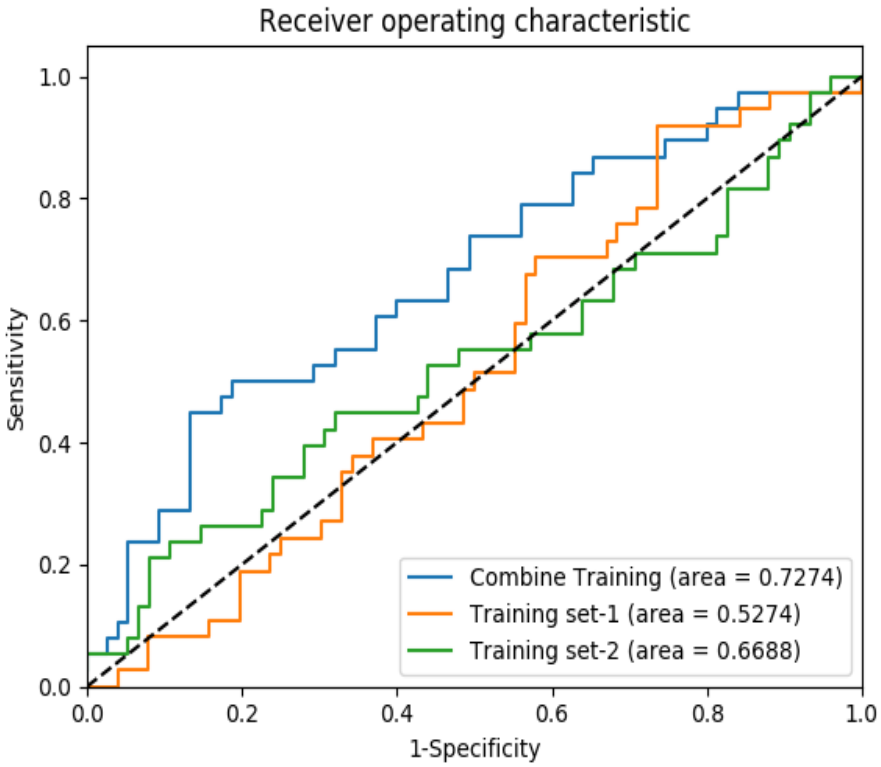

Figure 4.8: ROC analysis for the SVR model with different data sets

experimentally validated target databases, explained in the data set section. The positive microRNA-mRNA interactions are the interactions downloaded from the database. The negative interactions are obtained from the filtered data from the first stage of MicroTarget, indirect interactions inferred from the gene expression data. Table 4.3 shows two data sets that are used in the study. The third data set combines the two data sets in the table. Figure 4.8 shows how ROC curve for MicroTarget prediction with different data sets. The results from the ROC analysis indicate that MicroTarget has better target ranking with the combined data set over the other two data sets. Given the difference between results, in terms of the area under the curves, it only seemed natural that incorporating more interactions to the training data seems to improve our model performance.

Table 4.3: Positive and negative data sets for SVR analysis

	Positive	negative
Set 1	587	3706
Set 2	1634	4917

Testing SVR Kernel function

We then compare the performance of our SVR ranking model for each microRNA based on the number of validated targets with different kernel function. We create three models, one for each kernel function. As we have three models, with respect to each microRNA, we score each model using a number (called the M-ranking score) in the range of 1 to 3, with 3 indicating the best model and 1 the worst model. Finally, we calculate the M-ranking score of each model for the data set by summing up its scores for all microRNAs. The higher the ranking score of the model, the better the kernal function is. From Figure 4.9, we can see that the RBF (radial basis function) model outperforms the other models. Meanwhile, the other two models performance changes for the top 100, 200, and 300 scored targets.

4.4 Discussion

MicroTarget takes advantage of the fact that, for the microRNA to regulate its target, both have to be in the same tissue. When a microRNA regulates its targets, this regulation effect should propagate across the cell process. This effect can be better interpreted by integrating the expressions of genes and microRNA as well as the sequence data in the prediction. We have demonstrated that MicroTarget can be a valuable resource to improve the efficacy of microRNA target prediction. MicroTarget does not filter the prediction results like most of the prediction methods do. That helps in predicting novel targets for further experimental verification.

The result analysis highlights many cases in which microRNA families are predicted to regulate multiple members of breast cancer-related genes. In one case, our method predicts that the miR-200 family directly targets and regulates CCNE1, CDC16, ADAM10, and FOSL1. These genes are components of the Notch signaling pathway, especially, FOSL1 (Fos-Related Antigen 1) [13]. This pathway is involved in both the development and progression of breast cancer [1]. Also, miR-106b is predicted to directly target TGFBR2, CDKN1A, and

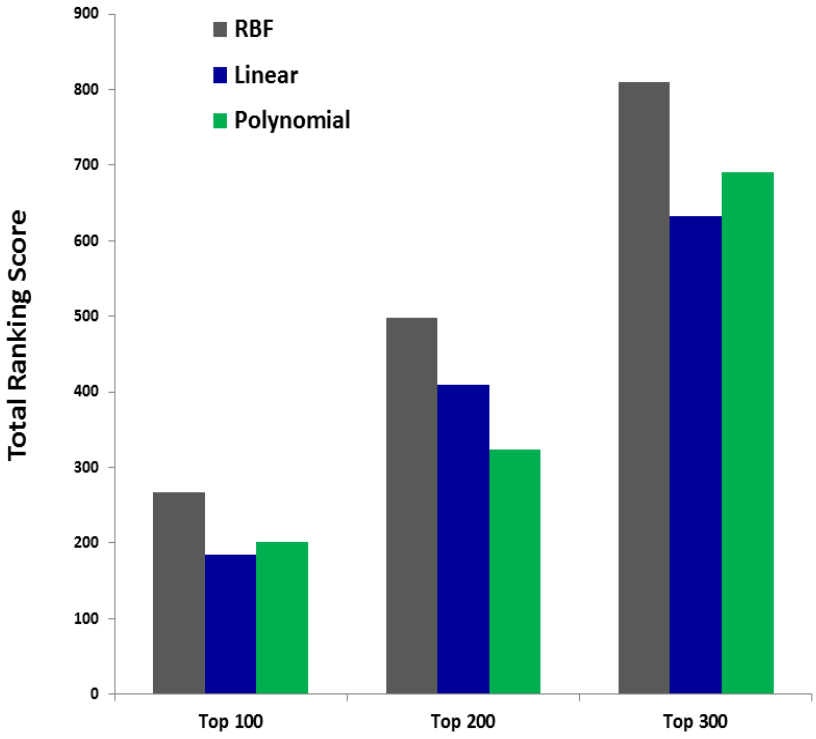

Figure 4.9: Total ranking score for the top 100, 200, and 300 scored target with different kernel functions for the SVR model.

DAB2. The TGFBR2 and DAB2 genes are components of the TGF-β signaling pathway, which is involved in many cellular processes including cell differentiation, cell growth, cellular homeostasis and apoptosis. This prediction is consistent with the hypothesis that miR-106 is oncogenic in breast cancer, and CDKN1A is known to regulate cell cycle progression [62].

MiR-17-5p is known to play a role in cancer cell proliferation [55]. It represses the translation of AIB1 mRNA, thereby inhibiting the function of E2F1 and ER α [83]. The down-regulation of AIB1 by miR-17-5p results in the suppression of estrogen stimulated proliferation and estrogen/ER-independent breast cancer cell proliferation. The regulatory interaction between miR-17-5p and AIB1 has been predicted by MicroTarget and mirWalk, while targetScan and GenMir++ fail to infer this interaction. Another interesting observation is the finding that the let-7 family regulates the expression of the RAS and HMGA2 gene in human breast cancer [81]. These interactions have been predicted by our approach, while the other three approaches have not. Also, miR-21 has been reported to be associated with

invasive and metastatic breast cancer and regulates HIF1A in breast cancer cells [148]. The co-regulation of miR-411 and miR-21 on HIF1A has been predicted by MicroTarget.

MicroTarget cannot accurately infer targets for microRNAs that are not expressed in the same tissue, because variation in expression for such microRNAs would in most cases not have an association with the target expression. The inferred microRNA-target interactions show the specificity of the prediction.

Chapter 5

Conserved Protein Complexes: Biological Background

The nucleus of every cell in an organisms contain a large DNA (deoxyribonucleic acid) molecule, which carries the genetic information of the organism. This DNA sequence contains instructions for the synthesis of every protein. A protein is a sequence of 20 different kinds of amino acids. Each amino acid is uniquely determined by three RNA nucleotides. Once we know the sequence of a gene, we can also know the sequence of the corresponding protein. Proteins are involved in many essential processes within the cell, such as gene regulation, metabolism, transmission of signals, and DNA repair [34]. Proteins rarely act alone. They interact together to form larger structures, such as protein complexes and pathways. Protein interactions play a basic role in most biological processes. Protein complexes that are conserved across species indicate core biological processes of cell machinery [18]. This chapter gives biological background on protein complexes, protein interaction networks, domains and domain interactions.

5.1 Protein-protein interaction

Proteins physically interact with each other to perform biological processes. A main step towards understanding the cellular machinery is to build a complete map of protein-protein interactions (PPIs) (sometimes called the interactome). Protein interactions can be categorized as stable or transient. Proteins interactions that are purified as subunit complexes

are the stable interactions, like core RNA polymerase proteins that interact to form a stable complex. Transient interactions on the other hand are temporary and often require a specific set of conditions to occur, such as that the interaction proteins must be located in specific area of the cell [117]. Transient interactions control major cell processes, such as cell cycling, protein modification, signaling, and protein folding.

A PPI network provides a conceptual view that describes a global mapping of protein interactions in a graphical framework. The nodes and edges of the network represent proteins and their interactions. Many PPI network databases have been constructed for a variety of organisms [137]. These networks are a collection of interactions from different experimental techniques. Many high throughput techniques have been developed over the last decade to detect protein interactions, for instance yeast-two-hybrid, and and tandem affinity purification coupled with mass spectrometry.

5.1.1 Identifying Protein Interactions

There are multiple experimental approaches to detect protein interactions. The most widely used one is the yeast-two-hybrid system (Y2H). In the Y2H technique, protein X, which is the protein of interest, is fused to the DNA binding domain and the complex is called the bait. Then the potential interacting protein Y is combine with the activation domain and the complex is called the prey. If the X and the Y actually interact, then their interaction will form a functional transcriptional activator that leads to recruiting the RNA polymerase II and subsequent transcription of a reporter gene. The Y2H technique has been enhanced into two main approaches for screening entire genomes. The first approach is a matrix approach, where all possible combinations between full-length open reading frames are systematically examined by performing direct mating of a set of baits versus a set of preys expressed in different yeast mating types. The defined position of each bait in a matrix allows rapid identification of interacting preys based on the expression of a reporter gene without sequencing [20]. The second approach is a library approach, which searches for pairwise interactions between the bait proteins and their interaction partners (preys) present in cDNA libraries or sub-pools of libraries, and the interacting proteins are determined by colony PCR analysis and DNA sequencing.

Another popular technique for detecting protein interactions is affinity purification coupled to mass spectrometry (AP-MS). In this technique, affinity tags are attached to a protein of

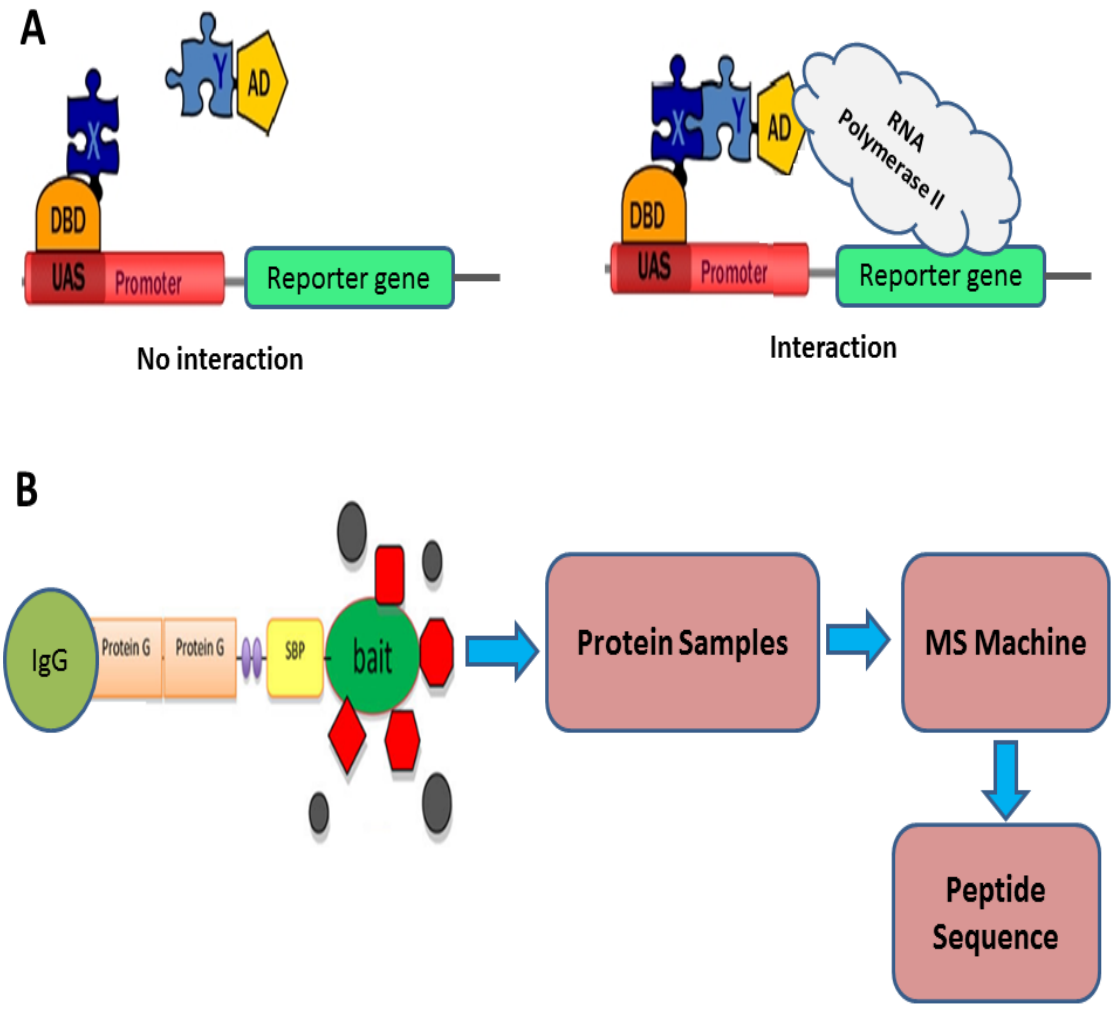

Figure 5.1: PPI identification methods; A) The yeast-two-hybrid system: If protein X and protein Y interact, then their DNA-binding domain (DBD) and activation domain (AD) will combine to form a functional transcriptional activator, UAS refers to upstream activator sequence of the promoter [20]. B) affinity purification coupled to mass spectrometry; first, tagged protein is pulled down via its tag together with the associated proteins and other non-specific interacting proteins. Then the protein samples collected are broken down into peptides and analyzed by mass-spectrometry. Finally, the list of peptide is sequenced and the proteins from each sample are reported as the interaction ones [141].

interest and systematic precipitation of the bait proteins is performed. Then, the proteins are separated according to their mass to detect purified protein complexes. Finally, the proteins are removed from the gel and analyzed by mass spectrometry techniques [137]. Figure 5.1 shows the general principle of the yeast-two-hybrid, and affinity purification processes. AP-

MS is less accessible than Y2H due to the expensive large equipment needed. AP-MS can determine all the components of a larger complex, which may not necessarily all interact directly with each other, while Y2H identifies the binary interactions.

Another technique for protein interaction identification is co-immunoprecipitation (Co-IP), which identify physiologically relevant PPIs by using target protein specific antibodies to indirectly capture proteins that are bound to a specific target protein [137]. This technique is working in the same manner as an immunoprecipitation of a single protein. The interacting protein is bound to the target antigen, which is bound by the antibody that is immobilized to the support. The proteins and their binding partners are then detected using western blot analysis. This technique is often used when the proteins under the experiment are related to the function of the target antigen at the cellular level.

A new important method for studying protein interactions is the pull-down technique. A pull-down assay is similar to co-immunoprecipitation, except that a bait protein is used instead of an antibody, where a tagged protein, called the bait, is used to capture a protein binding partner, called the prey [158]. Pull-down assays are mostly used for confirming the existence of a protein interaction predicted by other research techniques or as an initial screening assay for identifying unknown interactions.

Another proteomic method for identifying protein interactions is protein-fragment complementation assay (PCAs) [158]. PCAs can be used to detect PPI between proteins of any molecular weight and expressed at their endogenous levels. Protein microarrays can also be used to detect protein interactions and functions. A protein microarray is a piece of glass on which various protein molecules have been attached at separate locations in an ordered manner [30]. The objective behind the protein microarray technique is to achieve sensitive high-throughput protein analysis and to carry out large numbers of analysis in parallel. This method has seen much interest and become one of the biotechnology active areas of interest.

Synthetic lethality is also used for uncovering protein interactions. This method is based on the idea that genetic variation influences phenotype. First, it involves mutation of two genes that are capable of working successfully alone but cause lethality when combined in a cell under specific conditions. As these mutations are lethal, the two genes cannot be separated directly. They should be synthetically constructed. Then the methods tests if there is a physical interaction between the two gene products or not [42].

Even though these approaches identify many PPIs with high confidence, they still suffer

from high false positive and false negative rates [94]. Given the challenges in identifying PPIs experimentally, computational approaches have been proposed. These approaches are working on identifying a large network of thousands of protein interactions using statistical and machine learning techniques [120]. These approaches can be categorized based on the types of data they used for prediction as follows:

- Methods that infer protein interactions based on gene fusion events and conservation of gene neighborhood.
- Methods that use domain pairs or motif pairs observed in interacting protein pairs, along with structural information and sequence evidence about PPI interfaces.
- Methods that are based on the assumption that interacting proteins should undergo co-evolution in order to keep specific function shared between organisms. This type of methods are called *in-silico* two-hybrid (I2h) [114]. They also focus on analyzing physical closeness between residue pairs of the two individual proteins. The result from these methods indicate the possible physical interactions between the proteins.

5.2 Protein Structure

Each protein contains a polypeptide backbone that is attached to side-chains. Proteins deffer in their sequence and amino acid number. The sequence of the different side chains makes each protein distinct. The structure and shape of the proteins is relevant to determine their specific function [14]. Also the structural knowledge of proteins can help understanding of how a protein interacts with other molecules, which also gives important hints on protein functions.

Protein structure can be described at several levels. The primary structure corresponds to the linear amino-acid sequence. It describe the order of the backbone and the side-chains held together by covalent bonds. The sequence of these amino acids in the polypeptide chain determines the secondary structure of the protein. The tertiary structure is the path of the chain in 3-dimensions (3D) resulting from various long interactions [129]. Large proteins consist of several distinct structural units, called domains, that fold independently of each other. The Protein Data Bank (PDB) [128] has a large archive for the structural data of biological molecules. The available protein 3-dimensional structures in the PDB have been

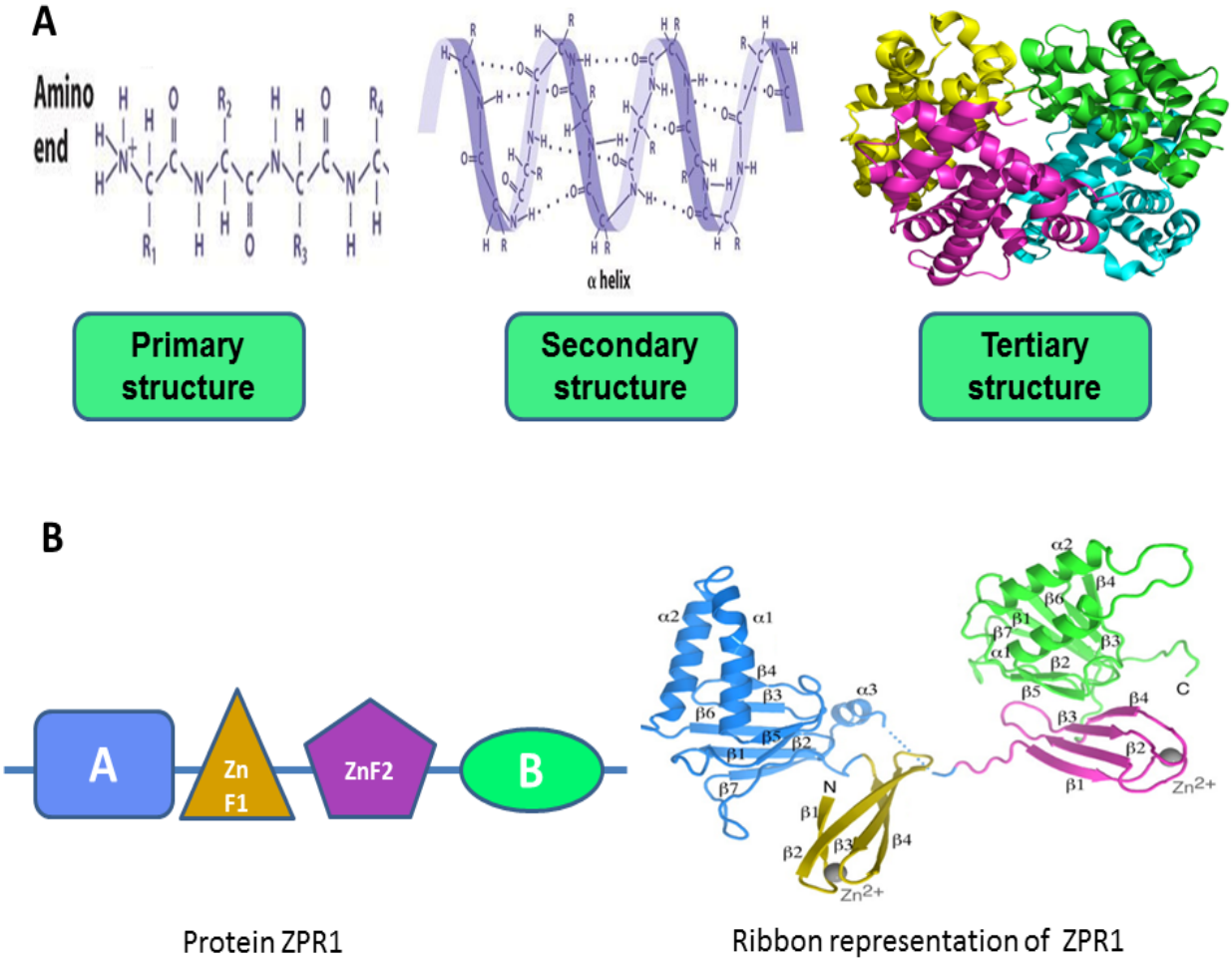

Figure 5.2: (A) type of protein structure [129]. (B) An example of domain organization tertiary structure of protein ZPR1 as in Pfam database; the schematic illustration of the modular architecture, and ribbon representation of the tertiary structure [39].

classified into more than one thousand unique folds. Each domain in the multi-domain protein has its own structure and function, and works with its neighboring domains to perform their tasks [10].

5.2.1 Structural domains

The term domain often relates to protein structure or function, our interest here is in the protein structure. Protein structural analysis begins with dividing the structure of the protein into its basic units, namely its structural domains. Protein can has a single domain or multiple domains. Protein domains are a set of simple and structurally meaningful units. The arrangement of domains in a protein is defined as its domain architecture [121]. To define which domains occur in which protein, we use the domain definitions from Pfam [39], which is projected onto the PDB structures. In Pfam, a structural domain is defined to be a compact structural unit that can fold independently of other domains. The Pfam database divides domains into two classes: Pfam-A which are manually curated and functionally assigned, and Pfam-B which are automatically generated based on the ProDom [19] database. Domains with the same fold may be functionally related to each other.

The idea of decomposing protein structure into domains was introduced by Wetlaufer [153]. Based on the criteria used for structural partitioning, some protein domains are annotated differently among databases. The interaction between two proteins usually involves a pair of constituent domains, one from each protein. The 3-dimensional structure is crucial for revealing how domains interact with each other, either in polypeptide chain level, or in complexes [40]. Additional criteria, along with the geometric definition, have been used to propose an automated methods for assigning structural domains, such as function, thermodynamic stability, and domain motions.

5.2.2 Domain-Domain Interactions

The binding interface of the proteins interaction is localized at the domains. As protein interactions generally occur via domains instead of the whole molecules, it is useful to know which specific domains of the proteins are interacting. To understand how domains interact at the molecular level, we need to know which amino acid residues and their atoms are interacting [12]. These data are available in the Protein Data Bank [128] database of protein structures. Experimentally identified 3-dimensional structures are a prime resource for understanding how interactions between domains are mediated. Therefore, it is widely used to obtain domain interactions, such as protein structure determination by X-ray crystallography. The iPfam [40] and 3did [103] are two databases that contain information on known

DDIs identified using the protein structure from PDB. The number of DDIs identified from structures is still fewer than the number of PPIs.

To accelerate the discovery of more DDIs, computational approaches have been proposed based on correlated sequence signatures and sequence co-evolution, gene fusion, phylogenetic profiling, gene ontology, and the parsimony principle [46]. Domain interactions can be divided into two types; heterotypic if the interaction involves two different domains, and homotypic if it involves two identical domains [61].

5.3 Protein complex

Many proteins perform their functions by integrating with other proteins to form protein complexes. A protein complex is a group of associated chains of polypeptides that are linked by non-covalent PPIs [112]. Protein complexes have a crucial role in biological processes, such as mRNA translation, DNA transcription, or signal transduction. Therefore, identifying protein complexes is important in molecular biology. Protein complexes can be identified using experimental techniques such as immunoprecipitation with high accuracy.

Some computational methods also have been applied to identify protein complexes from PPIs. One of the major challenges for detecting protein complexes computationally from PPI networks is that there is no mathematical formulation for protein complexes. Therefore, these methods depend on the observation that proteins within a complex interact closely with each other. Computational biologists usually use the idea that protein complexes form dense subgraphs and aim to search for dense regions in the PPI networks as protein complex candidates [138].

Chapter 6

Conserved Protein Complexes: Literature Review

Several methods have been proposed to search for a local mapping which illuminates conserved sub-structures in PPI networks. These sub-structures could be conserved protein complexes or pathways among the species of the PPI networks. There are two techniques for identifying conserved protein complexes from PPI networks. One is to compare the two PPI networks of the two corresponding species by aligning similar nodes and edges, then searching for potential regions in the aligned networks that could be conserved. The other is to use information from protein complexes of well-studied species, then match them to the network of a new species to identify subnetworks that are similar to the query complexes. The second technique is called network querying. In this chapter, we present computational methods used to define conserved protein complexes using network alignment.

6.1 PPI Comparative Analysis

As the amount of PPIs data for various species increases, comparative analysis of PPI networks across species is proving to be a valuable tool. This network analysis enables us to identify conserved functional components across species and perform high-quality ortholog prediction. Most comparative analysis approaches create a merged representation of the two networks being compared to facilitate the search for similarity between the two networks. The alignment may consist of one-to-one alignment, correspondence between two networks,

or many-to-many alignment, correspondence among multiple network.

The goal of network alignment is to find a mapping between the proteins and interactions of the networks. What makes the problem difficult is the trade-off involved in maximizing the overlap between the networks, while ensuring that the proteins mapped to each other are homologous. The network alignment problem can be formulated in various ways, depending on the kind of input and the scope of node mapping desired [139]. We can draw an analogy from the sequence alignment to differentiate between local and global network alignment:

- In global network alignment (GNA), the goal is to find the best overall alignment between the input networks (find a single consistent mapping covering all nodes across all input graphs). The mapping in a GNA should cover all of the input nodes. Each node in an input network is either matched to one or more nodes in the other networks or marked as a no-match [100, 113, 84]. Similar to global sequence alignment, GNA is used to compare interactomes and for understanding inter-species variations.

- In local network alignment (LNA), the goal is to find multiple, unrelated regions of isomorphism between the input networks, each region implying a mapping independent of the others. In contrast to GNA, an LNA algorithm is essentially intended for finding similar patterns between two networks where many independent local alignments are usually possible between two input networks. In fact, a protein can be mapped differently under each alignment. The motivations behind local sequence alignment and local network alignment are similar. The former is used to search for a conserved motif, while the latter is used to search for conserved functional components (for example pathways, or protein complexes) among species.

Local network alignment is the focus of our work. In general, LNA aims to align graphs in a way that display as much similarity as possible. There are several different definitions of what similarity between graphs might mean. LNA poses significant computational challenges, because it is related to the NP-complete subgraph isomorphism problem.

The most restricted definition of similarity between two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ is graph isomorphism. Two graphs G_1 and G_2 are isomorphic, if there exists a mapping $f : V_1 \rightarrow V_2$ that maps E_1 to E_2. The subgraph isomorphism problem is an extensions of the graph isomorphism problem to a more general case where the number of nodes is not equal. Subgraph isomorphism is known to belong to the class of NP-complete

problems [35]. The exponential time complexity of solving this problem encourages the researchers to propose general heuristic approaches to solving this problem for large graphs.

Conserved complex search strategy using LNA

Detecting conserved protein complexes between two or more species can be divided into two main steps. The first step includes organizing the PPIs data, and generates a network alignment graph, mostly based on protein homology data generated by methods such as BLAST [3]. The second step performs a search heuristic over the alignment graph and supplies a scoring model. Later, the results may be filtered to leave only the significant conserved protein sub-networks.

6.2 Existing LNA methods

In recent years many methods have been introduced for local network alignment. Local network alignment methods can be divided into two categories. One category starts with constructing an alignment graph, then uses this graph to find the conserved subgraphs between two or more networks. These methods either use seed and extend or clustering algorithms to find the conserved subgraphs. The other category of methods integrates biological information such as co-evolution or GO annotation to help with the alignment, we will call these information fusion methods. An overview of these methods is presented in the next sections.

6.2.1 Alignment graph based methods

Alignment graph methods start by building an alignment graph from the aligned networks, then search this graph for local alignments. Methods that use an alignment graph are based on the observation that complexes and functional modules correspond to highly interacting proteins. Therefore they are looking for sets of proteins that have more interactions among themselves than with the rest of the network [115]. Each of these methods impose a set of constraints on the topology of the aligned subgraphs.

Kelley et al. [66] proposed PathBLAST as a first method for local network alignment, with

the goal of aligning two PPI networks to identify the conserved pathways. The method identifies a set of high scoring alignments between pairs of pathways such that proteins in the first pathway map to their putative homologs in the same order in the second pathway. An alignment graph is first built in which a node represents a pair of putative homologous proteins, and an edge represents a conserved interaction. Gaps and mismatches are allowed in the edges. A match occurs when the two nodes are connected in the aligned networks. Otherwise, it is either a mismatch or a gap. A mismatch occurs when neither node is connected in the aligned network, and a gap occurs when only one of a pair of protein is connected. Then the highest scoring pathways are searched through the alignment graph using dynamic programming. The score is computed by decomposing the pathway similarity into a node scoring fraction and an edge scoring fraction. Using this scoring scheme, PathBLAST define an optimal alignment as one in which the pathway scoring function is optimized over all paths up to a user define length L for networks of size n. The presence of false negatives and positives on the PPI network leads to unreliable links in the alignment graph, causing PathBLAST to fail.

Kalaev et al. [135] extend PathBLAST into NetworkBLAST, which aims to identify not just simple linear pathways but also more complex subgraphs. It allows extraction of all conserved complexes across networks, as opposed to the single query model of PathBlast. It builds a weighted alignment graph by assigning a confidence value to each interaction [64]. Nodes in the alignment graph are allowed to be connected if the respective pairs of the orthologous proteins in the original network are at distance less than or equal to two. Then, the high-scoring seed nodes in the alignment graph are identified, and extension around the seeds in a greedy fashion approach is performed. NetworkBLAST has been generalized to NetworkBLAST-M [136] for identifying conserved subgraphs among multiple networks. It works with a layered alignment graph, in which each layer corresponds to a network. NetworkBLAST-M also uses a seed and extend strategy to identify high scoring alignments. The seeds nodes come from a set of connected subgraphs with each node coming from a different layer. These subgraphs are generated based on identical topology. Then, it performs an expansion around the seed by adding to the alignment a node that maximizes the current score, until no more nodes can be added or the alignment size exceeds the limit.

Koyuturk et al. [75] proposed the MaWISh alignment method using the same technique to build the alignment graph as previous methods. MaWISh proposes a scoring function that quantifies the evolutionary distance of the pair of interactions in the input networks.

Evolutionary information is encoded into the edge weights through the concepts of matches, mismatches, and duplication. A match corresponds to a conserved interaction between two orthologous protein pairs, and duplication is the duplication of a protein in the course of evolution. A node score is assigned based on the sequence similarity of the connected proteins. Then, the alignment problem is formulated into a maximum weight induced subgraph problem. Kim et al. [71] extend this method to work for multiple networks.

The previous methods only examine the direct neighborhood of each node; therefore, PPIs data noise causes them to yield bias results. AlignNemo [26] tries to solve this issue. It uses the concept of weighted alignment graph, in which nodes represent pairs of orthologous proteins, and edges are weighted via a scoring strategy that accounts for both direct and indirect interactions. For each pair of orthologous proteins, the number of short paths connecting them is used to evaluate how likely they are connected in the input network. AlignNemo takes into account the degree of each protein and penalizes paths that are passing through hubs. Then, a seed and extend algorithm is used on the alignment graph to find relatively dense groups of nodes that are the alignment solutions.

Mina et al. [101] propose AlignMCL to extend AlignNemo using the Markov clustering algorithm instead of seed and extend. Markov clustering is a graph clustering algorithm that simulates random walks using Markov chains iteratively. AlignMCL first builds a weighted alignment graph the same way AlignNemo does. Then, it applies Markov clustering to this graph to identify conserved protein modules. Considering the direct and indirect interactions in AlignNemo and AlignMCL reduces the impact of false positives on the construction of the alignment graph, since it is unlikely that many false interactions consistently form short redundant paths between two proteins. However, the mining heuristic implemented in AlignNemo is not scalable for the large size of current PPI networks. AlignMCL is still based on the idea of finding the subgraph as the collection of nodes that are more connected with each other than to the other network nodes.

6.2.2 Information Fusion Methods

In these methods, external information is added to the PPI data for the alignment. For instance, Flannick et al. [41] propose Graemlin to improve over previous methods by using evolutionary information. Graemlin finds with a seed and extend strategy a pairwise alignment of the two closest species based on their phylogenetic relationship. A scoring function

composed of two parts is employed. One part evaluates each equivalence class (a class consists of proteins evolved from a common ancestral protein). Scoring the equivalence classes is based on constructing the most ancestral history of their proteins. This construction is based on sequence mutations, insertions, deletions, duplication, and divergence among proteins in each class. The second part is edge scoring. Each edge is assigned a probability parametrized by its weight and node degree, based on the idea that two nodes of high degree are more likely to interact by chance than two nodes of low degree.

Hu et al. [57] present another method that uses phylogenetic information for the alignment called LocalAli. The method employs the input PPI networks and their proteins BLAST sequence similarity to construct a bipartite graph with interactions and homologous proteins. In the case of multiple alignment, the pairwise bipartite graphs are integrated into a k-layer graph (k is the number of PPI networks). Then, heuristic search is performed for the k-layer graph to find a set of refined seeds, using a seed and extend strategy. The induced subgraphs are set as the leaves of an evolutionary tree, which has the same topology and branch weights as the corresponding phylogenetic tree of the involved species. Using the maximum parsimony principle, the optimal or near optimal inner nodes of the tree are inferred using a simulated annealing algorithm. An alignment score of each resulting subgraph is calculated based on the evolutionary distance, and those scoring less than a threshold are filtered.

Another method that does not rely on building an alignment graph is GASOLINE [99]. It implements a new seed and extend strategy to extract shared complexes among a set of PPI networks. It starts with identifying a set of similar nodes by looking for homologous proteins and builds a set of seeds using a Gibbs sampling algorithm. This step is called the bootstrap phase. Then, it repeatedly either extends or removes nodes in the aligned sub-network, based on maximizing a similarity score. The similarity score for two protein is defined as either the bit score or the inverse of their BLAST E-value. An edge similarity score is based on the structure of its connected proteins. This step is iterated until the local density of the aligned sub-networks increases. The sub-network local density is measured through a defined degree ratio. The algorithm iterates the above steps producing a set of local alignments. Each local alignment consists of a set of similar subnetwork, in terms of both sequence and structure similarity. Finally, they rank each alignment according to an index called the index of structural conservation (ISC).

Seah et al. [134] propose DualAligner to recruit GO annotation information into the align-

ment. DualAligner divides the input networks into biologically related subgraphs. It aligns functional subgraphs of one network to functional subgraphs of another. A functional subgraph is a connected component of the network whose nodes share a particular biological role or function. First, functional subgraphs of the networks are identified. Then, an alignment between pairs of functional subgraphs is carried out, and high confidence protein pairs are identified based on the structural and sequence similarities of their underlying subgraphs.

6.2.3 Other Methods

Pache et al. [111] proposed NetAligner as an online tool to align the user defined query pathways or protein complexes to whole species PPI. The score of the alignment solution is computed as the weighted sum over all nodes and edges scores. A node score is estimated as the probability of the corresponding protein homology using BLAST E-value. Edge score is estimated as the weight of the interaction for its proteins. In addition, there are other works that try to detect functionally conserved sub-networks between species by using a combination of clustering algorithms and global alignment algorithms, such as PINALOG [119].

Luqman et al. [52] propose the PageRank-Nibble algorithm for local network alignment. The algorithm partitions one of the two input networks and maps these sub-networks to the other network. Then, a local extension is implemented to detect the connected components that consist of the homologous proteins in the other network. Using these connected components, the sub-networks are refined and the connected parts in them are extracted as conserved sub-networks.

Manikandan et al. [108] propose a match and split algorithm for aligning two networks. The method matches proteins of two networks according to a matching criterion, then splits the whole networks into connected components. It repeats this process recursively on those connected components and finally outputs the conserved sub-networks.

Current methods to network alignment suffer from several limitations. For instance, the heuristics used to speed up the alignment are coded into the implementation of the algorithms and are not easy to replace or modify specific components (e.g., the scoring function used for matching nodes across networks) of the alignment algorithms to meet the need for specific applications, such as transfer of biological knowledge across species [37] or aligning

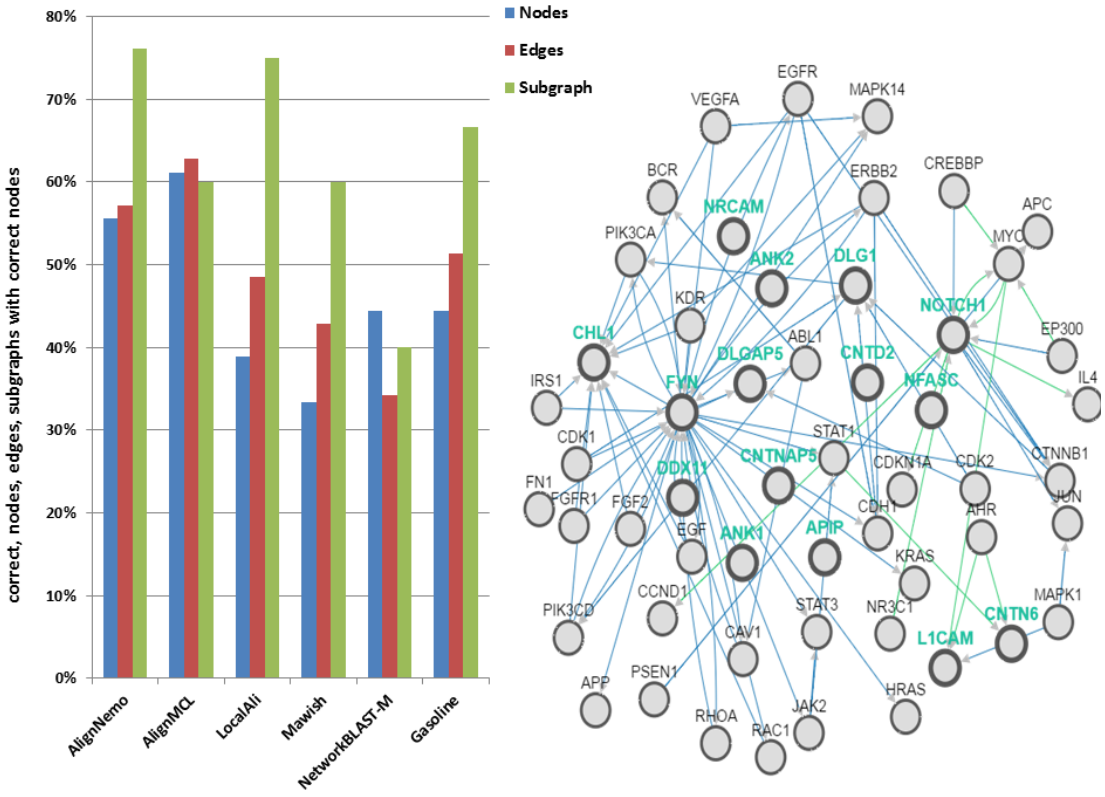

Figure 6.1: Evaluation analysis between the current methods on curated PPI that we know the real alignment in them between mouse and rat species, nodes with green colored name are the known conserved nodes.

networks that model multiple types of interactions between multiple types of molecular entities [140]. Also, some of the algorithms because of computational considerations, make some simplifying assumptions that are biologically inaccurate [36]. Because of network differences in edge densities and noise levels, methods that align one set of networks correctly might align another set of networks from a different database inaccurately. Another limitation is that the existing local alignment methods convert the problem of matching conserved nodes into grouping similar nodes into modules, and the heuristics used usually result in very different solutions. We have made a comparative study among five LNA methods to test their performance on two small networks with known conserved protein and interactions. Figure 6.1 shows the evaluation analysis that we made. We have curated two networks of 54 proteins and 240 interactions for mouse and rat. There are experimentally known 30 proteins and 158 interactions in each network to be conserved between the two species.

Chapter 7

DONA: Identifying Conserved Protein Complexes

Previous studies have shown that cross species protein-protein interactions (PPIs) comparison can uncover evolutionary related protein complexes. As PPI data accumulate, the challenges of identifying conserved protein complexes from PPIs have become very difficult. The purpose of our research here is to develop a new approach for identifying conserved protein complexes between two species. Unlike previous methods, we develop a machine learning approach that takes domains conservation of the PPIs into account. This allows us to enhance the accuracy of the predictions.

In this research, we developed DONA (Domain-Oriented Network Aligner), a new approach that detects conserved protein complexes between different species via local network alignment. This chapter gives a detailed description of DONA and its results. First, an identification of the problem is given, followed by a detailed description of the proposed approach. Finally, DONA results are analyzed to measure and compare its performance with the existing methods.

7.1 Problem Definition

A PPI network is represented as an undirected graph $G = (V, E)$, where V denotes the set of proteins, and $(u, v) \in E$ denotes an interaction between the two proteins $u, v \in V$.

The objective is to identify small and well defined units, such as protein complexes, that are similar between two PPI networks. Local network alignment is an effective way to comparatively analyze a pair of networks for conserved protein complexes discovery. In this section, we formally define the network alignment problem.

Local alignment seeks small sub-networks that are similar or conserved between the two networks, emphasizing regions of high confidence alignment. Conservation of sub-networks is measured in terms of similarity in protein homology (node similarity) and similarity in interactions patterns (network topology similarity). The local network alignment problem is related to the subgraph isomorphism problem and is NP-hard, which suggests the use of heuristics.

Given two PPI networks represented as graphs $G = (V, E)$ and $H = (U, W)$, the similarity between a pair of proteins, one from each network, can be defined by a similarity function $S : V \cup U \rightarrow R$. For any $u, v \in V \cup U$, $S(u, v)$ measures the degree of confidence in u and v being similar (homologous), where $0 \geq S(u, v) \leq 1$. We discuss the technique for measuring this similarity score for our approach in Section 7.2.3. A protein subset pair $P = (U', V')$, where $U' \subset U$ and $V' \subset V$, induces a pairwise local alignment $A(G, H, S, P) = (M, N)$ between networks G and H with respect to S. M is the set of matches, and N is the set of mismatches. A match corresponds to a conserved interaction between two orthologous protein pairs, which is rewarded by a match score that reflects the confidence in the conservation of this interaction. On the other hand, a mismatch is the lack of an interaction in the PPI network of one specie between a pair of proteins whose orthologs interact in the other organism. The biological analog of mismatch may correspond to PPIs data noise, the removal of a previously existing interaction in one of the species, or the appearance of a new interaction.

7.2 The proposed approach

With the purpose of applying network alignment to find conserved protein complexes from PPI networks, the network alignment problem is handled in our approach as a graph construction and search problem to find the similar sub-networks between two different species. This section explains our proposed approach, DONA, in detail.

7.2.1 DONA framework

Our approach is inspired by the analysis of yeast and human network conservation that was performed by et. al. [95], who discover that many cellular mechanisms have in fact evolved many fold in complexity, while several proteins in these mechanisms are conserved by sequence similarity, there are others that are unique to human. These unique proteins perform similar functions as their conserved counterparts but do not show high sequence similarity to any of the yeast proteins. An extensive investigation reveals that these proteins in fact contain conserved domains, for instance the BRCT domain which is present in yeast RAD9 and human hRAD9 proteins and is also present in the human BRCA1 and 53BP1 (non-conserved according to sequence similarity).

Therefore, integrating information on domain conservation can help to identify considerably conserved protein complexes more efficiently. To achieve this, we integrate multiple data sources to build an alignment graph among the input PPI networks. Rather than explicitly restrict our attention to align homologous proteins, we decomposes PPI networks in terms of their domains and employ their conservation along with PPI data to construct an alignment graph.

The general framework for our approach, DONA, is described in Figure 7.1. The local network alignment process of DONA is divided into four steps. First, the proteins of the two input PPI networks are mapped to their domains. Second, an alignment graph is constructed. The nodes of the alignment graph represent orthologous proteins between the two input networks that share one or more domain. The alignment graph has three types of edges: composite, simple-direct, and simple-indirect. Third, edges and nodes of the alignment graph are assigned weights. Fourth, DONA clusters the alignment graph with the MCL algorithm. The clustering results are extracted as the conserved subnetworks between the input PPI networks.

7.2.2 Alignment graph Construction

Here, the PPI network is represented as the graph $G = (V, E)$, whose nodes V are proteins and edges E are interactions among them, and domain-domain interactions data are represented as a graph $H = (D, I)$ with nodes D as domains and edges I are domain interactions. Given two undirected graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ corresponding to the pair of

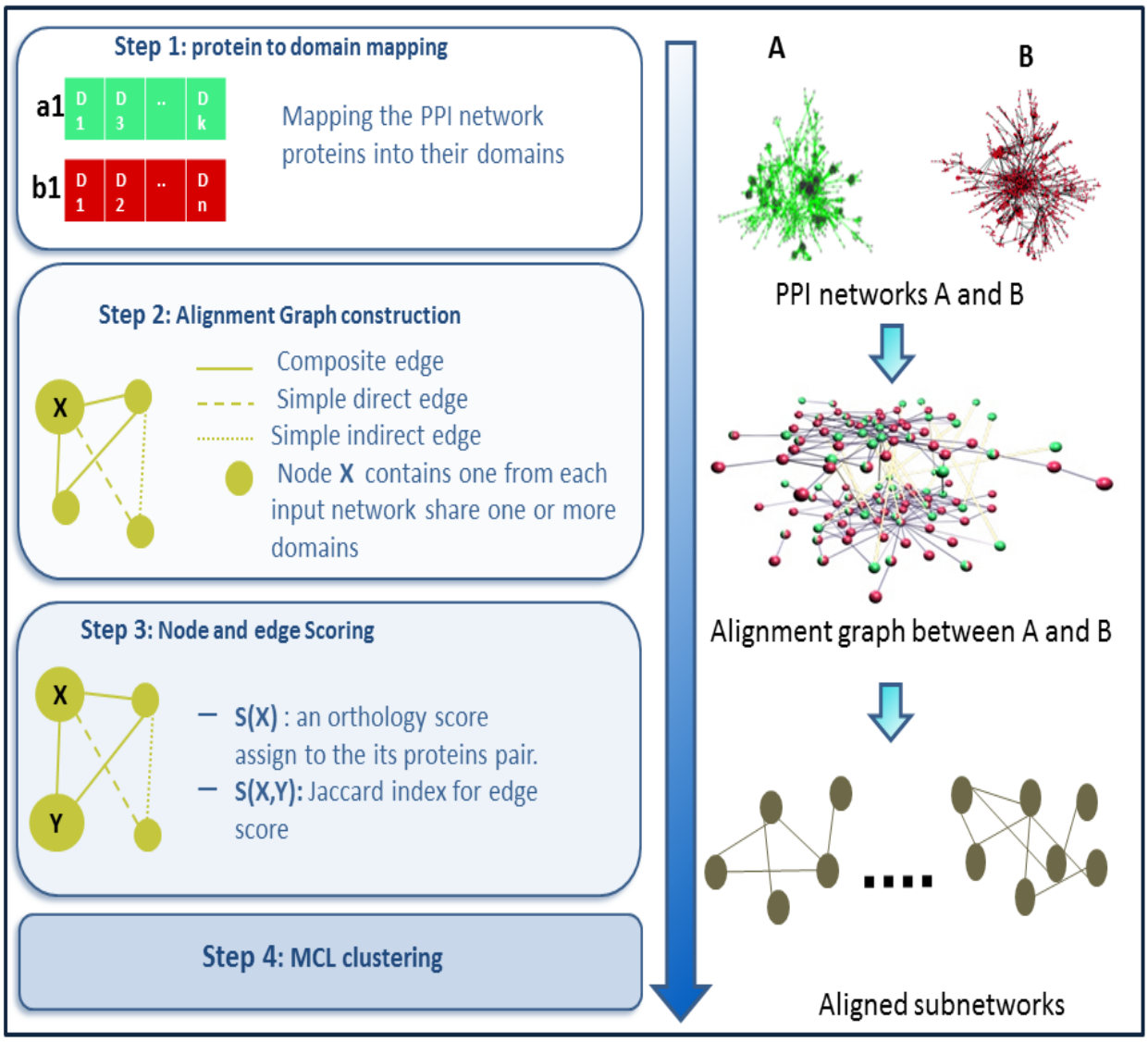

Figure 7.1: The general framework for DONA. Given two input PPI networks; (i) mapping the network proteins into their domain using Pfam database is performed, (ii) the alignment graph is built, (iii) scores are assigned to its nodes and edges, (iv) and the alignment graph is clustered.

input PPI networks belonging to two species, V_1, V_2 denote the node sets, E_1, E_2 denote the edge sets of the graphs. Let $M = \{(u, v, d), u \in V_1, v \in V_2, d \in D\}$ be the mapping between the nodes of G_1, G_2 and domains $d \in D$ of H. We aim to build an alignment graph that takes into account the structure of the input PPI and DDI networks.

Our approach first constructs an alignment graph of the input networks G_1, G_2 and H. The

purpose of the alignment graph is to merge all input data into a single graph. Nodes in the input networks are aligned based on their protein domains from mapping M. We say that a pair of nodes $v_i \in V_1$ and $v_j \in V_2$ is alignable if there exists a domain $d \in D$ shared between the proteins of these nodes. Each node n_l in the alignment graph $A = (N, E)$ contains an alignable pair (AP) of proteins, one node from each input network. In other words, we have a node in the alignment graph for each alignable pair in the original networks.

The alignment graph contains three type of edges, composite, simple-direct, and simple-indirect edges:

- A composite edge (CE) represents an edge between a pair of nodes n_1 and $n_2 \in N$ with both domain-domain interactions between their proteins' domains as well as protein-protein interactions. DONA allows an indirect match in one of the PPI network with the condition that the DDI is direct. This means that a composite edge connects two nodes even if there is one path of length less than or equal to 2 between the two nodes in one of the input PPI network as long as there exist a DDI between the proteins.
- A simple-direct edge (SDE) represents an edge between a pair of nodes n_1 and $n_2 \in N$ with a direct PPI between their nodes in the input networks of both species when no domain interactions can be found between their domains .
- A simple-indirect edge (SIE) is an edge between a pair of nodes n_1 and $n_2 \in N$ with a direct PPI interaction in one species and an indirect PPI interaction in the other species.

Figure 7.2 illustrates the three types of edges in our alignment graph. For simple-indirect edges, we also consider both direct and indirect proteins interactions, as a simple edge is put between two nodes in the alignment graph if the corresponding nodes have protein interactions with path length two. We choose the path length to not be greater than 2 for two reasons. First, adding edges only between directly connected node pairs is not robust against the false positive and false negative interactions in the original PPI networks, and it also does not support aligning the distantly related species. Second, considering edges between node pairs at a path length greater than 2 will increase the number of edges of the alignment graph.

Our analysis shows that the idea of using paths with length 2 for composite and simple-indirect edge improves the result, while using a path with length greater than 2 does not

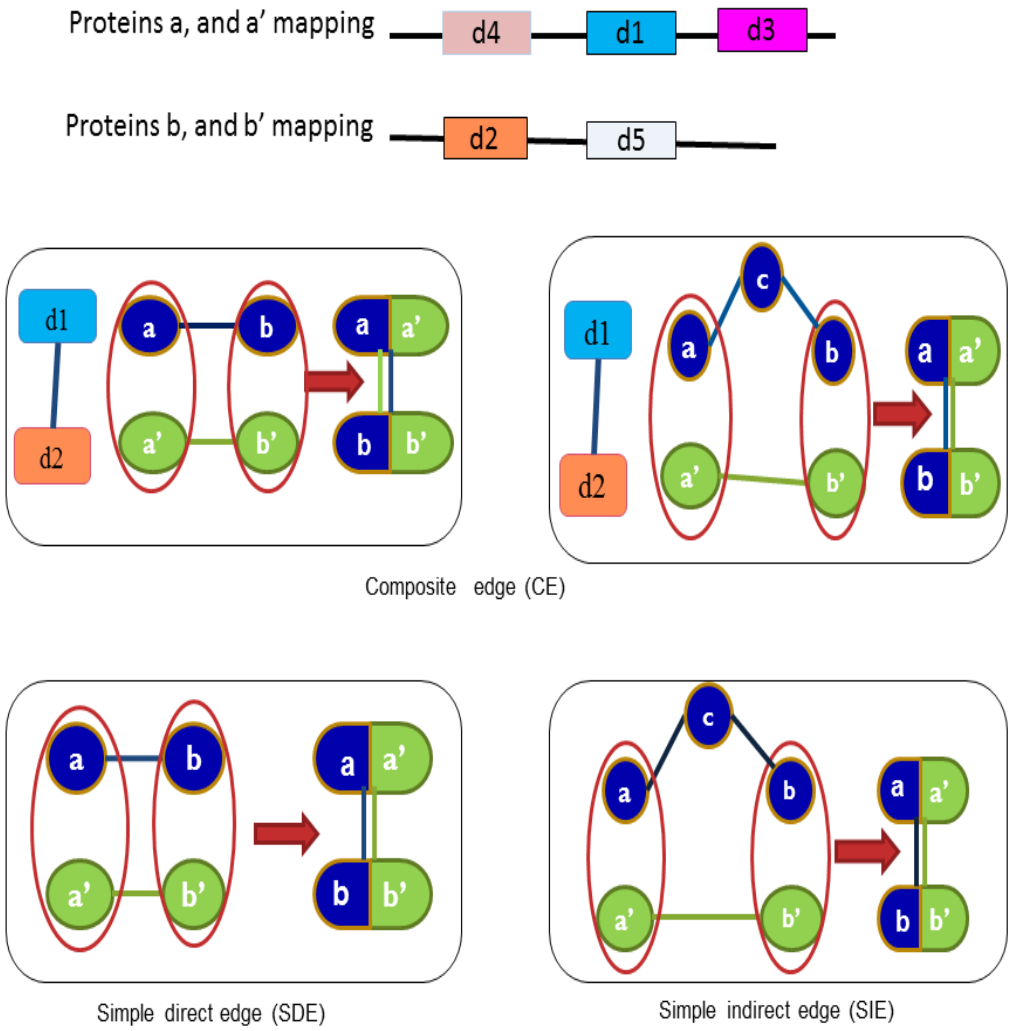

Figure 7.2: The types of edges in DONA alignment graph.

benefit the quality of results. These paths (indirect paths) have a major role in pinpointing the missing interactions in the input PPI networks. As not all of the indirect paths have the same importance, the existence of DDIs for composite edges provides evidence for the interaction of the proteins through their domains. In a simple-indirect edge, if the nodes with path length equal 2 have highly interacting proteins then the probability that there is a missing edge in the PPIs is high.

Formally, the alignment graph can be defined as a graph

$$A(H_1, H_2, M) = (N_A, E_A)$$

That has the following set of nodes:

$$N_A = \{(u, v, d) \in M\}$$

Each edge between two nodes in the alignment graph defines by one of the following cases:

i Composite edge

$$E_A(i,j) = \begin{cases} i = (u, v, d_1), j = (x, y, d_2) \in E_A, \&(d_1, d_2) \in I(u, x) \in E_1 \& (v, y) \in E_2. \\ i = (u, v, d_1), j = (x, y, d_2) \in E_A, \&(d_1, d_2) \in I \& (u, x) \in E_1 \| (v, y) \in E_2. \end{cases}$$

ii Simple-direct edge:

$$E_A(i,j) : \{i = (u, v), j = (x, y) \in E_A, \&(u, x) \in E_1 \& (v, y) \in E_2\}.$$

iii Simple-indirect edge:

$$E_A(i,j) : \{i = (u, v), j = (x, y) \in E_A, \&(u, x) \in E_1 \| (v, y) \in E_2\}.$$

The first case defines the composite edges. The next two cases define the simple-direct and indirect edges. The alignment graph construction goal is to consider the structure of the two PPI networks and the DDIs. We proposed a new scoring scheme for the edges of the alignment graph that incorporates topological information present in the original networks and DDIs data. The next section explains the alignment graph nodes and edges scoring.

7.2.3 Scoring the alignment graph

The alignment graph resulting from the above step is an unweighted graph. Each edge is weighted according to a scoring technique that incorporates the conservation and local significance of the interactions in the input PPI and DDI networks. The nodes of the alignment graph correspond to an alignable protein pair, and weight with an orthologous scores from. In this section, we briefly explains the scoring strategy that is used for measuring weights for each node and edge of the alignment graph.

Node scoring

To score the nodes of the alignment graph, we determined lists of orthologous proteins for all species combinations using the DIOPT [58] database version 5.3. DIOPT predicts putative orthologous proteins among various species. It use both phylogeny-based algorithms such as Compara and Phylome, and sequence similarity techniques such as InParanoid and orthoMCL to measure proteins orthology. Then, we estimate DIOPT scores for each alignable pair (AP) of the proteins in the nodes of the alignment graph.

Edge scoring

To score the alignment graph edges, we utilize a scoring strategy using the Jaccard index. The Jaccard index is a common similarity measure in information retrieval [85] that can be used to compute the similarity between two sets. It measures the probability that two variables x and y have a feature f_i, for a randomly selected feature f that either x or y has.

In DONA, Jaccard index is estimated as the proportion of the shared interactions between two nodes relative to the total number of interactions connected to them. Each edge in the alignment graph is scored based on the number of paths of length less than or equal two that connect its proteins in the input networks. Scores from domain interaction data are also considered for the composite edges.

The Jaccard index score of the edge $e(n_1, n_2)$ between two nodes in the alignment graph n_1 and n_2 is estimated by adding two terms, scores from direct paths and indirect paths in the input networks:

- For direct paths, the score is estimated as the ratio of the direct interactions that connect proteins of n_1 and proteins of n_2 in the input PPI networks divided by the number of all the direct interactions connecting proteins of n_1 or proteins of n_2 to any other node in the PPI network.

- For indirect paths, the score is estimated as the the ratio of the paths of length 2 that connect proteins of n_1 and proteins of n_2 in the input PPI networks divided by the number of all the paths of length 2 that connect the proteins of n_1 or proteins of n_2 to any other node in the PPI network.

We use the Jaccard index score for both direct and indirect paths to account for the local structure of the input networks and the significance of the aligned nodes.

If we have node n_1 containing an alignable protein pair (x, u) and the node n_2 containing an alingable proteins pair (y, v) in alignment graph, where $x, y \in G_1$ and $u, v \in G_2$. Let $P(x)$ be the number of paths of length k connecting the node x to its neighbors, and $P(y)$ be the number of paths of length k connecting the node y to its neighbors in the first input PPI network G_1. Let $L(u)$ be the number of paths of length k connecting the node u to its neighbors, and $L(v)$ be the number of paths of length k connecting the node v to its neighbors in the second input PPI network G_2.

Then a score estimated for every k as

$$S_k(n_1, n_2) = \frac{P_k(x) \cap P_k(y)}{P_k(x) \cup P_k(y)} + \frac{L_k(u) \cap L_k(v)}{L_k(u) \cup L_k(v)}.$$

As DONA calculated the edge score with $k = 1, 2$, the final score for the edge that connects n_1 and n_2 in the alignment graph is

$$S_f(n_1, n_2) = \sum_{k=1}^{2} S_k(n_1, n_2).$$

For composite edges, the existence of domain interactions strengthens the evidence for conservation of the protein interactions. To reflect the presence of the domain interaction on the composite edge score, we estimated a score for the interaction between the domains d_1 and d_2 in the DDI network $H = (D, I)$ also using Jaccard index as

$$JI(d_1, d_2) = \frac{E(d_1) \cap E(d_2)}{E(d_1) \cup E(d_2)},$$

where $E(d_1)$ is the number of paths connection the domain d_1 to its neighbors, and $E(d_2)$ is the number of paths connection the domain d_2 to its neighbors. If the edge has the domain interaction (d_1, d_2) (composite edge), then its score estimated as

$$\bar{S}_f(n_1, n_2) = S_f(n_1, n_2) + JI(d_1, d_2).$$

Once the alignment graph is constructed and weighted, the next step is to search this graph for conserved sub-networks.

7.2.4 Alignment graph Search

The next step for local network alignment after constructing the alignment graph is to search this graph to detect conserved protein complexes. This process is computationally difficult. Current methods propose heuristic search algorithms such as seed-and-extend. With the increase in size of PPI data in recent years, these heuristics algorithms are not scalable. Moreover, there is no mathematical definition to detect protein complexes from PPI networks, but it has been observed that proteins within a complex interact closely with each other. Therefore conserved protein complexes among different PPI networks mostly exist in the dense regions of the PPI networks [6].

Therefore, the problem of identifying conserved protein complexes is reduced to the problem of identifying high scoring subgraphs of the alignment graph. We propose to use the Markov cluster algorithm (MCL) [147] as a scalable approach to uncover the conserved complexes between the input PPI networks.

Markov Clustering Algorithm

The Markov cluster algorithm simulates a stochastic flow on graphs that resembles a set of random walks. The algorithm was proposed by Stijn van Dongen [147]. It is based on the idea that a region with many edges forms a cluster and the amount of flow within a cluster is stronger than the amount of flow between clusters. A cluster resulting from the algorithm is a collection of nodes that are connected to each other more than to the other nodes of the graph. MCL starts with a set of random walks within the whole graph to strengthen the flow where it is already strong and weaken it where it is weak. During these walks, the cluster structure eventually become visible, and the walks are ended when the clusters with strong internal flow are separated by boundaries having hardly any flow.

MCL simulates the walk or flow as a combination of simple algebraic operations on the stochastic matrix associated with the input graph. The first operation, called expansion, corresponds to normal matrix multiplication of a random walk matrix and models the extension of the flow as it becomes more homogeneous. The second algebraic operation, called

Algorithm 3 DONA approach pseudocode for Alignment graph construction.

Input: Given 2 PPI network $G_1(V_1, E_1)$, $G_2(V_2, E_2)$ and DDI network $H(D, I)$

Output: The alignment graph $A(N, E)$

Map the V_1 and V_2 in to D, proteins $\leftarrow$ domains

if $x \in V_1$ and $y \in V_2$ have $d_l \in D$ **then**
    $n_{x,y} \in N$
end if
Construct $A(N, E)$

for nodes $n_i, n_j \in N$ **do**
    search input network $G_1(V_1, E_1)$ and $G_2(V_2, E_2)$
    if $n_{x,u}, n_{y,v} \in N$ and there is $e(x, y) \in G_1$ and $e(u, v) \in G_2$ **then**
        $e(n_{x,u}, n_{y,v}) \in E$
    end if
end for
for nodes $n_i, n_j \in N$ **do**
    search input network $H(D, I)$
    if $e(d_l, d_2) \in D$ connect d_l of n_i, n_j **then**
        edge $e(n_1, n_2)$ is CE
    else
        Edge $e(n_1, n_2)$ is SDE or SIE
    end if
end for
Return $A(N, V)$

inflation, is a Hadamard power followed by a diagonal scaling of another random walk matrix. It models the contraction of the flow as it becomes thinner in regions of lower current and thicker in regions of higher current. Expansion and inflation are implemented sequen-

Algorithm 4 DONA approach pseudocode for scoring the alignment graph.

Input: Alignment graph $H(N, E)$

Output: Weighted $H'(N, E)$

Score $H(N, E)$

for nodes $n_i \in N$ search input network **do**

  score n_i by orthology score

  if $e(d_l, d_2) \in D$ connect d_l of n_i, n_j **then**

    $S'_f(n_1, n_2) = S_f(n_1, n_2) + JI(d_1, d_2)$

  else

    $S_f(n_1, n_2) = \sum_{k=1}^{2} S_k(n_1, n_2).$

  end if

end for

tially which causes the flow to extend within clusters and fade or disappear between clusters [34]. As these two operation are repeated, the initial distribution of flows becomes more non-uniform, and terminate when a steady state is reached. In an extensive comparison by Brohee and van Helden [17] between MCL and other graph clustering algorithms like RNSC [6] and MCODE [72], MCL out-performs other clustering algorithms in different conditions.

The inflation level r is the most important parameter of MCL. It represents the exponent used in the Hadamard powering operation. Changing the inflation parameter leads to finding clusters with different scales of granularity. Using a high inflation level deceases the average dimension of clusters, since the inflation step will increasingly penalize weaker flows. For weighted graphs, edges weights are considered when the first stochastic matrix is used in the iterative process. In our approach, we used the MCL implementation by van Dongen [147]. The weights of the alignment graph edges are taken into account in first stochastic matrix. From our analysis, we found that the best performance for DONA is achieved when the inflation is between 2.6 and 3.2, see Section 7.3.5 for more details on the effect of the inflation level change on the performance of our approach.

Algorithm 5 DONA approach pseudocode for Alignment graph clustering.

Input: Alignment graph $H(N, E)$

Output: Output clusters

1: Set inflation the parameters $r = 2.8$

2: MCL clustering for graph $H(N, E)$

i $A = A + I$ //add self loop to the vertices

ii $M = AD^{-1}$ // M is the canonical flow matrix

iii REPEAT

i Expand: $M := M * M$

ii Inflate: $M := M.^r$, re-normalize columns.

iii Prune: Saves memory by removing entries close to zero.

iv UNTIL M converges

v interpret M as the resulting clusters

Implementation

Our approach is implemented in two parts. The first one processes input PPI networks, DDIs data, and orthologous data to create the weighted alignment graph. This part is implemented with Python. The second part is the MCL clustering algorithm implemented in C++.

7.3 DONA Results

In this section, we evaluate the performance of DONA with five existing methods, AlignMCL, NetworkBLAST, Mawish, LocalAli, and DualAligner on data sets of five different species. We ran these methods on the same data sets, and for each method, we identify a set of solutions. Then, the solutions from each method are evaluated and compared.

7.3.1 Data sets

We combined multiple PPI data sets to enhance the coverage of PPI networks. In particular we built extensive data sets of PPI networks for five species: *Drosophila melanogaster* (fly), *Saccaromices cerevisiae* (yeast), *Homo sapiens* (human), *Rattus norvegicus* (rat), and *Mus musculus* (mouse). Up-To date PPIs have been downloaded from the STRING [142] database and combined with i2D version 2.9 [18] and BioGRID [24] Release 3.4.145 data, with self interactions or repeated interactions removed. These databases integrate several data sources to build more complete and reliable networks from high throughput experiments, such as yeast two-hybrid (Y2H) assays or affinity purification coupled to mass spectrometry (AP/MS).

For mapping the proteins in each species to their domains, we use the Pfam [39] database version 29.0. We chose Pfam because it is the largest protein domain database. Then, for the proteins that have no record in Pfam, we use CDD [93]. The 3DID [103], Domine [123], and iPfam [40] databases contain a large number of domain interactions. They differ slightly in their DDI definition, and therefore they overlap in only about 70% of the DDIs. We combine the DDIs data from these databases and filter the interactions that do not exist in at least two of these databases. Statistics for the PPI networks and DDIs data are reported in Table 7.1.

For scoring the nodes of the alignment graph, we downloaded the score for the putative orthology associations between proteins of each node in the different species from DIOPT (Integrative Ortholog Prediction Tool) [58]. Some of the evaluation algorithms require BLASTP [98] data, we performed a BLASTP sequence alignment between the proteins of the different species. We used the default parameters of BLASTP. We perform proteome-wide all-against-all BLASTP searches with $E - value \leq 10^{10}$ and considered only hits in the top ten of the BLASTP output.

Table 7.1: Statistics of PPI networks used.

	PPIs data		**DDIs data**	
Species	Proteins	Interactions	Domains	Interactions
Human	47,625	120,560	9,900	15,634
Mouse	8,726	20,898	5,163	8,229
Rat	7,028	16,837	4,062	7,166
Yeast	4,928	15,528	4,349	9,194
Fly	7,446	11,013	2,948	8,465

Table 7.2: The number of complexes available in databases for evaluating DONA.

Species	Database	No. of Complexes
Human	CORUM	1043
Mouse	CORUM	330
Rat	CORUM	251
Yeast	CYC2008	399
Fly	DroID	356

Protein Complex data set

To detect conserved protein complexes, we need a benchmark data set to compare our results with. We retrieved the known complexes for each species from databases that identify complexes from small scale experiments and literature mining. Table 7.2 shows the data set of protein complexes we used for the five species in our study. These databases are CORUM [131] for human mouse, and rat complexes, CYC2008 [122] for yeast, and DroID [164] for fly. We noticed that around 25% of CYC2008 and CORUM complexes have complexes with size less than 3 proteins. Such small complexes might lead to biased statistical measures, since one solution can overlap with more than one complex and hence be counted more than once. Therefore, we restrict our analyses to protein complexes that have at least 3 proteins.

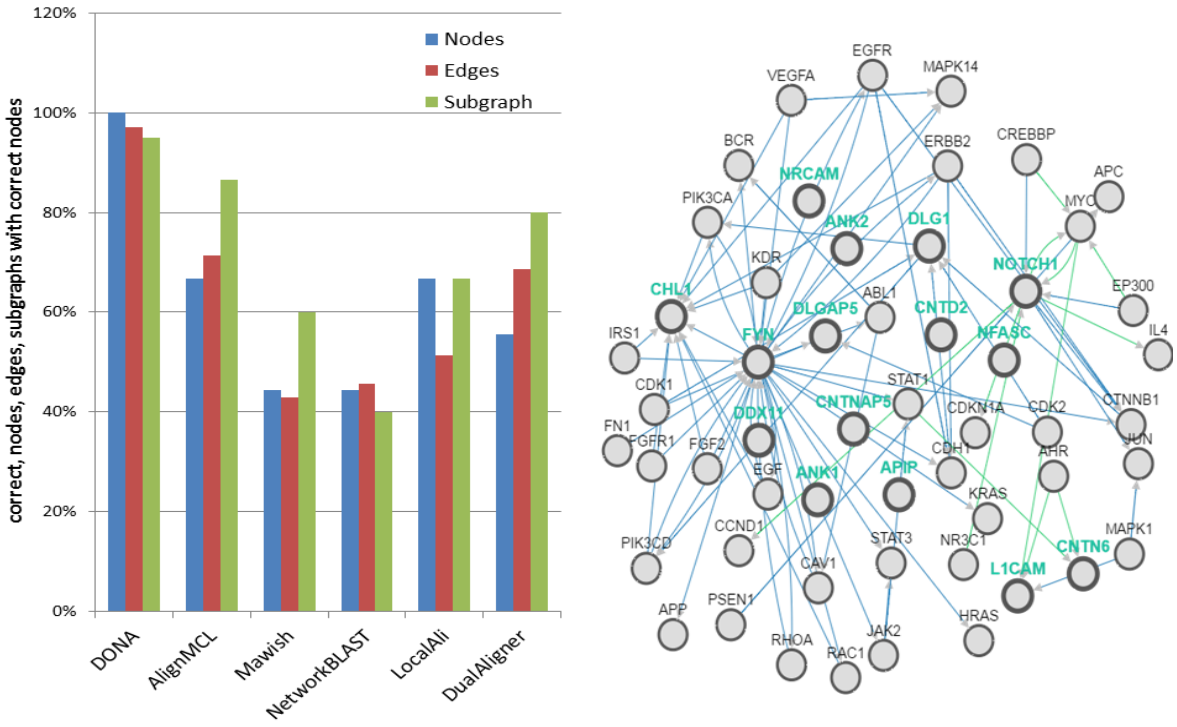

Figure 7.3: Comparing our approach DONA with the existing approach in a case study.

7.3.2 Case study

We have curated two networks of 54 proteins and 166 interactions for both mouse and rat. In this small network, there are experimentally known to be 31 proteins and 98 interactions in each network conserved between the two species. Figure 7.3 shows the performance of DONA compared with the other methods in term of the number of conserved proteins identified, the number of conserved interactions and the number of solutions that identify the known conserved sub-network or subset of it. We found that DONA out-performed the other methods as it is able to identify all the conserved proteins and 96 out of the 98 conserved interactions. Also DONA generates a sub-network as one of its solutions that contains all the known conserved proteins.

7.3.3 Comparison with other methods

We evaluated DONA performance over the extensive data sets we created in Section 7.3.1, to avoid over-fitting and examine its performance in different alignments. Table 7.3 shows

Table 7.3: Each cell shows the symbol used to represent the different alignment throughout the chapter.

Species	Human	Mouse	Rat	Yeast	Fly
Human	-	H-M	H-R	H-Y	H-F
Mouse	H-M	-	M-R	M-Y	M-F
Rat	H-R	M-R	-	R-Y	R-F
Yeast	H-Y	M-Y	R-Y	-	Y-F
Fly	H-F	M-F	R-F	Y-F	-

the symbols used to represent the different alignments throughout the chapter. We compare DONA performance with five LNA methods: AlignMCL, Mawish, NetworkBLAST, LocalAli, and DualAligner. Each of these methods is executed on the same data set for each alignment. There are other local alignment methods that are not taken into consideration in our assessment. For instance, the current Graemlin [41] version is outdated and does not compile, and CAPPI [30] was only compatible for particular design. After performing DONA and the other methods on the data sets, we obtained a set of solutions from each method. Table 7.4 presents the number of solutions produced for each alignment from the different methods.

Known complex detection

Since the goal of DONA is to discover conserved protein complexes, it is essential to evaluate how well its solutions produced known protein complexes in the aligned species. Given a solution and a known complex, we measures the overlap between the solution and the complex using two measurements; precision p and recall r. Precision is defined as the fraction of proteins in the solution that are also present in the complex. Recall measures the ratio of proteins in the complex that are in common with the solution. Then, we integrate these two measures into F-score to measure the harmonic mean of precision and recall. These measures are defined as follows

$$p = \frac{TP}{TP + FP}$$

Table 7.4: The number of solutions produced for each alignment in the different methods.

Alignment	**Number of solutions**					
	DONA	AlignMCL	Mawish	NetworkBLAST	LocalAli	DualAligner
M-R	854	805	830	725	267	561
H-M	965	830	1057	934	693	756
H-R	1020	750	1161	1014	203	646
H-Y	1220	941	890	820	498	772
H-F	845	701	724	861	630	823
M-Y	952	834	563	620	491	410
M-F	734	530	400	650	528	340
R-Y	930	632	530	767	501	298
R-F	701	439	529	498	320	256
Y-F	873	752	630	567	431	398

$$r = \frac{TP}{TP + FN}$$

where TP (true positive) is the number of proteins found in the solution that are also in the complex. FP (false positive) is the number of proteins in the solution that are not in the complex. FN (false negative) is the number of proteins in the complex that are found in the solution. And F-score estimated as

$$F - score = \frac{2p * r}{p + r}$$

The F-score value range is $[0, 1]$, with 1 represent a perfect match between the solution and the complex.

First, we match each known complex of a species to all the solutions of a given alignment, and we select the best matched solution with its F-score. Then, we compare DONA performance with other methods in terms of each approach's ability to identify the known protein complexes in the two aligned species. To assess our approach robustness we considered the degree of variation of the number of complex hit over 20 runs for DONA and AlignMCL

Table 7.5: The number of known complexes hit with F-score 0.3 in the different methods, and standard error over 20 runs for DONA and AlignMCL, the number in parentheses.

Alignment	**Number of Complexes** $F-score=0.3$					
	DONA	AlignMCL	Mawish	NetworkBLAST	LocalAli	DualAligner
M-R	143 (0.02)	103 (0.05)	48	25	85	52
H-M	130 (0.038)	123(0.1)	29	15	63	65
H-R	170 (0.3)	97 (0.05)	76	21	72	41
H-Y	112 (0.08)	96 (0.4)	88	23	30	35
H-F	88 (0.1)	89 (0.5)	72	21	66	54
M-Y	113 (0.04)	92 (0.1)	45	69	78	61
M-F	78 (0.09)	65 (0.3)	40	54	28	37
R-Y	93 (0.1)	63 (0.4)	34	48	42	39
R-F	89 (0.05)	67 (0.12)	49	43	32	55
Y-F	139 (0.07)	92 (0.02)	56	42	53	63

as they both use clustering algorithms for alignment graph search. Tables 7.5, 7.6, and 7.7 offer a wide comparison among the different methods for the number of complex hit with F-score cutoff equal 0.3 , 0.5 and 0.07 respectively. In the tables, we list the number of protein complexes found by each method and the standard error for DONA and AlignMCL .

DONA uncovered a higher number of complexes with respect to the other methods with good quality. We observe that AlignMCL and LocalAli behave well on most alignments with low F-score cutoff but have some problems in dealing with the higher F-score cutoff. Both DONA and AlignMCL perform better on closely related species alignment, with the latter having overall higher values of protein complex hit. Even with the large number of solutions found by Mawish and NetworkBLAST, they have in general low precision and fail to recover most proteins in a complex. DONA and AlignMCL have close trend for mouse-yeast and

Table 7.6: The number of known complexes hit with F-score 0.5 in the different methods, and the standard error over 20 runs for DONA and AlignMCL, the number in parentheses.

Alignment	**Number of Complexes** $F - score = 0.5$					
	DONA	AlignMCL	Mawish	NetworkBLAST	LocalAli	DualAligner
M-R	102 (0.01)	97 (0.4)	37	16	41	29
H-M	98 (0.02)	89 (0.01)	18	8	50	61
H-R	84 (0.2)	73 (0.03)	39	18	47	32
H-Y	94 (0.03)	81 (0.01)	41	15	24	35
H-F	47 (0.01)	46 (0.009)	35	13	31	20
M-Y	36 (0.03)	34 (0.01)	36	11	29	41
M-F	43 (0.009)	39 (0.0)	34	27	31	40
R-Y	49 (0.01)	37 (0.4)	14	8	22	19
R-F	32 (0.2)	17 (0.1)	9	6	13	15
Y-F	39 (0.3)	29 (0.08)	11	22	13	23

human-mouse alignments with F-score cutoff equal 0.5. However, the standard error for the change in number of complex hit with 20 runs shows the consistence in DONA performance. We also noticed that, while Mawish performs similarly well for the mouse-yeast alignment with $F - score = 0.3$, the majority of solutions produced by Mawish have small size, most of them consisting of 2 to 4 proteins only.

We analyze the F-score cutoff range for each method. Figure 7.4 summarizes the performance of the 6 methods in term of the number of recovered complexes with different F-score cutoff reveals. The representation used in Figure 7.4 is useful for summarizing how each method is affected by the F-score cutoff in the different alignments. In most cases, DONA achieves better results. In fact, even though DONA and AlignMCL appear to have more resemblance in the number of complex hit DONA achieves better performance with high F-score cutoff.

Figures 7.5 and Figure 7.6 report the performance of DONA, Mawish and NetworkBLAST in terms of precision and recall separately. A positive note is the fact that most DONA solutions are concentrated in the top-right area, while MaWish and NetworkBLAST ones

Table 7.7: The number of known complexes hit with F-score 0.7 in the different methods, and the standard error over 20 runs for DONA and AlignMCL, the number in parentheses.

Alignment	**Number of Complexes** $F - score = 0.7$					
	DONA	AlignMCL	Mawish	NetworkBLAST	LocalAli	DualAligner
M-R	21 (0.05)	19 (0.5)	9	7	11	12
H-M	17 (0.1)	9 (0.0)	3	-	8	11
H-R	18 (0.25)	7 (0.3)	-	1	5	9
H-Y	21 (0.03)	11 (0.1)	-	6	4	5
H-F	20 (0.01)	16 (0.5)	2	-	9	11
M-Y	16 (0.03)	8 (0.01)	-	-	-	5
M-F	15 (0.09)	9 (0.1)	-	7	2	10
R-Y	9 (0.05)	7 (0.4)	7	-	2	-
R-F	14 (0.1)	7 (0.1)	-	-	1	5
Y-F	18 (0.02)	19 (0.5)	1	-	9	3

are more in the bottom-left area. That explains the degrading in their performance with high F-score. The figure show that DONA have a high number of high quality solutions that match known complexes with an F-score greater than 0.5.

7.3.4 Biological relevance of conserved subnetworks

To further validate our approach, we investigate biological relevance between the identified conserved subnetworks, from now on we will call them modules, which is measured by the average of functional similarity among all proteins in them. Functional similarity of two proteins refers to the semantic similarity of their Gene Ontology (GO) annotations [5]. Two measures have been used to evaluate the functional similarity of the aligned modules: purity and GO enrichment. These two measures have been suggested in several LNA studies [49, 116].

A module is called pure if it satisfies two conditions. First, it has to contain at least three

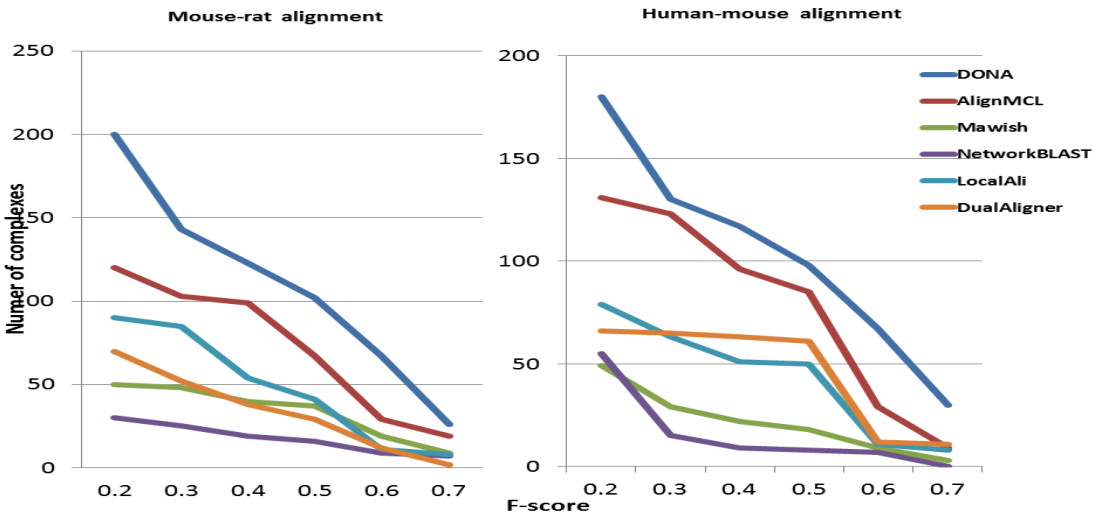

Figure 7.4: Methods comparison based on the change of the predicted complexes with F-score.

annotated proteins in the CORUM database, and, second, the module must cover $\geq 75\%$ of a known complex in CORUM. Purity is computed as the number of pure modules divided by the total number of modules with at least three CORUM annotated proteins. The purity measure uses the known protein complexes from CORUM as the gold standard. Therefore, only mouse-rat, human-mouse, and human-rat alignments are considered here.

GO enrichment measures the functional coherence of the proteins of the identified modules with respect to the molecular function annotation of GO. The GO:TermFinder [16] tool is used to calculate the significance of GO annotations for each identified module. The modules that have one or more enriched GO terms with $p - value < 0.05$ are regarded as functionally coherent modules. For each species, we calculate the fraction of functionally-coherent modules. Tables 7.8 and 7.9 compare the performance of DONA and the 5 other methods in term of the purity and GO enrichment. DONA identified more functionally-coherent modules than the other methods. It achieved the highest score on almost all the evaluation measure in the considered alignments. The quality of DualAligner results is more variable, with few high quality modules in the alignment of mouse-rate. These high quality modules do not emerge when evaluating the other two alignments, suggesting stronger sensitivity to the aligned species.

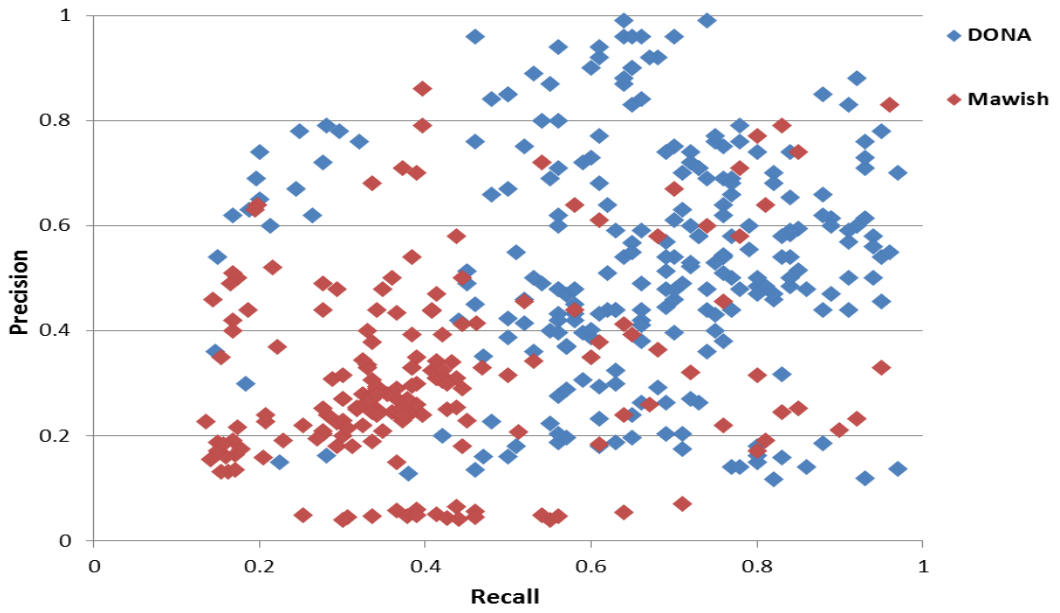

Figure 7.5: Precision and recall for the detected complexes in human-yeast alignment.

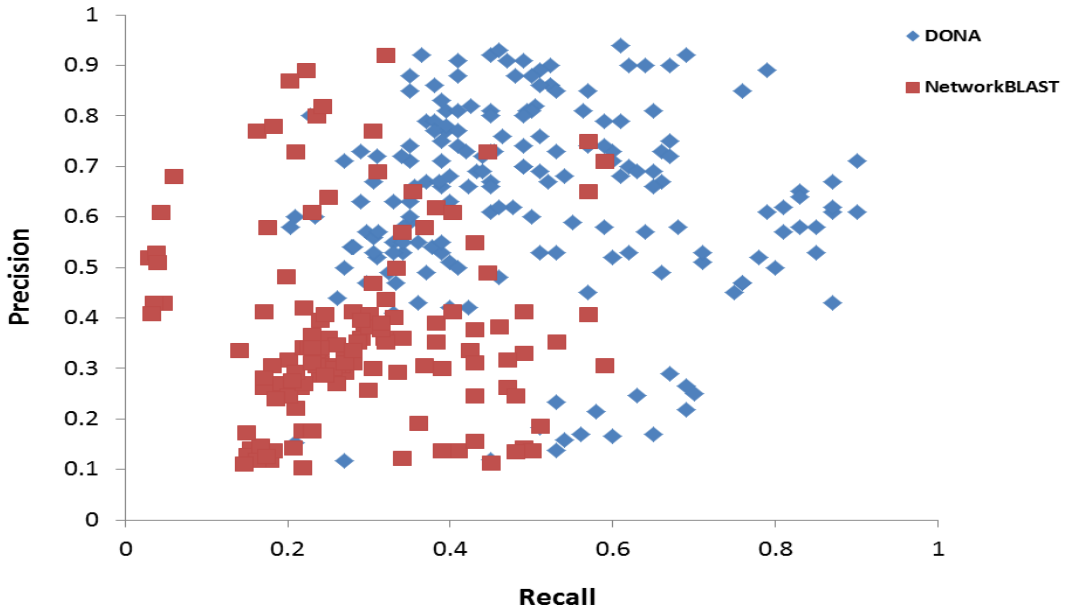

Figure 7.6: Precision and recall for the detected conserved complexes in Mouse-Rat alignment.

Table 7.8: Purity and GO enrichment analysis for mouse-rat and human-mouse alignments.

Method	**mouse-rat alignment**			**human-mouse alignment**		
	Purity %	**GO enrichment**		Purity	**GO enrichment**	
		mouse %	rat %		human %	mouse %
DONA	78.0	94.8	89.0	71.0	84.8	79.0
AlignMCL	66.5	75.3	66.0	59.5	62.3	59.0
Mawish	40.0	69.02	65.8	31.0	59.02	42.8
NetworkBLAST	42.8	63.5	60.9	42.8	40.5	31.9
LocalAli	58.4	81.0	69.2	58.4	53.0	61.2
DualAligner	60.0	81.4	89.0	57.0	72.4	59.0

7.3.5 The effect of MCL parameter on the performance

Inflation parameter regulates the MCL clustering algorithm. The impact of varying the inflation level on the prediction of the conserved complexes is tested here. The best performance is achieved when inflation ranges between 2.6 and 3.2, as DONA is quite stable within this range. When the inflation level is below 2.6, we found quick degradation of the performance, and a slow degradation when the inflation increases over 3.2. Figure 7.7 shows how the inflation level changes the number of protein complex hit in different alignments.

Running time

In comparing DONA running time with the time of the other methods, DONA is the fastest alignment tool. As shown in Figure 7.8, DONA finished all the pairwise alignments within 2 hours using a 2.2Ghz processor with RAM of 12gb. In contrast, Mawish and NetworkBLAST which spent about 8.8 hours on the mouse-rate alignment and 24 hours on the human-mouse alignment. To construct the alignment graph, Figure 7.8-B, DONA is faster than AlignMCL.

Table 7.9: Purity and GO enrichment analysis human rat alignment.

Method	**human-rat alignment**		
	Purity %	**GO enrichment**	
		human %	rat %
DONA	78.0	94.8	89.0
AlignMCL	66.5	75.3	66.0
Mawish	40.0	69.02	65.8
NetworkBLAST	42.8	63.5	60.9
LocalAli	58.4	81.0	69.2
DualAligner	60.0	81.4	89.0

7.4 Discussion

Our approach uses local network alignment based on both PPI and DDI data and leads to several improvements. It produced better results in terms of the agreement with known protein complexes. DONA often provides a more comprehensive means for biologically interpreting the aligned sub-networks, as protein domains are directly related to their proteins function. For the functional coherence of the detected alignments, DONA performs better than other alignment methods. Therefore, recruiting DDIs in the alignment process improves identifying the conservation across species. Also, employing scalable clustering algorithm like MCL improves the results by increasing the solution set size.

Some conserved modules found in human-mouse alignment by our approach have noisy interaction data in their regions in the original PPI networks, thereby reducing their topological significant when identified only by PPI data; adding DDI data helps to identify these modules. See Figure 7.9 for examples of these modules that are identified by DONA while other methods failed to identify them. Their conservation is verified by NetAligner [111]). Moreover, DONA is able to detect conserved protein complexes that might be deemed by other methods to be insignificant.

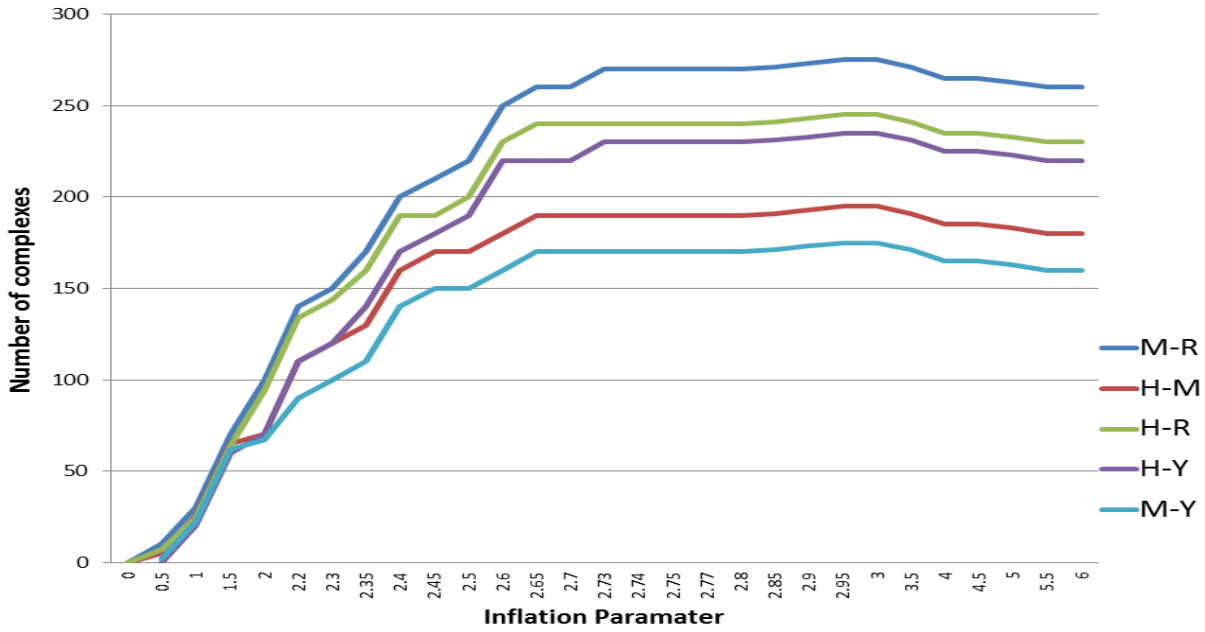

Figure 7.7: Number of complexes detected with different inflation level in different alignment, refer to table 7.3 for the name of the alignment.

An example: Exocyst and F0F1 ATP synthase complexes

Let us focus specifically on a few complexes of CORUM for mouse-rat alignment to better assess the different methods' performance. Here, we discuss two complexes: a small one Exocyst with 8 proteins and a large one F0F1 ATP synthase complex with 17 proteins and many interactions. Table 7.10 shows the number of proteins that have been correctly associated and recovered in the mouse-rat alignment with the precision and recall. DONA is able to identify 7 out of 8 proteins conserved between mouse and rat for the Exosyst complex. Other methods either failed to detect the conservation or only recover a small part of the complex.

Also, GO functional coherence of the aligned proteins in both complexes is higher for DONA than the other methods, indicating an improvement in biological quality. The functional coherence of the F0F1 APT synthase mouse complex proteins is significant, for instance, threonine-type peptidase activity has $P - value \cong 10^{-5}$, and transporter activity has $P - value \cong 10^{-6}$. This complex has not been reported by either Mawish, NetworkBLAST, LocalAli, or DualAligner to be involved in alignment with rat. DONA is able to identify

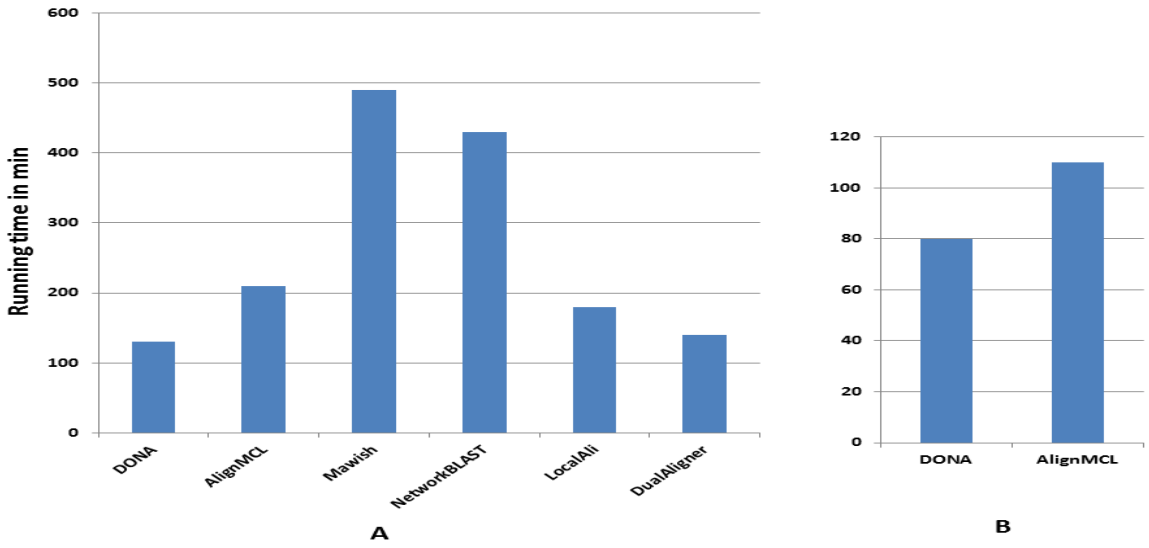

Figure 7.8: Number of complexes detected with different inflation level in different alignment.

13 out of the 17 proteins for this complex, while AlignMCL only identified 7 conserved proteins. DONA solution extends beyond the proteins of F0F1 APT synthase complex due to the high level of interactions of its proteins. To verify the quality of the solution, we search for enriched GO terms of all the proteins in the solution. We found that 20 out of 21 mouse proteins and 18 out of 19 rat proteins in our solution are enriched for the same GO terms with $P - value \cong 10^{-4}$.

An example: Arp 2/3, TFIID, and 20S proteasome complexes

Table 7.11 shows the performance of DONA along with other methods in terms of their ability to correctly identify these complexes in the human-fly alignment. For instance, the Arp2/3 complex contains 7 proteins and plays an important role in the regulation of the actin cytoskeleton [32]. The level of its protein interactions found to be high in human PPI network, while very low in other species especially fly. This incomplete information makes this complex challenging to recover. DONA is able to identify 6 out of 7 proteins of this complex in human-fly alignment, while other methods like AlignMCL only found 2 proteins or failed completely in finding any solution.

Table 7.10: Comparing the best matching solutions for Exocyst, and F0F1 ATP synthase complexes in mouse-rat alignment.

Complex name: Exocyst		**Complex size: 8 proteins**	
	DONA	AlignMCL	DualAligner
Predicted Solution size	7	2	2
Precision	0.5833	0.1428	0.0869
Recall	0.875	0.25	0.25
Complex name: F0F1 ATP synthase		**Complex size: 17 proteins**	
	DONA	AlignMCL	DualAligner
Predicted Solution size	13	7	0
Precision	0.52	0.5833	0
Recall	0.7647	0.4117	0

Table 7.11: Comparing the best matching solutions for Arp 2/3, TFIID, and 20S proteasome complexes in human-fly alignment.

Complex name: Arp 2/3		**Complex size: 7 proteins**	
	DONA	AlignMCL	DualAligner
Predicted Solution size	6	2	0
Precision	0.5833	0.1904	0
Recall	0.8571	0.2857	0
Complex name: TFIID		**Complex size: 13 proteins**	
	DONA	AlignMCL	DualAligner
Predicted Solution size	11	5	2
Precision	0.6875	0.3913	0.2105
Recall	0.8461	0.6923	0.3076
Complex name: 20S proteasome		**Complex size: 14 proteins**	
	DONA	AlignMCL	DualAligner
Predicted Solution size	14	7	6
Precision	1	0.465	0.45
Recall	1	0.715	0.6428

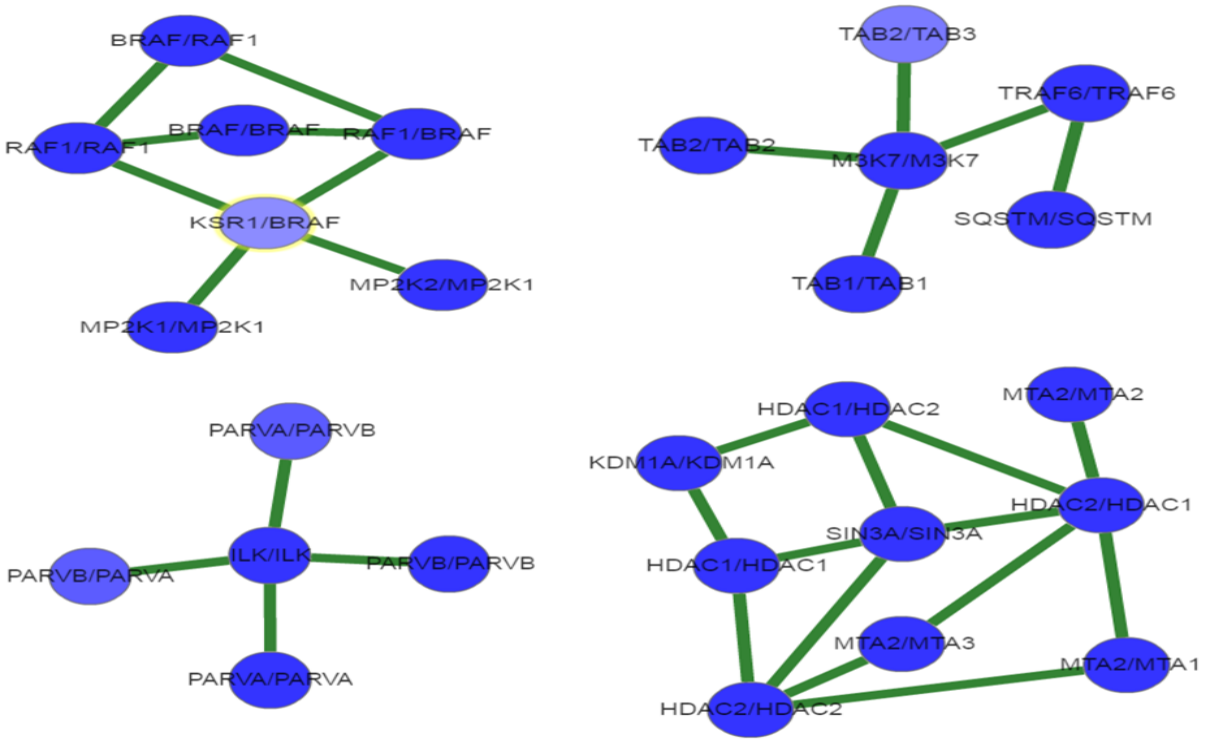

Figure 7.9: Some examples of conserved modules found in human-mouse alignment by our approach. The original PPI networks in these modules regions include several noisy interactions, thereby reducing their topological significant when identified only by PPIs data, adding DDI improve the performance.

Chapter 8

Conclusions and Future Directions

In this chapter, we summarize our contributions for solving the two problems in this dissertation, along with proposed future research directions.

8.1 MicroRNA target prediction

MicroRNAs are small non-coding RNAs. They regulate their target gene by binding to sites located in the 3′-UTR of the transcript. This association results in either cleavage or translation repression of the target, depending on the degree of base pairing between the microRNA and the mRNA. Perfect complementarity results in cleavage, whereas imperfect base pairing leads to translation repression. These alternative effects impose challenges for identifying microRNA targets. Increasing efforts have been made to identify the specific targets of microRNAs, leading to speculation that microRNA may regulate at least 30% of human genes. As the number of identified microRNAs grows, using experimental approaches becomes more limited since these methods are costly and time consuming. Computational methods, on the other hand, can provide a genome-wide prediction of microRNA targets.

During the past decade, many microRNA target prediction methods have been developed. The vast majority of these methods use sequence determinants to predict the target genes of microRNAs. Many performance evaluation studies have shown that current sequence features alone cannot provide accurate prediction of microRNA targets.

It is of great interest to utilize different information sources to discover the regulatory network

of microRNAs. In this dissertation, a new approach, MicroTarget, has been developed for predicting microRNA targets. MicroTarget uses expression data to predict the candidate targets. Then, it focuses on the sequence data to identify the direct targets and their ranking scores. MicroTarget identifies microRNA and mRNA interactions that are believed to be expressed in the same tissue. MicroTarget was applied on an expression data set for human breast cancer. The results show that our approach provides better predictive estimates than those reported by the state-of-art target prediction methods. The main contributions of this dissertation in this domain can be summarized as:

- We take advantage of the expression data profiles for microRNAs and mRNAs, as microRNA and its target have to be expressed in the same tissue to interact.
- Several individual scores were calculated to rank microRNAs targets: (i) thermodynamic stability score based on the free energy estimated of associated between microRNA and its targets, (ii) conservation score based on the level of conservation in four species, (iii) a set of context scores based on the properties and overall complementary between a microRNA and its target.
- A composite score was estimated for each target by SVR ranking model from the individual criteria scores described above.
- Spearman rank correlation coefficient is computed between the scoring features to evaluate their dependence.

MicroTarget does not filter out the prediction results with the targeting features like most of other methods do. The prediction of validated targets as the top ranked targets in our approach show good consistency of our approach performance with the factor of using expression data. In addition, the analysis of feature relevance suggests that the model built upon the feature set presents the most balanced ranking results in terms of specificity and sensitivity. The comparative study for our approach performed in this research show that MicroTarget adds to the field of target prediction in the sense of providing promising candidate target for further experimental validation.

8.1.1 Future direction

Further research in this direction may be needed to a gain better understanding of the role of microRNA in the cell machinery. Analysis of miRNAs and their target genes is expected to shed light on the potentially diverse and important biological functions of miRNAs within living systems. For instance, microRNAs can act as oncogenes or tumor suppressors to inhibit the expression of cancer related-genes and to promote or suppress the tumors in various tissues. Therefore, using microRNA to target oncogenes might improve the therapeutic outcomes in human cancers. Once microRNA regulatory interactions are predicted with good accuracy, the next step is to use these results for therapeutic applications. In future work, we will use MicroTarget to predict microRNA interactions that defer in different cancer type.

Upon degradation of the complex mRNA-miRNA, miRNA molecules can be recycled with a ratio. That is, one miRNA can work for several rounds of target recognition and cleavage per miRNA before it is degraded [60]. Also, it has been shown this recycle ratio is a very important factor for the dynamic of RNA-miRNA reciprocal regulation with theoretical analysis [144]. However, there is no such as tool which can predict or measure this recycle ratio. This recycling of microRNA regulation cannot be discovered from the sequence data; the gene expression data is the best candidate information to do so. Time series expression data can be used to predict the microRNA recycle ratio. In future work, we will work on time-series expression data to measure the recycle ratio of the microRNA regulation.

Other interesting future work for our research is adding new functions for our prediction approach based on the competitive endogenous RNA (ceRNA) hypothesis. The ceRNA hypothesis proposes that mRNAs with shared microRNA binding sites compete for post-transcriptional control. The central mechanism underlying the ceRNA hypothesis is the idea that mRNAs may have indirect interactions among themselves that are mediated by competition and depletion of shared microRNA pools. In other word, when a ceRNA such as a pseudogene, remains transcriptionally silent, the parent mRNA is transcribed and exported to the cytoplasm where it is targeted by the microRNA, resulting in decreasing the expression level of the parent gene. But, when the pseudogene with competing target sites becomes active, it competes for binding with the microRNA. This drives microRNAs away from the parent gene and leads to an increase in the parent gene expression [143]. We suggest to predict these indirect interactions in a form of ceRNA network. The ideas for providing evidence

for competition of microRNA regulation can be collected by constructing a genome-level network of microRNA-mediated interactions.

8.2 Identifying conserved complexes

Protein complexes are key functional units in many biological processes. The recent advances in high throughput experimental techniques provide large protein-protein interactions (PPIs) data for many species. Identifying conserved complexes between species is a fundamental step towards learning the conserved mechanisms among different species, as well as transferring knowledge from model organisms to others. Researchers obtain PPI networks as input and provide computational methods to detect conserved protein complexes. Current methods based on PPI networks do not work well in identifying conserved complexes. They are severely limited by the lack of true interactions and presence of large amounts of false interactions in PPI data.

We integrate multiple data sources to build an alignment graph among PPI networks of two species. Rather than explicitly restrict our attention to align homologous proteins, we decompose PPI networks in terms of their domains and employ their conservation along with PPI data to construct an alignment graph. The nodes of the alignment graph represent orthologous proteins between the two input networks that share one or more domains. The alignment graph has three types of edges composite, simple-direct, and simple-indirect. Then, edges and nodes of the alignment graph are assigned weights. The final step of DONA is to cluster the alignment graph with the MCL algorithm. The main contributions of this dissertation in this problem can be summarized as:

- We first presented a case study evaluation for the current computational methods for identifying conserved protein complexes. A brief overview on the current methods and the evaluation study are given in Chapter 6.
- We developed a novel approach, DONA, which is based on a new strategy for building an alignment graph to identify the conserved complexes.
- As protein evolution can be understood through domains, we add data sets that consider domain conservation.

- We developed a new scoring scheme to measure the conservation level between proteins and their interaction.
- We demonstrate that integrating domain interaction data significantly enhances the quality of the alignment.
- We build an extensive testing data set for identifying the conserved protein complexes between five different species. A collection of conserved sub-networks among these species is identified. As currently there is no benchmark data set for conserved protein complexes in the literature, we hope that this data set could be useful.

Our experiments on the data sets revealed that DONA can identify conserved sub-networks more efficiently than existing methods in term of precision and recall. DONA produced better results in terms of the agreement with known protein complexes. Recruiting DDIs in the alignment process performed well in identifying the conservation across species. Moreover, DONA provides a more comprehensive means for biologically interpreting the aligned sub-networks, as protein domains are directly related to protein function. All the analyses for identifying conserved protein complexes were performed on pairwise alignments of five species: human, mouse, rat, fly, and yeast. This is because we need to study the performance of our approach in closely as well as distantly related species.

8.2.1 Future direction

In our future work, we will concentrate on understanding the function and evolution of the proteins interactions among more than two species by many-to-many alignment. DONA provides pairwise alignment. A careful modification for DONA is needed to analyze the conserved interactions among group of species. Such an update would be helpful in understanding the similarity of networks in multiple species and evolutionary events that might have taken place among these species. Expanding DONA to multiple alignment will be our next target. This can be performed by pairwise alignment of networks along a phylogenetic tree. The result of multiple alignments would identify the types of protein complexes that are common across a number of species.

Another future research direction for DONA can be adapting it to align other types of networks, such as, gene interaction networks. These types of networks are often presented

as directed graphs. Therefore, further work to modify DONA to utilize on direct graph is required, such as, redefining the edge scoring function to satisfy the properties of these networks. Moreover, some of these networks are sparser than PPI networks; therefore the clustering method might needed to be rethought. Farther future direction could be improving the usability of DONA by developing an online system for it. Where users could upload their PPI network for alignment. In this case a function could be added to DONA to estimate the impact of varying the inflation level on MCL clustering and provide the user with the inflation parameter range that generate the best performance [145].

Another interesting future work is predicting protein functions. Proteins that are found in a structural complex are functionally related. This leads us to tentative functional assignments, which is called annotation transfer. Future work for our research could be directed in this way. Here is one idea. Given a set of proteins in a complex, we can predict new protein functions when a set of requirements are fulfilled. For instance, the set of proteins in the conserved complex is significantly enriched for a particular GO annotation with very low corrected $p - value$, at least 80% of the proteins are annotated with this GO annotation, and the GO annotation is in a high level in the GO tree, and other requirements could be added. Then all the proteins in the set could be considered to have this GO annotation.

Bibliography

[1] Hamed Al-Hussaini, Deepa Subramanyam, Michael Reedijk, and Srikala S. Sridhar. Notch signaling pathway as a therapeutic target in breast cancer. *Molecular Cancer Therapeutics*, 10(1):9–15, 2011.

[2] Maria I. Almeida, Rui M. Reis, and George A. Calin. MicroRNA history: Discovery, recent applications, and next frontiers. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, 717(1-2):1–8, 2011.

[3] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[4] Victor Ambros. microRNAs: Tiny regulators with great potential. *Cell*, 107(7):823–826, 2001.

[5] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene Ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[6] Gary D Bader and Christopher W V Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4:2, 2003.

[7] Onureena Banerjee, Laurent El Ghaoui, and Alexandre D'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.

[8] D Bartel. MicroRNAs genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297, jan 2004.

[9] Doron Betel, Anjali Koppal, Phaedra Agius, Chris Sander, and Christina Leslie. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biology*, 11(8):R90, 2010.

[10] Ramachandra M Bhaskara and Narayanaswamy Srinivasan. Stability of domain structures in multi-domain proteins. *Scientific reports*, 1:40, 2011.

[11] D Bhaumik, G K Scott, S Schokrpur, C K Patil, J Campisi, and C C Benz. Expression of microRNA-146 suppresses NF-kappaB activity with reduction of metastatic potential in breast cancer cells. *Oncogene*, 27(42):5643–5647, 2008.

[12] Patrik Bjorkholm and E. L L Sonnhammer. Comparative analysis and unification of domain-domain interaction networks. *Bioinformatics*, 25(22):3020–3025, 2009.

[13] T. Borggrefe and F. Oswald. The Notch signaling pathway: Transcriptional regulation at Notch target genes. *Cellular and Molecular Life Sciences*, 66(10):1631–1646, 2009.

[14] Peer Bork, Lars J. Jensen, Christian Von Mering, Arun K. Ramani, Insuk Lee, and Edward M. Marcotte. Protein interaction networks from yeast to human. *Current Opinion in Structural Biology*, 14(3):292–299, 2004.

[15] Stephen Boyd. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2010.

[16] Elizabeth I. Boyle, Shuai Weng, Jeremy Gollub, Heng Jin, David Botstein, J. Michael Cherry, and Gavin Sherlock. GO::TermFinder - Open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710–3715, 2004.

[17] Sylvain Brohee and Jacques van Helden. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC bioinformatics*, 7:488, 2006.

[18] K R Brown and I Jurisica. Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome biology*, 8(5):R95, 2007.

[19] Catherine Bru, Emmanuel Courcelle, Sébastien Carrère, Yoann Beausse, Sandrine Dalmar, and Daniel Kahn. The ProDom database of protein domain families: More emphasis on 3D. *Nucleic Acids Research*, 33(DATABASE ISS.):212–215, 2005.

[20] Anna Brückner, Cécile Polge, Nicolas Lentze, Daniel Auerbach, and Uwe Schlattner. Yeast two-hybrid, a powerful tool for systems biology. *International Journal of Molecular Sciences*, 10(6):2763–2788, 2009.

[21] Tony Cai, Weidong Liu, and Xi Luo. A constrained L1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

[22] Yimei Cai, Xiaomin Yu, Songnian Hu, and Jun Yu. A brief review on the mechanisms of miRNA regulation. *Genomics, Proteomics and Bioinformatics*, 7(4):147–154, 2009.

[23] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, 2011.

[24] Andrew Chatr-Aryamontri, Bobby Joe Breitkreutz, Rose Oughtred, Lorrie Boucher, Sven Heinicke, Daici Chen, Chris Stark, Ashton Breitkreutz, Nadine Kolas, Lara O'Donnell, Teresa Reguly, Julie Nixon, Lindsay Ramage, Andrew Winter, Adnane Sellam, Christie Chang, Jodi Hirschman, Chandra Theesfeld, Jennifer Rust, Michael S. Livstone, Kara Dolinski, and Mike Tyers. The BioGRID interaction database: 2015 update. *Nucleic Acids Research*, 43(D1):D470–D478, 2015.

[25] Marina Chekulaeva and Witold Filipowicz. Mechanisms of miRNA-mediated post-transcriptional regulation in animal cells. *Current Opinion in Cell Biology*, 21(3):452–460, 2009.

[26] Giovanni Ciriello, Marco Mina, Pietro H. Guzzi, Mario Cannataro, and Concettina Guerra. AlignNemo: A local network alignment method to integrate homology and topology. *PLoS ONE*, 7(6), 2012.

[27] Bryan R. Cullen. Transcription and processing of human microRNA precursors. *Molecular Cell*, 16(6):861–865, 2004.

[28] Patrick Danaher, Pei Wang, and Daniela M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 76(2):373–397, 2014.

[29] Jun Ding, Xiaoman Li, and Haiyan Hu. TarPmiR: A new approach for microRNA target site prediction. *Bioinformatics*, 32(18):2768–2775, 2016.

[30] Janusz Dutkowski and Jerzy Tiuryn. Identification of functional modules from conserved ancestral protein-protein interactions. *Bioinformatics*, 23(13):149–158, 2007.

[31] Harsh Dweep, Carsten Sticht, Priyanka Pandey, and Norbert Gretz. MiRWalk - Database: Prediction of possible miRNA binding sites by walking the genes of three genomes. *Journal of Biomedical Informatics*, 44(5):839–847, 2011.

[32] Amy B Emerman, Zai-Rong Zhang, Oishee Chakrabarti, and Ramanujan S Hegde. Compartment-restricted biotinylation reveals novel features of prion protein metabolism in vivo. *Molecular biology of the cell*, 21(24):4325–4337, 2010.

[33] Espen Enerly, Israel Steinfeld, Kristine Kleivi, Suvi Katri Leivonen, Miriam R. Aure, Hege G. Russnes, Jo Anders Rønneberg, Hilde Johnsen, Roy Navon, Einar Rødland, Rami Mäkelä, Bjørn Naume, Merja Perälä, Olli Kallioniemi, Vessela N. Kristensen, Zohar Yakhini, and Anne Lise Børresen-Dale. miRNA-mRNA integrated analysis reveals roles for mirnas in primary breast tumors. *PLoS ONE*, 6(2), 2011.

[34] 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.

[35] David Eppstein. Subgraph isomorphism in planar graphs and pelated problems. *Proceedings of the Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, 3(3):632–640, 1995.

[36] Fazle E Faisal, Lei Meng, Joseph Crawford, and Tijana Milenković. The post-genomic era of biological network alignment. *EURASIP Journal on Bioinformatics and Systems Biology*, 2015:3, 2015.

[37] Fazle Elahi Faisal, Han Zhao, and Tijana Milenkovic. Global network alignment in the context of aging. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(1):40–52, 2015.

[38] Kyle Kai-How Farh, Andrew Grimson, Calvin Jan, Benjamin P Lewis, Wendy K Johnston, Lee P Lim, Christopher B Burge, and David P Bartel. The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science*, 310(5755):1817–1821, 2005.

[39] Robert D Finn, Penelope Coggill, Ruth Y Eberhardt, Sean R Eddy, Jaina Mistry, Alex L Mitchell, Simon C Potter, Marco Punta, Matloob Qureshi, Amaia Sangrador-Vegas, Gustavo A Salazar, John Tate, and Alex Bateman. The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Research*, 44(D1):D279–D285, 2015.

[40] Robert D. Finn, Benjamin L. Miller, Jody Clements, and Alex Bateman. IPfam: A database of protein family and domain interactions found in the Protein Data Bank. *Nucleic Acids Research*, 42(D1):364–373, 2014.

[41] Jason Flannick, Antal Novak, Balaji S. Srinivasan, Harley H. McAdams, and Serafim Batzoglou. Graemlin: General and robust alignment of multiple large interaction networks. *Genome Research*, 16(9):1169–1181, 2006.

[42] Hunter B Fraser, Aaron E Hirsh, Dennis P Wall, and Michael B Eisen. Coevolution of gene expression among interacting proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 101(24):9033–8, 2004.

[43] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[44] Robin C. Friedman, Kyle Kai How Farh, Christopher B. Burge, and David P. Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19(1):92–105, 2009.

[45] David M Garcia, Daehyun Baek, Chanseok Shin, George W Bell, Andrew Grimson, and David P Bartel. Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nature Structural & Molecular Biology*, 18(10):1139–1146, 2011.

[46] Alvaro J Gonzalez, Li Liao, Alvaro J Gonzalez, and Li Liao. Predicting domain-domain interaction based on domain profiles with feature selection and support vector machines. *BMC Bioinformatics*, 11:537–550, 2010.

[47] Sam Griffiths-Jones, Russell J Grocock, Stijn van Dongen, Alex Bateman, and Anton J Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, 34(Database issue):D140–D144, 2006.

[48] Andrew Grimson, Kyle Kai How Farh, Wendy K. Johnston, Philip Garrett-Engele, Lee P. Lim, and David P. Bartel. MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing. *Molecular Cell*, 27(1):91–105, 2007.

[49] Xin Guo and Alexander J. Hartemink. Domain-oriented edge-based alignment of protein interaction networks. *Bioinformatics*, 25(12):240–246, 2009.

[50] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 402(6761 Suppl):C47–C52, 1999.

[51] Mallory a. Havens, Ashley a. Reich, Dominik M. Duelli, and Michelle L. Hastings. Biogenesis of mammalian microRNAs by a non-canonical processing pathway. *Nucleic Acids Research*, 40(10):4626–4640, 2012.

[52] Luqman Hodgkinson and Richard M. Karp. Algorithms to detect multiprotein modularity conserved during evolution. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(4):1046–1058, 2012.

[53] Ivo L. Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Research*, 31(13):3429–3431, 2003.

[54] Mingyi Hong and Zhi-Quan Luo. On the linear convergence of the alternating direction method of multipliers. *Mathematical Programming Series*, 23:49–85, 2012.

[55] Anwar Hossain, Macus T Kuo, and Grady F Saunders. Mir-17-5p regulates breast cancer cell proliferation by inhibiting translation of AIB1 mRNA. *Molecular and Cellular Biology*, 26(21):8191–8201, 2006.

[56] Sheng Da Hsu, Yu Ting Tseng, Sirjana Shrestha, Yu Ling Lin, Anas Khaleel, Chih Hung Chou, Chao Fang Chu, Hsien Da 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 Yuan Huang. MiRTarBase update 2014: An information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Research*, 42(D1):78–85, 2014.

[57] Jialu Hu and Knut Reinert. LocalAli: An evolutionary-based local alignment approach to identify functionally conserved modules in multiple networks. *Bioinformatics*, 31(3):363–372, 2014.

[58] Yanhui Hu, Ian Flockhart, Arunachalam Vinayagam, Clemens Bergwitz, Bonnie Berger, Norbert Perrimon, and Stephanie E Mohr. An integrative approach to ortholog prediction for disease-focused and other functional studies. *BMC Bioinformatics*, 12:357, 2011.

[59] Jim C Huang, Tomas Babak, Timothy W Corson, Gordon Chua, Sofia Khan, Brenda L Gallie, Timothy R Hughes, Benjamin J Blencowe, Brendan J Frey, and Quaid D Morris. Using expression profiling data to identify human microRNA targets. *Nature Methods*, 4(12):1045–1049, 2007.

[60] Gyorgy Hutvagner and Phillip D Zamore. A microRNA in a multiple- turnover RNAi enzyme complex. *Science*, 297(September):2056–2060, 2002.

[61] Zohar Itzhaki, Eyal Akiva, Yael Altuvia, and Hanah Margalit. Evolutionary conservation of domain-domain interactions. *Genome Biology*, 7(12):R125, 2006.

[62] Irena Ivanovska, Alexey S Ball, Robert L Diaz, Jill F Magnus, Miho Kibukawa, Janell M Schelter, Sumire V Kobayashi, Lee Lim, Julja Burchard, Aimee L Jackson, Peter S Linsley, and Michele a Cleary. MicroRNAs in the miR-106b family regulate p21/CDKN1A and promote cell cycle progression. *Molecular and Cellular Biology*, 28(7):2167–2174, 2008.

[63] Bino John, Anton J. Enright, Alexei Aravin, Thomas Tuschl, Chris Sander, and Debora S. Marks. Human microRNA targets. *PLoS Biology*, 2(11), 2004.

[64] Maxim Kalaev, Mike Smoot, Trey Ideker, and Roded Sharan. NetworkBLAST: Comparative analysis of protein networks. *Bioinformatics*, 24(4):594–596, 2008.

[65] Brian P Kelley, Roded Sharan, Richard M Karp, Taylor Sittler, David E Root, Brent R Stockwell, and Trey Ideker. Conserved pathways within bacteria and yeast as revealed by global protein network alignment. *Proceedings of the National Academy of Sciences of the United States of America*, 100(20):11394–11399, 2003.

[66] Brian P. Kelley, Bingbing Yuan, Fran Lewitter, Roded Sharan, Brent R. Stockwell, and Trey Ideker. PathBLAST: A tool for alignment of protein interaction networks. *Nucleic Acids Research*, 32(WEB SERVER ISS.):83–88, 2004.

[67] Michael Kertesz, Nicola Iovino, Ulrich Unnerstall, Ulrike Gaul, and Eran Segal. The role of site accessibility in microRNA target recognition. *Nature Genetics*, 39(10):1278–1284, 2007.

[68] Mohsen Khorshid, Jean Hausser, Mihaela Zavolan, and Erik van Nimwegen. A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. *Nature Methods*, 10(3):253–5, 2013.

[69] Rimpi Khurana, Vinod Kumar Verma, Abdul Rawoof, Shrish Tiwari, Rekha a Nair, Ganesh Mahidhara, Mohammed M Idris, Alan R Clarke, and Lekha Dinesh Kumar. OncomiRdbB: a comprehensive database of microRNAs and their targets in breast cancer. *BMC Bioinformatics*, 15(1):15, 2014.

[70] Sung-Kyu Kim, Jin-Wu Nam, Je-Keun Rhee, Wha-Jin Lee, and Byoung-Tak Zhang. miTarget: microRNA target gene prediction using a support vector machine. *BMC Bioinformatics*, 7:411, 2006.

[71] Yohan Kim, Shankar Subramaniam, Wojciech Szpankowski, and Ananth Grama. Detecting conserved interaction patterns in biological networks. *Journal of Computational Biology*, 13(7):1299–1322, 2006.

[72] A. D. King, N. Pržulj, and I. Jurisica. Protein complex prediction via cost-based clustering. *Bioinformatics*, 20(17):3013–3020, 2004.

[73] Rhoda J. Kinsella, Andreas Kähäri, Syed Haider, Jorge Zamora, Glenn Proctor, Giulietta Spudich, Jeff Almeida-King, Daniel Staines, Paul Derwent, Arnaud Kerhornou, Paul Kersey, and Paul Flicek. Ensembl BioMarts: A hub for data retrieval across taxonomic space. *Database*, 2011:1–9, 2011.

[74] Marianthi Kiriakidou, Peter T. Nelson, Andrei Kouranov, Petko Fitziev, Costas Bouyioukos, Zissimos Mourelatos, and Artemis Hatzigeorgiou. A combined computational-experimental approach predicts human microRNA targets. *Genes and Development*, 18(10):1165–1178, 2004.

[75] Mehmet Koyut. Pairwise local nlignment of protein interaction. *Pacific Symposium on Biocomputing*, 108(2):48–65, 2005.

[76] Ana Kozomara and Sam Griffiths-Jones. MiRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research*, 42(D1):1–6, 2014.

[77] Azra Krek, Dominic Grün, Matthew N Poy, Rachel Wolf, Lauren Rosenberg, Eric J Epstein, Philip MacMenamin, Isabelle da Piedade, Kristin C Gunsalus, Markus Stoffel, Nikolaus Rajewsky, 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. *Nature Genetics*, 37(5):495–500, 2005.

[78] Oleksii Kuchaiev and Nataša Pržulj. Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics*, 27(10):1390–1396, 2011.

[79] Markus Landthaler, Dimos Gaidatzis, Andrea Rothballer, Po Yu Chen, Steven Joseph Soll, Lana Dinic, Tolulope Ojo, Markus Hafner, Mihaela Zavolan, and Thomas Tuschl. Molecular characterization of human Argonaute-containing ribonucleoprotein complexes and their bound target mRNAs. *RNA*, 14(12):2580–2596, 2008.

[80] Minh T N Le, Peter Hamar, Changying Guo, Emre Basar, Ricardo Perdigão-henriques, Leonora Balaj, and Judy Lieberman. miR-200 — containing extracellular vesicles promote breast cancer cell metastasis. *The Journal of Clinical Investigation*, 124(12):5109–5128, 2014.

[81] Yong Sun Lee and Anindya Dutta. The tumor suppressor microRNA let-7 represses the HMGA2 oncogene. *Genes and Development*, 21:1025–1030, 2007.

[82] Benjamin P. Lewis, Christopher B. Burge, and David P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20, 2005.

[83] Hongling Li, Chunjing Bian, Lianming Liao, Jing Li, and Robert Chunhua Zhao. miR-17-5p promotes human breast cancer cell migration and invasion through suppression of HBP1. *Breast Cancer Research and Treatment*, 126(3):565–575, 2011.

[84] Chung Shou Liao, Kanghao Lu, Michael Baym, Rohit Singh, and Bonnie Berger. IsoRankN: Spectral methods for global alignment of multiple protein networks. *Bioinformatics*, 25(12):253–258, 2009.

[85] David Liben-Nowell and Jon Kleinberg. The Link Prediction Problem for Social Networks. *Proceedings of the Twelfth Annual ACM International Conference on Information and Knowledge Management (CIKM)*, (November 2003):556–559, 2003.

[86] Lee P Lim, Nelson C Lau, Philip Garrett-Engele, Andrew Grimson, Janell M Schelter, John Castle, David P Bartel, Peter S Linsley, and Jason M Johnson. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, 433(7027):769–773, 2005.

[87] Yat-Yuen Lim, Josephine a Wright, Joanne L Attema, Philip a Gregory, Andrew G Bert, Eric Smith, Daniel Thomas, Angel F Lopez, Paul a Drew, Yeesim Khew-Goodall, and Gregory J Goodall. Epigenetic modulation of the miR-200 family is associated with transition to a breast cancer stem-cell-like state. *Journal of Cell Science*, 126(Pt 10):2256–66, 2013.

[88] Chen-Chung Lin, Ling-Zhi Liu, Joseph B Addison, William F Wonderlin, Alexey V Ivanov, and J Michael Ruppert. A KLF4-miRNA-206 autoregulatory feedback loop can promote or inhibit protein translation depending upon cell context. *Molecular and Cellular Biology*, 31(12):2513–2527, 2011.

[89] Hui Liu, Dong Yue, Yidong Chen, Shou-Jiang Gao, and Yufei Huang. Improving performance of mammalian microRNA target prediction. *BMC Bioinformatics*, 11:476, 2010.

[90] Ronny Lorenz, Stephan H Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.

[91] William H Majoros, Parawee Lekprasert, Neelanjan Mukherjee, Rebecca L Skalsky, David L Corcoran, Bryan R Cullen, and Uwe Ohler. MicroRNA target site identification by integrating sequence and binding information. *Nature Methods*, 10(7):630–633, 2013.

[92] Ray M Mar\'in, Ji\'i Van\'iek, Ray M. Marín, and Jií Vaníek. Efficient use of accessibility in microRNA target prediction. *Nucleic Acids Research*, 39(1):19–29, 2011.

[93] Aron Marchler-Bauer, Myra K. Derbyshire, Noreen R. Gonzales, Shennan Lu, Farideh Chitsaz, Lewis Y. Geer, Renata C. Geer, Jane He, Marc Gwadz, David I. Hurwitz,

Christopher J. Lanczycki, Fu Lu, Gabriele H. Marchler, James S. Song, Narmada Thanki, Zhouxi Wang, Roxanne A. Yamashita, Dachuan Zhang, Chanjuan Zheng, and Stephen H. Bryant. CDD: NCBI's conserved domain database. *Nucleic Acids Research*, 43(D1):D222–D226, 2015.

[94] E M Marcotte, M Pellegrini, M J Thompson, T O Yeates, and D Eisenberg. A combined algorithm for genome-wide prediction of protein function. *Nature*, 402(6757):83–6, 1999.

[95] Joseph A Marsh, Helena Hernandez, Zoe Hall, Sebastian E Ahnert, Tina Perica, Carol V Robinson, and Sarah A. Teichmann. Protein complexes are under evolutionary selection to assemble via ordered pathways. *Cell*, 153(2):461–470, 2013.

[96] Aida Martinez-Sanchez and Chris L Murphy. MicroRNA target identification-experimental approaches. *Biology*, 2(1):189–205, 2013.

[97] T. G. McDaneld. MicroRNA: mechanism of gene regulation and application to livestock. *Journal of Animal Science*, 87(14 Suppl), 2009.

[98] Scott McGinnis and Thomas L. Madden. BLAST: At the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*, 32(WEB SERVER ISS.):20–25, 2004.

[99] Giovanni Micale, Alfredo Pulvirenti, Rosalba Giugno, and Alfredo Ferro. GASOLINE: A greedy and stochastic algorithm for optimal local multiple alignment of interaction Networks. *PLoS ONE*, 9(6), 2014.

[100] Tijana Milenković, Weng Leong Ng, Wayne Hayes, and Nataša Pržulj. Optimal network alignment with graphlet degree vectors. *Cancer Informatics*, 9:121–137, 2010.

[101] Marco Mina and Pietro Hiram Guzzi. AlignMCL: Comparative analysis of protein interaction networks through Markov clustering. *2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops*, pages 174–181, 2012.

[102] Prasun J Mishra. MicroRNAs as promising biomarkers in cancer diagnostics. *Biomarker Research*, 2(1):19, jan 2014.

[103] Roberto Mosca, Arnaud Ceol, Amelie Stein, Roger Olivella, and Patrick Aloy. 3did: A catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Research*, 42(D1):374–379, 2014.

[104] M. M. Mukaka. Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, 24(3):69–71, 2012.

[105] Su Naifang, Qian Minping, and Deng Minghua. Integrative approaches for microRNA target prediction: combining sequence information and the Paired mRNA and miRNA expression profiles. *Current Bioinformatics*, 8(1):37–45, 2013.

[106] Viswam S. Nair, Colin C. Pritchard, Muneesh Tewari, and John P a Ioannidis. Design and analysis for studying microRNAs in human disease: A primer on-omic technologies. *American Journal of Epidemiology*, 180(2):140–152, jul 2014.

[107] Jin Wu Nam, Olivia S. Rissland, David Koppstein, Cei Abreu-Goodger, CalvinH Jan, Vikram Agarwal, Muhammed a. Yildirim, Antony Rodriguez, and David P. Bartel. Global analyses of the effect of different cellular contexts on microRNA targeting. *Molecular Cell*, 53(6):1031–1043, 2014.

[108] Manikandan Narayanan and Richard M. Karp. Comparing protein interaction networks via a graph. *Journal of Computational Biology*, 14(7):1–15, 2007.

[109] Cydney B Nielsen, Noam Shomron, Rickard Sandberg, Eran Hornstein, Jacob Kitzman, and Christopher B Burge. Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA*, 13(11):1894–910, 2007.

[110] Andersson Orom and Anders H. Lund. Isolation of microRNA targets using biotinylated synthetic microRNAs. *Methods*, 43(2):162–165, 2007.

[111] Roland A. Pache, Arnaud Céol, and Patrick Aloy. NetAligner: a network alignment server to compare complexes, pathways and whole interactomes. *Nucleic Acids Research*, 40(W1):157–161, 2012.

[112] Philipp Pagel, Matthias Oesterheld, Oksana Tovstukhina, Norman Strack, Volker Stümpflen, and Dmitrij Frishman. DIMA 2.0 - Predicted and known domain interactions. *Nucleic Acids Research*, 36(SUPPL. 1):651–655, 2008.

[113] Rob Patro and Carl Kingsford. Global network alignment using multiscale spectral signatures. *Bioinformatics*, 28(23):3105–3114, 2012.

[114] Florencio Pazos and Alfonso Valencia. In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins: Structure, Function and Genetics*, 47(2):219–227, 2002.

[115] Wei Peng, Jianxin Wang, Fangxiang Wu, and Pan Yi. Detecting conserved protein complexes using a dividing-and-matching algorithm and unequally lenient criteria for network comparison. *Algorithms for Molecular Biology*, 10:21, 2015.

[116] J. B. Pereira-Leal, E. D. Levy, and S. A. Teichmann. The origins and evolution of functional modules: Lessons from protein complexes. *Philosophical Transaction of Biology*, 361(1467):507–517, 2006.

[117] James R. Perkins, Ilhem Diboun, Benoit H. Dessailly, Jon G. Lees, and Christine Orengo. Transient protein-protein interactions: Structural, functional, and network properties. *Structure*, 18(10):1233–1243, 2010.

[118] Sarah M. Peterson, Jeffrey A. Thompson, Melanie L. Ufkin, Pradeep Sathyanarayana, Lucy Liaw, and Clare Bates Congdon. Common features of microRNA target prediction tools. *Frontiers in Genetics*, 5(FEB):1–10, 2014.

[119] Hang T T Phan and Michael J E Sternberg. PINALOG: A novel approach to align protein interaction networks-implications for complex detection and function prediction. *Bioinformatics*, 28(9):1239–1245, 2012.

[120] Sylvain Pitre, Alamgir James, and R Green Michel. Computational methods For predicting protein-protein interactions. *Advances in Biochemical Engineering/Biotechnology.*, (January):247–267, 2008.

[121] Guillaume Postic, Yassine Ghouzam, Romain Chebrek, and Jean-christophe Gelly. An ambiguity principle for assigning protein structural domains. (January), 2017.

[122] Shuye Pu, Jessica Wong, Brian Turner, Emerson Cho, and Shoshana J. Wodak. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Research*, 37(3):825–831, 2009.

[123] Balaji Raghavachari, Asba Tasneem, Teresa M. Przytycka, and Raja Jothi. DOMINE: A database of protein domain interactions. *Nucleic Acids Research*, 36(SUPPL. 1):656–661, 2008.

[124] Marc Rehmsmeier, Peter Steffen, Matthias Höchsmann, Robert Giegerich, and Matthias Ho. Fast and effective prediction of microRNA / target duplexes. *Bioinformatics*, (2003):1507–1517, 2004.

[125] B J Reinhart, F J Slack, M Basson, a E Pasquinelli, J C Bettinger, a E Rougvie, H R Horvitz, and G Ruvkun. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, 403(6772):901–906, 2000.

[126] William Ritchie and John E. J. Rasko. Refining microRNA target predictions: Sorting the wheat from the chaff. *Biochemical and Biophysical Research Communications*, 445(4):780–784, 2014.

[127] Harlan Robins, Ying Li, and Richard W Padgett. Incorporating structure to predict microRNA targets. *Proceedings of the National Academy of Sciences of the United States of America*, 102(11):4006–9, 2005.

[128] PeterW. Rose, Andreas Prli, Ali Altunkaya, Chunxiao Bi, Anthony R. Bradley, H. Cole Christie, Luigi Di Costanzo, Jose M. Duarte, Shuchismita Dutta, Zukang Feng-Green Rachel Kramer, David S. Goodsell, Brian Hudson, Tara Kalro, Robert Lowe, Ezra Peisach, Christopher Randle, Alexander S. Rose, Chenghua Shao, Yi-Ping Tao, Valasatava Yana, Maria Voigt, Huangwang John D.Westbrook JesseWoo Yang, Jasmine Y. Young, Christine Zardecki, Helen M. Berman, and Stephen K. Burley. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Research*, 45(October 2016):1–15, 2016.

[129] Kristian Rother, Magdalena Rother, Michał Micha\l Boniecki, Tomasz Puton, and Janusz M. Bujnicki. RNA and protein 3D structure modeling: Similarities and differences. *Journal of Molecular Modeling*, 17(9):2325–2336, 2011.

[130] J Graham Ruby, Calvin H Jan, and David P Bartel. Intronic microRNA precursors that bypass Drosha processing. *Nature*, 448(7149):83–6, 2007.

[131] Andreas Ruepp, Brigitte Waegele, Martin Lechner, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Gisela Fobo, Goar Frishman, Corinna Montrone, and H. Werner Mewes. CORUM: The comprehensive resource of mammalian protein complexes-2009. *Nucleic Acids Research*, 38(SUPPL.1):497–501, 2009.

[132] Catherine Sanchez, Corinne Lachaize, Florence Janody, Bernard Bellon, Laurence Roder, Jerome Euzenat, Francois Rechenmann, and Bernard Jacq. Grasping at molecular interactions and genetic networks in Drosophila melanogaster using FlyNets, an Internet database. *Nucleic Acids Research*, 27(1):89–94, 1999.

[133] Stefanie Sassen, Eric a. Miska, and Carlos Caldas. MicroRNA—Implications for cancer. *International Journal of Pathology*, 452(1):1–10, 2008.

[134] Boon Siew Seah, Sourav S. Bhowmick, and C. Forbes Dewey. DualAligner: A dual alignment-based strategy to align protein interaction networks. *Bioinformatics*, 30(18):2619–2626, 2014.

[135] Roded Sharan, Trey Ideker, Brian Kelley, Ron Shamir, and Richard M Karp. Identification of Protein complexes by comparative analysis of yeast and bacterial protein interaction data. *Journal of computational biology*, 12(6):835–846, 2005.

[136] Roded Sharan, Silpa Suthram, Ryan M Kelley, Tanja Kuhn, Scott McCuine, Peter Uetz, Taylor Sittler, Richard M Karp, and Trey Ideker. Conserved patterns of protein interaction in multiple species. *Proceedings of the National Academy of Sciences of the United States of America*, 102(6):1974–1979, 2005.

[137] Benjamin A. Shoemaker and Anna R. Panchenko. Deciphering protein-protein interactions. Part I. Experimental techniques and databases. *PLoS Computational Biology*, 3(3):0337–0344, 2007.

[138] Erik L. L. Sonnhammer, Sean R. Eddy, Ewan Birney, Alex Bateman, and Richard Durbin. Pfam: Multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res. *Nucleic Acids Research*, 26(1):320–2, 1998.

[139] Balaji S Srinivasan and Serafim Batzoglou. Automatic parameter learning for multiple local network alignment. *Journal of Computational Biology*, 16(8):1001–1022, 2009.

[140] Balaji S. Srinivasan, Nigam H. Shah, Jason A. Flannick, Eduardo Abeliuk, Antal F. Novak, and Serafim Batzoglou. Current progress in network research: Toward reference networks for key model organisms. *Briefings in Bioinformatics*, 8(5):318–332, 2007.

[141] Xiaoyun Sun, Pengyu Hong, Meghana Kulkarni, Young Kwon, and Norbert Perrimon. PPIRank — an advanced method for ranking protein-protein interations in TAP/MS data. *Proteome Science*, 11(Suppl 1):S16, 2013.

[142] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos,

Kalliopi P. Tsafou, Michael Kuhn, Peer Bork, Lars J. Jensen, and Christian Von Mering. STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(D1):D447–D452, 2015.

[143] Daniel W Thomson and Marcel E Dinger. Endogenous microRNA sponges: evidence and controversy. *Nature Reviews Genetics*, 17(5):272–283, 2016.

[144] Xiao-jun Tian, Hang Zhang, Jingyu Zhang, and Jianhua Xing. Reciprocal regulation between mRNA and microRNA enables a bistable switch that directs cell fate decisions. *FEBS Letters*, 590(19):3443–3455, 2016.

[145] S. van Dongen. Performance criteria for graph clustering and Markov cluster experiments. *Technical Report INS-R0012, National Research Institute for Mathematics and Computer Science*, page 36, 2000.

[146] Stijn van Dongen, Cei Abreu-Goodger, and Anton J Enright. Detecting microRNA binding and siRNA off-target effects from expression data. *Nature Methods*, 5(12):1023–1025, 2008.

[147] Stijn van Dongen, Cei Abreu-Goodger, Stijn van Dongen, and Cei Abreu-Goodger. Using MCL to Extract Clusters from Networks. *Methods in Molecular Biology*, 804:281–295, 2012.

[148] Eleni van Schooneveld, Hans Wildiers, Ignace Vergote, Peter B Vermeulen, Luc Y Dirix, and Steven J Van Laere. Dysregulation of microRNAs in breast cancer and their potential role as prognostic and predictive biomarkers in patient management. *Breast Cancer Research*, 17(1):1–15, 2015.

[149] Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91, 2006.

[150] V. Vijayan, V. Saraph, and T. Milenković. MAGNA++: Maximizing accuracy in global network alignment via both node and edge conservation. *Bioinformatics*, 31(14):2409–2411, 2015.

[151] Jeppe Vinther, Mads M. Hedegaard, Paul P. Gardner, Jens S. Andersen, and Peter Arctander. Identification of miRNA targets with stable isotope labeling by amino acids in cell culture. *Nucleic Acids Research*, 34(16):2–7, 2006.

[152] Yonghua Wang, Yan Li, Zhi Ma, Wei Yang, and Chunzhi Ai. Mechanism of microRNA-target interaction: Molecular dynamics simulations and thermodynamics analysis. *PLoS Computational Biology*, 6(7):5, 2010.

[153] Donald B Wetlaufer. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proceedings of the National Academy of Sciences of the United States of America,*, 70(3):697–701, 1973.

[154] Erno Wienholds and Ronald H. Plasterk. MicroRNA function in animal development. *FEBS Letters*, 579(26):5911–5922, 2005.

[155] Bruce Wightman, Thomas R. Bürglin, Joseph Gatto, Prema Arasu, and Gary Ruvkun. Negative regulatory sequences in the lin-14 3-untranslated region are necessary to generate a temporal switch during Caenorhabditis elegans development. *Genes and Development*, 5(10):1813–1824, 1991.

[156] Daniela M Witten and Robert Tibshirani. Covariance-regularized regression and classification for high-dimensional problems. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 71(3):615–636, 2009.

[157] Feifei Xiao, Zhixiang Zuo, Guoshuai Cai, Shuli Kang, Xiaolian Gao, and Tongbin Li. miRecords: An integrated resource for microRNA-target interactions. *Nucleic Acids Research*, 37(SUPPL. 1):105–110, 2009.

[158] Shuping Xing, Niklas Wallmeroth, Kenneth W Berendzen, and Christopher Grefen. Techniques for the analysis of protein-protein interactions in Vivo. *Plant Physiology*, 171(2):727–58, 2016.

[159] Jin Xu, Rui Zhang, Yang Shen, Guojing Liu, Xuemei Lu, and Chung-i Wu. The evolution of evolvability in microRNA target sites in vertebrates. *Genome Research*, pages 1810–1816, 2013.

[160] Wenlong Xu, Anthony San Lucas, Zixing Wang, and Yin Liu. Identifying microRNA targets in different gene regions. *BMC bioinformatics*, 15 Suppl 7(7):S4, 2014.

[161] Andrew Yates, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Konstantinos Billis, Denise Carvalho-Silva, Carla Cummins, Peter Clapham, Stephen Fitzgerald, Laurent Gil, Carlos Garcoa Giron, Leo Gordon, Thibaut Hourlier, Sarah E. Hunt, Sophie H.

Janacek, Nathan Johnson, Thomas Juettemann, Stephen Keenan, Ilias Lavidas, Fergal J. Martin, Thomas Maurel, William McLaren, Daniel N. Murphy, Rishi Nag, Michael Nuhn, Anne Parker, Mateus Patricio, Miguel Pignatelli, Matthew Rahtz, Harpreet Singh Riat, Daniel Sheppard, Kieron Taylor, Anja Thormann, Alessandro Vullo, Steven P. Wilder, Amonida Zadissa, Ewan Birney, Jennifer Harrow, Matthieu Muffato, Emily Perry, Magali Ruffier, Giulietta Spudich, Stephen J. Trevanion, Fiona Cunningham, Bronwen L. Aken, Daniel R. Zerbino, and Paul Flicek. Ensembl 2016. *Nucleic Acids Research*, 44(D1):D710–D716, 2016.

[162] Andrew Yates, Kathryn Beal, Stephen Keenan, William McLaren, Miguel Pignatelli, Graham R S Ritchie, Magali Ruffier, Kieron Taylor, Alessandro Vullo, and Paul Flicek. The Ensembl REST API: Ensembl data for any language. *Bioinformatics*, 31(1):143–145, 2015.

[163] Jianxin Yin and Hongzhe Li. A sparse conditional gaussian graphical model for analysis of genetical genomics data. *The Annals of Applied Statistics*, 29(6):997–1003, 2012.

[164] Jingkai Yu, Svetlana Pacifico, Guozhen Liu, and Russell L Finley. DroID: the Drosophila Interactions Database, a comprehensive resource for annotated gene and protein interactions. *BMC Genomics*, 9:461, 2008.

[165] Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[166] Teng Zhang and Hui Zou. Sparse precision matrix estimation via Lasso penalized D-trace loss. *Biometrika*, 101(1):103–120, 2014.