Accurate Prediction of Human miRNA targets via Graph Modeling and Machine Learning Approaches

by

MOHAMMAD MOHEBBI

(Under the Direction of Liming Cai)

Abstract

miRNAs are small endogenous non-coding RNA molecules that have a critical function in suppressing genes and they also correlate with many diseases and cancers. Due to the importance of their effects in several cell activities, discovering their mechanisms is an important task. Because the functionality of miRNAs tightly connected to the way they recognize their targets miRNA target prediction has received a lot of attentions in research. Despite that, most of current methods suffer from high false positive rates and they are not able to provide much insight to the actual process of miRNA targeting.

In this dissertation, we present two novel approaches aimed at addressing existing issues in miRNA target prediction; one approach to improve false positive rate and the other to substantiate multiple hypotheses pertaining to biological mechanism of miRNA targeting and to provide insight into the actual mechanism. To address the first issue, we present Correlation Graph model that captures nucleotide correlations between miRNA sequence and the target. This model makes it possible to characterize nucleotide correlations other than Watson-Crick base pairings between two parts of the duplex. We designed an SVM based algorithm and tested our model on human data and it achieved a sensitivity of 86% with a false positive rate below 13% which is a significant performance improvement in

comparison to the state-of-the-art methods miRanda and RNAhybrid.

The second part of this dissertation addresses the issue of understanding the mechanism of miRNA targeting. It contains a multi-hypothesis learner algorithm that utilizes features collected from literature pertaining to the mechanisms of targeting. These features enable the algorithm to partition data in a way very relevant to the biological features. The algorithm uses these partitions to learn multiple hypotheses. Our evaluations on human and mouse datasets show our method has comparable performance to that of high performance classifiers such as $RandomForest$. Moreover, feature selection on the resulting partitions confirms that the partitioning mechanism is compatible with biological mechanisms. These partitions could be used for further in vivo experiments to verify the currently proposed targeting approaches and to discover the new mechanisms.

Index words: miRNA target prediction, miRNA, Machine Learning, Graph Modeling, Feature selection, Re-sampling

Accurate Prediction of Human miRNA targets via Graph Modeling and Machine Learning Approaches

by

MOHAMMAD MOHEBBI

B.S., Yazd University, Iran, 2002

M.S., Sharif University of Technology, Iran, 2005

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2017

Accurate Prediction of Human miRNA targets via Graph Modeling and Machine Learning Approaches

by

MOHAMMAD MOHEBBI

Major Professor:	Liming Cai
Committee:	Russell L. Malmberg
	Khaled Rasheed

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2017

Acknowledgments

First and foremost, I want to thank you my advisor Dr. Liming Cai and Dr. Russell L. Malmberg for being very encouraging, positive, and patient during my PhD program and research. They have been very encouraging for every inch of progress that I made. I am thankful to have them as examples for my future career as an assistant professor and for the way I need to encourage students for their courses and researches.

I would also like to express my appreciation to Dr. Khaled Rasheed for being always available for me to discuss questions either for the machine learning course or research projects for this dissertation. In addition, I would like to thank Dr. Liang Ding, our former colleague at RNA-Informatics lab, who had been very helpful to everyone and me during different research projects. Finally , I would like to thank Ms. Khin Khine for kindly proofreading some chapters of this dissertation and very helpful suggestions she has made.

Table of Contents

List of Figures

Page

List of Tables

Chapter 1

Introduction

In the last two decades, several hundred genes in mammalians and plants have been discovered that contain a double-stranded RNA consisting of 60-80 nucleotides, called precursor-microRNA (pre-miRNA). Pre-miRNAs form a hair-pin structure, which are cleaved by Dicer after being transferred from nucleus to the cytoplasm. Mature miRNAs, or simply miRNAs, are about 22 nucleotides length RNAs made from pre-miRNAs [7].

miRNA genes compose about 1%-2% of genes in eukaryotes [34]. miRNAs have been correlated with many critical cell processes such as proliferation, differentiation, cell death, growth control, and developmental timing. Therefore, discovering their functions is very important [4].

1.1 miRNA Functionality Mechanism

miRNAs are thought to down-regulate the translation of messenger RNAs (mRNAs). Their mechanism to regulate a gene includes complementarity binding to the 3'-untranslated regions (3'-UTR) of the mRNA. This binding is complete in plants and partial in mammalians [5]. Their mechanism for regulating genes is either through inhibiting translation of messenger or cleaving the targeted gene [32, 17]. The exact mechanism of regulating repression is not fully understood yet [32].

Binding between miRNA and mRNA gene prevents ribosomes from becoming associated with the mRNAs, which blocks protein production by suppressing protein synthesis and/or

by initiating mRNA degradation [32]. Since most target sites on the mammalians mRNA have only partial base complementarity with their corresponding miRNA, each miRNA may potentially target about 300 to 400 different mRNAs [72]. In addition, each mRNA gene may have multiple binding sites for different miRNAs.

1.2 miRNA Functionality Study

The study of miRNA functionality relies heavily on the identification of their targeted genes and the location of site on the target gene. In spite of their significance, only about 1000 human miRNA target genes have been recognized in vivo. In comparison to the potential number of human gene targets, mechanisms of most of miRNAs are still poorly understood [77].

Using in vivo methods to recognize targets are very slow and costly; therefore it cannot be the only source of miRNA target identification. This drawback of experimental methods can potentially be remedied by computational and bioinformatics methods. In the last decade dozens of algorithms, with a variety of approaches and techniques have been developed [75]. These methods are either specific for a few species or general for any species; Methods for vertebrates are TargetScan and TargetScanS [48, 46], miRanda [20, 34], DIANA-microT [39] and RNAhybrid [63] for flies. Some of general tools are miTarget [38] and MicroInspector [68].

1.3 miRNA Target Prediction Is Difficult

Interaction of miRNA and mRNAs is a complex and largely unknown phenomenon and researchers believe that just complementary sequence matching of miRNA and mRNAs may not enough for target recognition. This problem can be even worse in vertebrates as they do not have continuous binding and perfect base pairing in the duplex between miRNA and target site. This makes computational prediction of miRNA targets a challenging task in

vertebrates causing almost all computational methods suffer from high false positive rate.

1.4 Current Approaches

To address this problem, one approach is to include more parameters such as UTR sequence context, free energy of complexes, and evolutionarily related sequence comparison. The idea for the later one, is that orthologous sequences of miRNA target sites are conserved in evolutionarily distant species such as human, mouse or fish. Friedman et al. proposed that conserved miRNAs have preserved sites in most of mammalian targets [24]. However, there are two main critics to this approach. One, some conserved 3'-UTR might have a number of non-conserved target sites. Second, there is a huge number of non-conserved miRNAs, which indicates this method does not work.

Another approach is to feed computational methods with better and more comprehensive datasets. This happens in light of the emergence of more advanced sequencing technologies and high throughput experimental methods. Data driven methods, which mainly are based on machine learning and statistical methods, could also be improved by development of more accurate algorithms.

1.5 Data Set Selection

To choose our dataset, we searched for an experimentally validated and well-known database. mirTarBase [30] with first release on 2010 and second on 2014, have been cited more than 600 times to date. mirTarBase contains more than three hundred and sixty thousand miRNA-target pairs for about twenty species, such as human, mouse, rat, chicken, etc. Pairs are selected from articles manually and with systematic text processing methods. Experimental methods used in articles for finding targets include iPAR-CLIP, PAR-CLIP, CLASH, HITS-CLIP, reporter assay, qRT-PCR, western blot, microarray, next-generation sequencing experiments, pSILAC and others. mirTarBase may be the most up-to-date resource for

miRNA targets [29].

1.6 Contribution of The Dissertation

In this dissertation, we address two challenges regarding the miRNA target prediction problem, i.e, high false positive rate and lack of understanding of mechanism of miRNA targeting. We present two novel approaches; one is Correlation Graph model that aims to improve false positive rate and the other approach to learn multiple hypotheses pertaining to the biological mechanism of targeting. The goal of this research is to provide insight into the actual mechanism, in addition to improving prediction performance.

1.7 Dissertation Outline

We first review some preliminary terminologies in Chapter 2. In Chapter 3, we introduce Correlation Graph model, which is a bipartite graph to characterize interactions across two sequences miRNA and target. This model aims to capture nucleotide correlations beyond RNA secondary structure base-pairs. The graph is used as input to an SVM model which, together with a re-sampling approach, boosts the performance of the model. Chapter 4 introduces our Multi-hypothesis learner algorithm. The algorithm is designed for the cases that there are several underlying and unknown hypotheses to be learned according to the several patterns of the data. Moreover, we selected features from biology literature which exist in different mechanisms of miRNA targeting. Given the fact that current hypotheses for targeting are not well descriptive, it is expected that our model could help clarifying and verifying the proposed ideas. In addition, the sequences provided by the model (which do not correlate with proposed mechanisms) could be utilized for lab experiments to discover new targeting methods. Chapter 5 concludes the dissertation.

Chapter 2

Preliminaries

2.1 RNA

RNAs (Ribonucleic Acids) are biomolecules that are composed of nucleotides. Nucleotides are building blocks of nucleic acids, which are in four types; Adenine (A), Guanine (G), Cytosine (C), and Uracil (U) [61]. DNA, RNA and proteins are fundamental and vital molecules for majority of living cells on the planet [35]. DNA sequences contain information about the cell development and it's function. RNAs carry the information to biological machines inside cell which produce proteins based on the RNA information. Here is an example of an RNA sequence:

UGGGAUGAGGUAGUAGGUUGUAUAGUUUUAGGGUC

2.2 Base-pair

A nucleotide can interact and bind to other nucleotides through their chemical bonds, in such a case, the two nucleotides form a pair, called $base-pair$. Pairing of A to U, C to G and vice versa are called $Watson-Crick$ base-pairs. Nucleotide G can bind to U to form $Wobble$ base-pair. Watson-Crick base-pairs are shown with a line connecting nucleotides, while for Wobble base-pairs they are shown by a colon :, on the other words they are denoted with A-U, C-G, G:U and vice versa. Figure 2.1 illustrates base-pairs within a sequence structure.

Watson-Crick and Wobble base-pairs are called *canonical* base-pairs. However, *in vivo*

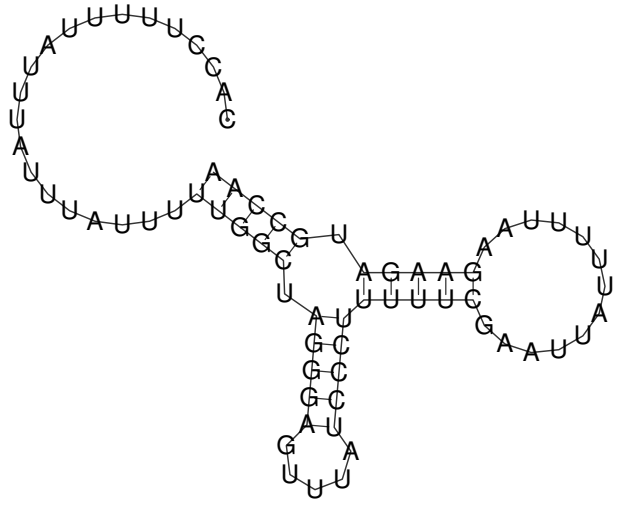

Figure 2.1: Secondary structure of an RNA sequence composed of *stem*s and *loop*s. It has been created by RNAfold web-server.

nucleotides can pair with each other in all possible 16 combinations of these four nucleotides A, C, G and U. Those other base-pairs beyond the canonical ones are named *non−canonicals*.

2.3 Hierarchy of RNA Structures

RNA structures are studied in three levels of abstractions: primary, secondary and tertiary. Representation of a sequence, as a list of bases, is called primary structure. Base-pairs make a sequence to fold over itself and represent additional information about the molecule related to the sequence. If we allow just canonical base-pairs, the structure is called *secondary structure* which contains *loop*s and *stem*s.

An *stem* is the part of structure with *stacked* base-pairs: i.e. adjacent base-pairs forming a ladder shape. A *loop* is a subsequence that none of its bases is part of a base-pair [18]. An example of secondary structure is depicted at figure 2.1.

In *vivo*, an RNA molecule forms a complex 3D shape. To model such a structure canonical

base-pairs are not enough and we need to consider non-canonical interactions. The structure of a sequence containing possible canonical and non-canonical interactions, is called *tertiary structure*.

2.4 Minimum Free Energy (MFE)

Based on the second law of thermodynamics, closed systems tend to reach to an equilibrium form which has minimum internal energy or so called Minimum Free Energy (MFE) [57]. It is believed that a structure with minimum free energy is in stable state. The minimum free energy structure of a sequence is the structure with the lowest value of Gibbs free energy; and it is the most likely structure in theory and necessarily it is not the structure that is formed in nature [23].

2.5 MicroRNA (miRNA)

MicroRNAs or miRNAs are small non-coding RNA molecules with about 22 nucleotides which regulate genes in mammalians, plants and some viruses [58]. They bind and interact with mRNAs to control proteins generated from the information in mRNAs, in other words they *regulate* genes.

2.6 mRNA

Messenger RNAs (mRNAs) are RNA molecules that contain genetic information from DNA to convey to the ribosomal RNA (rRNA), the protein-manufacturing machinery. The ribosome structure generates proteins based on the information from mRNA sequence [56].

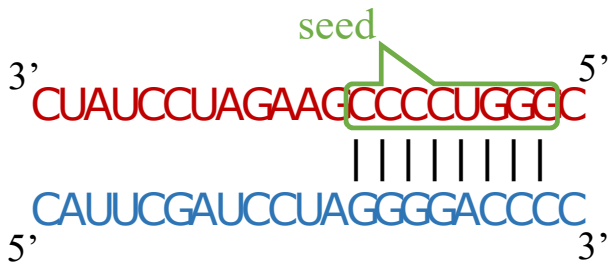

Figure 2.2: miRNA Seed; the eight consecutive nucleotides from second index starting on 5'side of the miRNA sequence is called *seed*.

2.7 miRNA Target and miRNA Duplex

miRNA sequence could bind through complementary base-pairs over several locations on mRNA sequence. The subsequence of mRNA that a miRNA binds over is called the miRNA target or miRNA target site. In this dissertation both terms are used interchangeably and equivalently. When miRNA binds to a target site, they form a duplex, named $miRNAduplex$.

2.8 miRNA Seed

Eight consecutive nucleotides from position 2 to 9 on 5'end of the miRNA sequence is called *seed* (Figure 2.2) [48]. Biologists believe that the process of binding start from seed part of a miRNA [69]. When the binding in seed region is continuous and it is about 6 to 8 base-pairs, it is called a canonical seed, otherwise it is called non-canonical [51]. Seed binding is considered as the most important identifier for miRNA targets in mammals [62].

2.9 Current computational methods and algorithms

The early computational methods for target recognition were rule based, using a set of discriminative rules derived from experimental and biological knowledge, such as MFE, duplex

binding pattern or target accessibility [77]. Some popular rule based tools are RNAhybrid [63], TargetScan [48] and miRanda [20, 34]. In the last several years, data driven methods became popular due to advances in sequencing technology and emergence of relevant data sets. These methods use sophisticated machine learning and statistical models to learn more discriminative features for target identification [77]. Some of data driven methods are TargetSpy [72], miRanda-mirSVR [9] and Avishkar [26].

In the last decade many research on animal's miRNA target prediction has been conducted and a variety of statistical models such as Bayesian networks, hidden Markov models and machine learning approaches have been applied, yet benchmark evaluations show they have mediocre performances [16].

2.10 Tools and Programs Referred In This Dissertation

2.10.1 miRanda

miRanda is a program that detects potential miRNA target sites for a given miRNA sequence and an mRNA sequence. Potential targets could be discriminated by two scoring values provided by the tool; one is a minimum free energy value and the other is the miRanda score. While the lower minimum free energy is a better identifier of targets, for the miRanda score the higher is the stronger evidence for the target. Usually targets have a miRanda score between 100 to 140 and beyond. The program outputs a list of potential candidates sorted in descending order by miRanda score. Choosing some cut-off thresholds in program options could shorten this list for large mRNA sequences [59].

2.10.2 RNAfold

RNAfold is a program that calculates the minimum free energy secondary structures for a given RNA sequence. It has the ability to predict the secondary structure of several

sequences; the structures are printed in *dot – bracket* notations. Pairing brackets show nucleotides interacting with each other and a dot means the corresponding nucleotide does not pair with any other bases. It is a very renown, powerful and flexible tool for RNA secondary prediction. With a "*constraints*" input file, one can set certain conditions to restrict some bases from interacting with others or with specific locations of the sequence. Similarly, the user can force some nucleotides to pair and let the program to predict the rest of structure. RNAfold can print out the structure in PostScript format too, given the appropriate argument options [65].

2.10.3 RNAcofold

It is a variation of RNAfold which predicts the secondary structure of two RNA sequences as a dimer structure. These sequences must be concatenated by the separator character "&". The output options are similar to that of RNAfold [64].

2.10.4 RNAhybrid

RNAybrid finds the minimum free energy hybridization structure of a short and a long RNA sequences. To do so, RNAhybrid finds the subsequence of the long RNA to unwrap its hydrogen bonds, and let the short sequence to bind on the subsequence, such that the whole structure reach to the lowest possible energy. The main purpose of this program is to predict miRNA targets on a mRNA [41, 66].

2.11 Performance Evaluation methods and metrics

The problem of miRNA target prediction is a binary classification question. In other words for a sample containing a pair of miRNA and a candidate sequence as target we are interested to know if they bind or not, i.e. if the sample is labeled positive or negative. For evaluation of a binary classification method usually two parameters sensitivity and specificity are used.

2.11.1 Sensitivity and Specificity

Sensitivity also called the true positive rate, the probability of detection [71] or recall, is the percentage of correctly predicted positives out of total number of available positives. Specificity, also called the true negative rate as well, represents the proportion of negatives correctly predicted as such, from the set of all available negatives in the evaluation set.

For some bioinformatics problems, such as miRNA target prediction two other parameters also are used: the false positive rate and false negative rate. The former represents the number of cases predicted as target while lab experiments reject them, and the latter measures the ratio of targets exist *in vivo*, which could not be predicted by the computational method. In search for all potential targets for a given miRNA, having higher sensitivity and tolerating larger false positive rate are preferred. On other hand, increasing specificity could be helpful to check miRNAs regulating a single gene[53].

2.11.2 MCC (Matthews correlation coefficient)

MCC is used to measure the quality of a binary classifier and it can be used for very imbalanced two-class datasets. The MCC is the correlation coefficient between observed and predicted labels and it is a real value between −1 to +1. A value +1 means a perfect prediction. On the other hand, −1 implies that the prediction method works completely in reverse: all positives are predicted as negative and all negatives is labeled as positive. An MCC = 0 means the classifier is as good as random labeling of the evaluation set [55].

2.11.3 ROC (Receiver Operating Characteristic) Curve

ROC curve is a graphical plot that illustrates the performance of a binary classifier when the separating threshold is varied. As depicted in figure 2.3, the vertical axis for the curve is true positive rate versus the horizontal axis is false positive rate. These axes are also labeled as sensitivity or recall versus (1-specificity). It is used for comparing the performance of

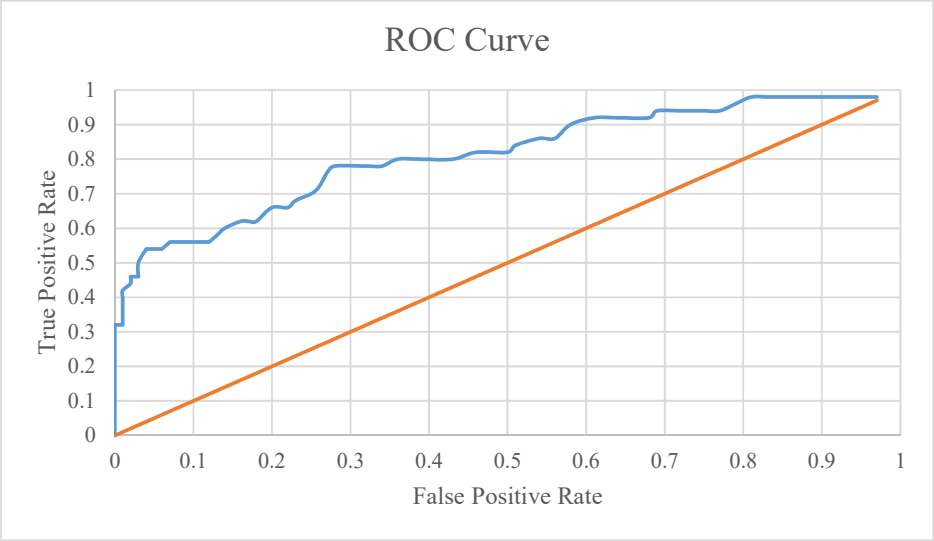

Figure 2.3: ROC (Receiver Operating Characteristic) Curve, a graphical plot illustrating the performance of a binary classifier in terms of true positive rate as a function of false positive rate.

several classifiers while the discriminating threshold varies identically for all of them [42].

2.11.4 AUC

AUC is the abbreviation for Area Under the Curve of ROC. As true positive rate and false positive rate, range from 0 to 1, then the maximum AUC is 1 for a perfect predictor; a classifier with true positive rate one and zero false positive rate. An AUC value 0.5 implies a random classifier and any value < 0.5 shows that predictor works with some degree of disagreement with the observed labels. AUC = 0 means a total disagreement between predicted and observed labels [67].

2.12 Graph and Bipartite Graphs

In graph theory, graphs are represented as $G = (V, E)$, where V is the set of vertices and E is the set of edges connecting vertices. A graph is simple when there is no such an edge

connecting a node to itself [15]. A simple graph $G = (V, E)$ is a bipartite graph if $E \subseteq V_1 \times V_2$ where $V_1 \cup V_2 = V$ and $V_1 \cap V_2 = \varnothing$. The bipartite graph G is weighted if every edge is associated with a numerical weight $\omega : E \to R$ [11].

2.13 Machine Learning

Machine learning is the subfield of computer science that enables computers to do some jobs without explicitly programmed. In the past decade, machine learning has found its way to many research areas such as pattern recognition, image processing, speech recognition, self-driving cars and bioinformatics [43]. There are two types of learning: Supervised and Unsupervised.

2.13.1 Supervised Learning

Supervised learning is the task of calculating a function from labeled training data. In other words, we have a set of samples which each sample is a vector of characteristics plus a label. The supervised learning is about to learn a function that predicts the sample label based on the characteristics. Some of the popular supervised learning algorithms are Support vector machines (SVM), RandomForests, Artificial Neural Network (ANN), Decision Tree Learner, Nave Bayes Classifier and etc [44].

2.13.2 Unsupervised Learning

Unsupervised Learners are types of machine learning algorithms aim to draw inferences from a training dataset which samples are not labeled. The major method for unsupervised learning is clustering and as examples of algorithms in this category, one can refer to K-Means algorithm, Hierarchical clustering, Hidden Markov models and so on [45].

2.13.3 Kernel Function

Generally, and in introductory level of machine learning, it is assumed that all samples have a fixed-size feature vector. In reality we may face problems that samples could not be represented by a fixed length of features, for example when samples are molecular structures, protein sequences or evolutionary trees. One approach to solve this problem is using some functions, so called $KernelFunction$, to map these variable length samples into a space with a fixed dimension [60].

Another application of kernel functions is when a classifier such as SVM cannot find an appropriate hyperplane hypothesis to separate classes. In such a case, a kernel function could be used to transfer the data into higher dimensions so that the hyperplane with decent performance could be calculated with the SVM tool [70]. A commonly used kernel with SVMs is RBF (Radial Basis Function) [36].

Chapter 3

ACCURATE PREDICTION OF HUMAN MIRNA TARGETS VIA GRAPH MODELING OF MIRNA-TARGET DUPLEX[1]

[1]M. Mohebbi, L. Ding, R. L. Malmberg, C. Momany, K. Rasheed and L. Cai. Submitted to *Journal of Bioinformatics and Computational Biology*, 04/19/2017.

Abstract

miRNAs are involved in many critical cellular activities through binding to their mRNA targets, e.g., in cell proliferation, differentiation, death, growth control, and developmental timing. Accurate prediction of miRNA targets can assist efficient experimental investigations on the functional roles of miRNAs. The prediction, however, remains a challenge task due to the lack of experimental data about the tertiary structure of miRNA-target binding duplexes. In particular, correlations of nucleotides in the binding duplexes may not be limited to the canonical Watson Crick base pairs as they have been perceived; methods based on secondary structure prediction (typically minimum free energy) have only had mix success. In this work, we characterized miRNA binding duplexes with a graph model to capture the correlations between pairs of nucleotides on a miRNA and its target sequences. We developed machine learning algorithms to train the graph model to predict the target sites of miRNAs. In particular, because imbalance between positive and negative samples can significantly deteriorate the performance of machine learning methods, we designed a novel method to re-sample available dataset to produce more informative data the learning process.

We evaluated our model and miRNA target prediction method on human miRNAs and target data obtained from mirTarBase, a database of experimentally verified miRNA-target interactions. The performance of our method in target prediction achieved a sensitivity of 86% with a false positive rate below 13%. In comparison with the state-of-the-art methods miRanda and RNAhybrid on the test data, our method outperforms both of them by a significant margin.

The source codes, test sets and model files all are available at
http://rna-informatics.uga.edu/?f=software&p=GraB-miTarget.

3.1 Introduction

In the last two decades, several hundred genes have been characterized that contain a double-strand RNA of length 60-80 nucleotides, called precursor-microRNA (pre-miRNA). Mature

miRNAs, or simply miRNAs, are about 22 nucleotides long RNAs made from pre-miRNAs [50]. miRNAs have been found to be involved in a number of critical cellular processes such as proliferation, differentiation, cell death, growth control and developmental timing. Therefore discovering their functions is very important [31].

miRNAs are believed to down-regulate the translation of messenger RNAs (mRNAs). Their mechanism to regulate a gene, includes complementarity binding to the 3'-untranslated regions (3'-UTR) of the mRNA. This binding is complete in plants and partial in mammals [6]. Their mechanism for regulating genes is either through inhibiting translation of messenger or cleaving the targeted gene [40]. Binding between miRNA and mRNA gene prevents ribosomes from becoming associated with the mRNAs, blocking protein production by suppressing protein synthesis and/or by initiating mRNA degradation [22]. Since most target sites on the mammal's mRNAs have only partial base complementarity with their corresponding miRNA, each miRNA may potentially target about 300 to 400 different mRNAs [72]. In addition, each mRNA gene may have multiple binding sites for different miRNAs.

In contrasts to the potential number of human gene targets, the functions of most miRNAs are still poorly understood. The study of miRNA functionality relies heavily on the identification of their targeted genes and the location of the binding site on the target gene. In spite of their significance, only about 1000 human miRNA target genes have been verified in vivo [77]. In vivo methods to recognize targets are very slow and costly; therefore they cannot be the only source of miRNA target identification. This drawback of experimental methods can potentially be covered by computational and bioinformatics methods. In particular, in the last decade dozens of algorithms, with a variety of approaches and techniques, have been developed. These methods are either specific for a few species or general for any species. For example, methods for vertebrates include TargetScan [1], miRanda [10], DIANA-microT [54], and RNAhybrid [63], while general tools for broader groups of organisms include miTarget [38] and PITA [37].

Due to the lack of a full understanding of the biological process and mechanism in miRNA

functioning, the state-of-the-art computational methods for miRNA target prediction have largely been data-driven and machine learning based, with such representative tools as TargetSpy [72], miRanda-mirSVR [9], and Avishkar [26]. This is indeed the case for organisms like vertebrates where non-continuous bindings and non-perfect base-pairings between miRNAs and mRNA targets have often been found. Rules for such binding patterns remain elusive. Though sophisticated machine learning techniques have been adopted with the goal to overcome the difficulty of deterministic rules, these previous methods for miRNA target prediction usually yield false positive rates much higher than expected.

Without doubt, the quality of modeling based on machine learning relies heavily on availability of pertinent data, typically positive and negative samples for classification. As more and more efforts have been invested in miRNA research, the knowledge about miRNA and their targets has grown dramatically. For example, mirTarBase [30] is an experimentally validated miRNA target database that contains more than 360,000 miRNA-target interactions for 18 species (including human), these duplexes are natural candidates for positive samples. On the other hand, however, negative samples usually are not directly available. In theory, any stretch other than the real target of an appropriate length in the 3'-UTR of a targeted mRNA gene may be considered a negative target of the corresponding miRNA. The previous work has mostly chosen sequences as negative targets which were randomly sampled in this manner.

We have analyzed all such negatives for each positive target sample from mirTarBase in terms of minimum free energy (MFE) and the number of canonical base pairs in the involved secondary structure. Our study reveals a serious issue with such negative samples which may explain the under-performance of the previous methods in miRNA target prediction. Figures 3.1 and 3.2, respectively, show the distribution of MFE and the number of canonical base pairs for the secondary structure of all negatives versus positives. These data show that random selection of negatives is more prone toward selecting negatives with a higher MFE or a fewer number of canonical base pairs involved. Such sequences are easier to be

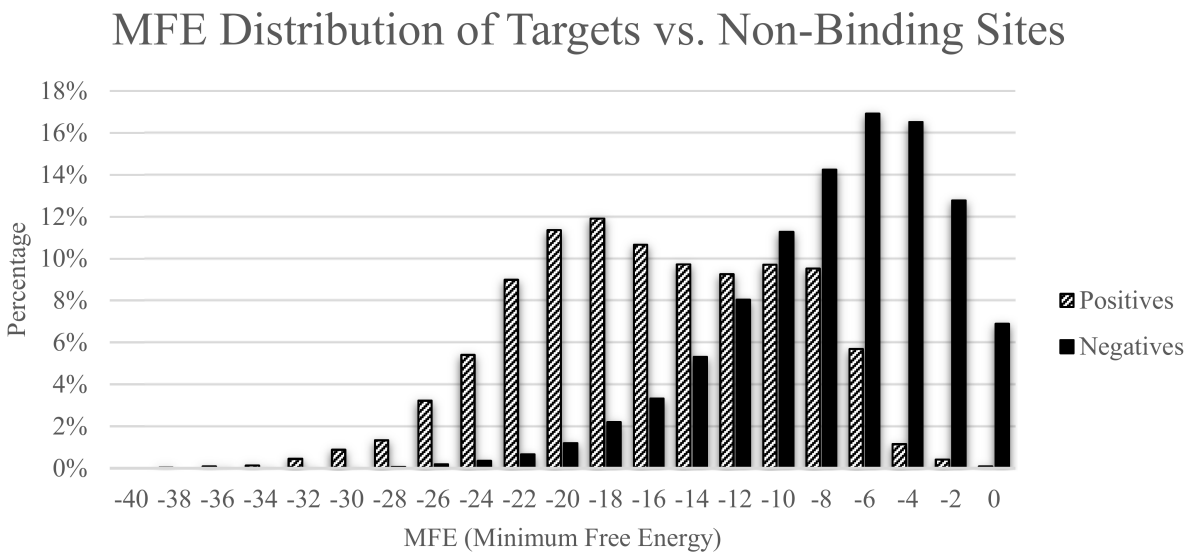

Figure 3.1: Minimum free energy (MFE) distribution of positive samples vs. all possible negative samples. Positive samples are experimentally verified target sites from MirTarBase and negative samples are all other locations on the corresponding mRNA sequences which do not bind to a miRNA. MFE of binding duplex of miRNA sequence and a site were computed by RNAcofold. The large overlap percentage of positive and negative samples demonstrates the limit of MFE based methods for miRNA target prediction.

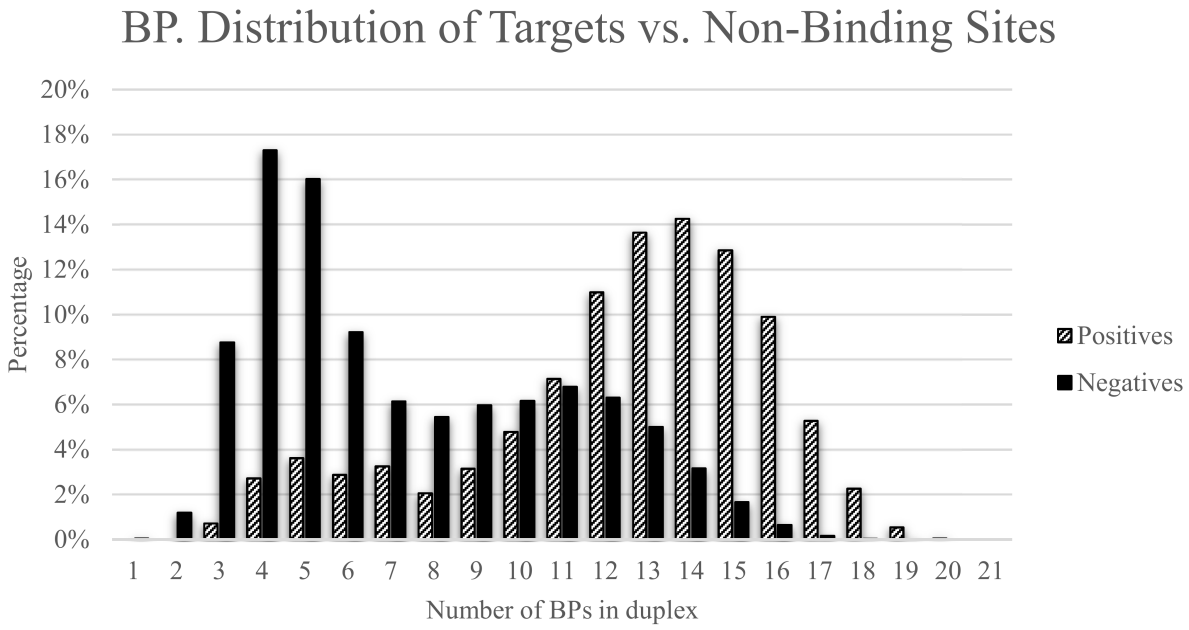

Figure 3.2: Number of BPs (Base-Pair) distribution for positive samples vs. all possible negative samples. Positive samples are real target sites from MirTarBase and negative samples are non-binding locations on corresponding mRNAs. For a given miRNA and site sequence, RNAcofold predicts secondary structure of binding. The number of BPs in secondary structure of a binding were counted for real targets and non-binding sites. Overlap of positive and negative bars shows the limitation of current secondary structure prediction methods for target prediction.

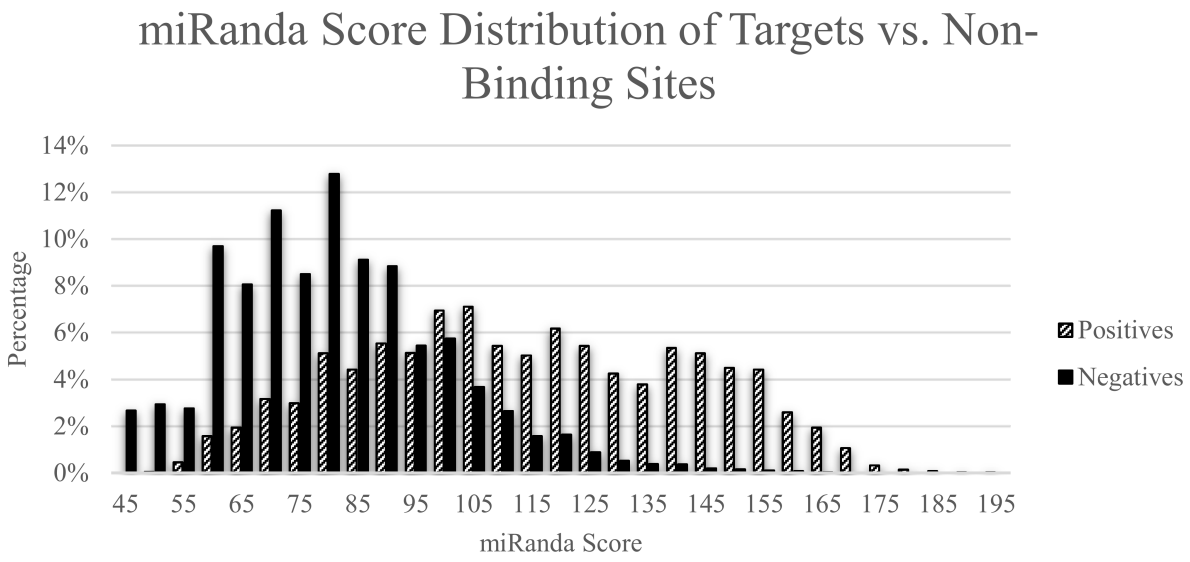

Figure 3.3: A result of miRanda on experimentally verified human target sites (from MirTarBase) and non-binding locations on corresponding mRNAs. A large overlap of scores for positive and negative sites shows that miRanda can also suffer from high false positive rate on an experimental dataset.

distinguished from positives; we call such negative samples easy negatives. Nevertheless, there are negatives with MFE and patterns of canonical base pairs much more similar to a positive sample. Such samples prove very difficult to be distinguished from positive samples as shown by the previous miRNA target prediction methods based on MFE or binding pattern of miRNA-target duplex. We call such samples hard negatives. These hard negatives have caused the previous methods to yield high false positive rates. For example, Figure 3.3 shows a large overlap of miRanda [10] scores between positive and negative samples. Negative samples contributing to this overlap region apparently deteriorate the performance of the prediction methods.

The computational prediction of miRNA targets in vertebrates is challenging due to the non-continuous binding and non-perfect base-pairing in the duplex between a miRNA and its mRNA target. It is likely the reason that almost all computational methods suffer from high

false positive rate. The emergence of more advanced sequencing techniques, and therefore better related data sets, along with recent advances in machine learning methods, could lead to the development of more accurate algorithms.

The early computational methods for target recognition were rule based, i.e., they had a set of discriminative rules derived from experimental and biological knowledge, such as minimum free energy (MFE), duplex binding pattern, or target accessibility. Some popular rule based tools are RNAhybrid, TargetScan, and miRanda. In the last several years, in light of emerging relevant data sets, data driven methods became popular. These methods use sophisticated machine learning and statistical models to learn more discriminative features for target identification [77]. TargetSpy [72], miRanda-mirSVR [9] and Avishkar [26].

In this work, we introduce a new machine learning based method for accurate prediction of human miRNA targets. This method consists of two novel components: the identification of more informative, hard negative miRNA-target duplex samples and the extraction of more relevant features from experimentally validated samples for machine learning. We demonstrate that these strategies make it possible to yield effective prediction of human miRNA targets with a performance exceeding those of the state-of-the-art methods by a significant margin.

3.2 The Method and Model

Our method for target prediction for human miRNA is based machine learning where both positive and negative samples are used to construct an effective model for the duplex between miRNAs and their targets. In particular, our method has two main novel components: negative sample identification and a graphical model for the duplex, upon which an effective machine learning algorithm can be developed for miRNA target prediction.

3.2.1 Identification of Hard Negatives

We established our dataset based on mirTarBase [30], an experimentally validated miRNA target database. First released in 2010 and second in 2014, mirTarBase contains more than 360,000 miRNA-target duplexes for 18 species. Duplexes were selected from research publications via manual curation as well as systematic text processing methods. mirTarBase may be the most up-to-date resource for miRNA targets and it has been cited more than 600 times to date. We extracted 3,684 experimentally supported human miRNA-target duplexes whose secondary structures have been provided in research articles. Such duplexes were selected as positive samples for our machine learning method.

Negative samples are not directly available from miTarbase. Based on our analyses in the previous section, we have developed a novel method to generate effective negative samples. Theoretically, any stretch of an appropriate length other than the real target in the 3'-UTR of a targeted mRNA gene can be considered a negative target of the corresponding miRNA. In most previous work, such negative targets were sampled randomly. To develop a more effective way for identifying negative samples, we analyzed all such negatives for each positive target sample for MFE and the number of canonical base pairs of the involved secondary structure. Figures 3.1 and 3.2, respectively, show the distribution of MFE and the number of canonical base pairs for the secondary structure of all negatives versus positives. These data show that random selection of negatives is more prone toward selecting negatives with a higher MFE or a lesser number of canonical base pairs, which are easier to be distinguished from positives. We call such negative samples easy negatives. Nevertheless, there are negatives with MFE and patterns of canonical base pairs more similar to a positive sample. Such samples prove very difficult for target prediction methods based on MFE or binding pattern of duplex to distinguish from positive targets. We call such samples hard negatives. These hard negatives cause rule-based methods, such as miRanda, to have high false positive rates. Figure 3.3 shows a large overlap of miRanda scores between positive and negative samples.

Negative samples related to this overlap deteriorate the performance of miRanda on the current dataset. Our work has strived to include hard negatives as well as easy negatives. Hard negatives are very important for two reasons. First, from a machine learning point of view, more information can be learned from hard negative samples that appear similar to a positive sample than from those easy negatives that are drastically different from positives.

Second, hard negatives have resulted in high false positive rates for most of the rule based methods that use MFE or canonical base pair binding patterns for target identification. Figures 3.1 and 3.2 suggest that more features are necessary and feature selections can benefit greatly from hard negatives.

We define sequence dissimilarity between a positive target P and a negative site N of the same length randomly sampled from the 3'-UTR of the corresponding targeted mRNA gene such that $P \cap N = \varnothing$. In particular, formula (3.1) defines the dissimilarity as a weighted distance function $\mathrm{DIFF}(P, N)$, where p_i and n_i are the ith nucleotide in P and N respectively, $1 \leq i \leq |P|$, and the value of $\mathrm{diff}(p_i, n_i)$ is 1 if $p_i = n_i$, otherwise it is 0. The weight value w_i is the probability that the ith nucleotide in an miRNA target sequence is involved in a canonical base pair. The DIFF function measures how different is the given negative site from the positive site. Negative sites with lower DIFF values are harder to distinguish from a positive site; they are hard negatives while easy negatives are those with high DIFF value. The distribution of DIFF function values is shown in Figure 3.4. It shows the reason that random selection of negatives might not be the best strategy, because easy negatives have a much higher chance to be selected.

$$\mathrm{DIFF}(P, N) = \sum_{i=1}^{|P|} \mathrm{diff}(p_i, n_i) \times w_i \tag{3.1}$$

To ensure that harder negatives are also properly included for training purposes, we sort all negative samples by the non-decreasing order of their DIFF values with the corresponding positive target samples. The DIFF values are then partitioned into I intervals, where integer I is a chosen parameter, with the intervals being indexed by $k = 0, 1, \ldots, I-1$. Let $\mathcal{N}$ be the

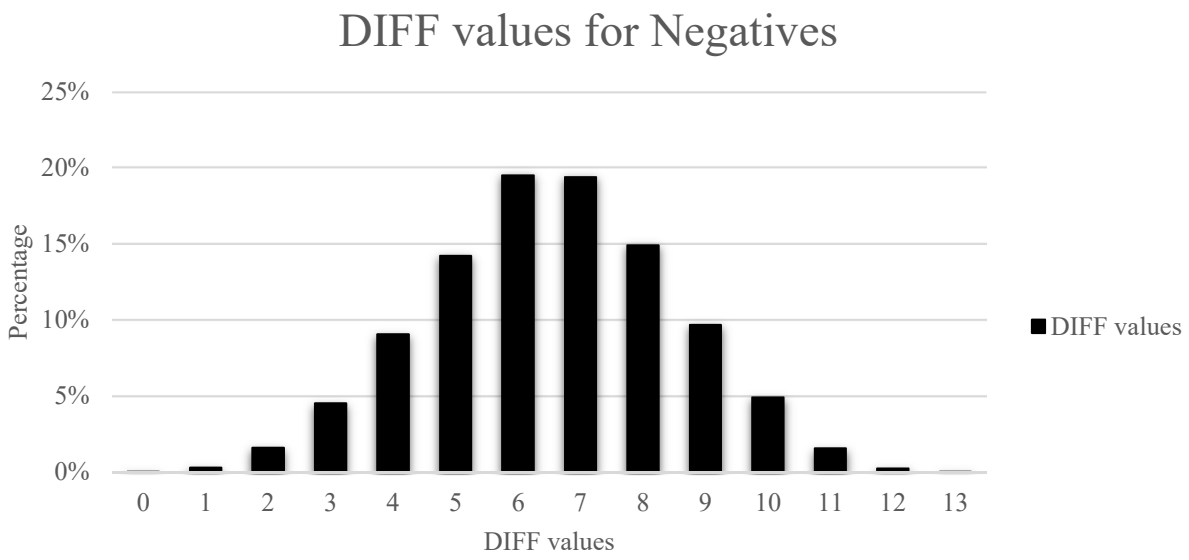

Figure 3.4: Distribution of DIFF values between the negative targets and the corresponding positive targets. The negative samples were obtained based on the method introduced in section II.A. This is a normal distribution with mean of 6.05 and standard deviation of 1.93. We define hard negatives as those with 2 standard deviations away from the mean value.

number of negative samples per positive target site. To ensure to learn from all variety of negatives, we choose at least two samples from every interval. The rest of $(\mathcal{N}-2\times I)$ negative samples are chosen by the Power Law distribution from the I intervals, with $(\mathcal{N}-2\times I)/(4+k)$ being the number of negative samples chosen from the interval with index k. This selection approach favors hard negatives, i.e., those with lower DIFF values (with their corresponding positives).

3.2.2 Correlation Graph Model

Our method exploits the informational data of experimentally validated miRNA-target duplexes to extract features correlations between nucleotides across a miRNA and its target. The seed of a miRNA consists of the nucleotides 2-9 from the 5' end of the miRNA [1]. It is believed that the process of nucleotide binding between the miRNA and its mRNA target

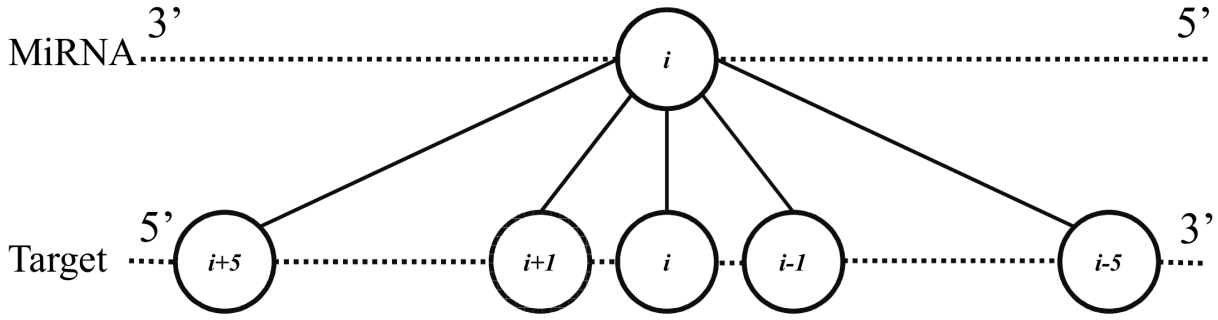

Figure 3.5: The correlation graph model for miRNA-target duplex: Each vertex represent a nucleotide on miRNA or target site. Vertices are increasingly indexed from 5' to 3' for the miRNA sequence and increasingly from 3' to 5' for the target site sequence. Vertices with the same index in miRNA and target site, are across from each other. Each nucleotide i on miRNA has edges to all nucleotides from $i-5$ to $i+5$ on target site to model all possible interactions including Watson-Crick, wobble, and other tertiary interactions.

starts from the 3' end of the target sequence [69]. When the binding in the seed region is continuous for 6 to 8 bps, it is called a canonical seed; otherwise it is called non-canonical [51]. Though the seed binding is considered the most important identifier for miRNA targets in mammals [62], in our study we consider correlations that occur not just in the seed region but also in all other regions across a miRNA and its target.

A simple graph $G = (V, E)$ is a bipartite graph if $E \subseteq V_1 \times V_2$ where $V_1 \cup V_2 = V$ and $V_1 \cap V_2 = \varnothing$. The bipartite graph G is weighted if every edge is associated with a numerical weight $\omega : E \to R$.

Definition A correlation graph model for miRNA target duplex is a weight bipartite graph $G = (V_1 \cup V_2, E)$ in which $V_1 = \{v_1, \ldots, v_n\}$ and $V_2 = \{u_1, \ldots, u_m\}$ such that vertices in V_1 represent nucleotides in the miRNA sequence, indexed increasingly from the 5' side, and vertices in V_2 represent nucleotides in the mRNA target site sequence, indexed increasingly from the 3' side. In addition, edges $(v_i, u_j) \in E$ for every i and j that satisfy $i - 5 \leq j \leq i + 5$.

We also generalize the notion of correlation graph to a profile correlation graph model in which we associate every edge (v_i, u_j) with a probability distribution $P(i, j)$, instead of a

single value, between the ith nucleotide of a miRNA and the jth nucleotide of a target of the miRNA. Technically, such a probability distribution can be calculated from any set of duplex samples of interest, and it can be represented as a 4×4 matrix, detailing the joint probability between every pair of the 4 nucleotide `A`, `C`, `G`, and `U` which may occur simultaneously in the position i of the miRNA and position j of the target.

For any specific set of miRNA-target duplex samples, a profile correlation graph model can be built for training a support vector machine (SVM) classifier that will predict target sites for given miRNAs. In particular, for a pair (g, s), of a known miRNA g and an unknown sequence s, a correlation graph $G_{g,s}$ can be constructed in which the weight of every edge (v_i, u_j) is drawn from the probability distribution of the corresponding edge in the profile correlation graph. Then $G_{g,s}$ will be evaluated by the SVM and s be predicted as either a putative target site or otherwise. In the next section, we will describe the method in detail.

3.2.3 A SVM-Based Method

We developed a machine learning method for effective miRNA target prediction based on the technique of support vector machine (SVM). We used the LibSVM package [13] and chose RBF (Radial Basis Function) as the kernel function for its performance over others. Two parameters γ and C need to be decided: γ is a kernel parameter and C is a cost parameter for training. A large C results in a high performance on the training set and low on the test set. By reducing C, we can generalize the model to improve performance on unseen data. To find the best γ and C, we did a grid search [8] over these two parameters; as recommended in the LibSVM manual, increasing C and γ exponentially leads to better parameter estimation. We searched for C^x by x, starting from -5 with step +2 until 15 and for γ^y with y starting from 3 step -2 and end -15. The optimum C and γ for each class of training sets are shown in the table 3.1. With γ and C chosen, we trained a support vector regression (SVR) model and then chose a threshold T, optimum for a false positive rate < 20 and the highest possible recall rate.

3.2.4 Duplex Classification

We classified miRNA-target duplexes by their seed binding patterns. Recall that the seed of an miRNA consists of nucleotides 2-9 from the 5' end of the miRNA. A seed binding pattern between the miRNA and its target refers to the pattern of the nucleotide interactions within this region of 8 bps across the miRNA and the target. It is a canonical seed if and only if the binding contains at least 6 contiguous canonical bps.

For duplexes containing canonical seeds, they can be classified into a minimally sufficient set of six classes, which provide better recall and specificity for target recognition than the previous prediction algorithms, as the earlier works considered seeds with length just seven or eight to improve their performance [19]. Based on this classification, we obtained six classes of duplexes with canonical seeds.

For duplexes with non-canonical seeds, i.e., seeds with patterns of fewer than 6 contiguous canonical bps, we created the following classes. One class for duplexes of seeds with just one mismatch bp in the binding region. For other duplexes of seeds with more than one mismatch bp, we used the k-mean clustering algorithm [3] to classify them into another six classes. The size of each of these classes is between 200 to 500 positive samples. For each positive sample we have ten negatives for training, therefore training dataset for each class has a size between 2200 to 5500 samples. Training time was about between 5 to 17 hours on a Red Hat 4.8.2-7 server with 4 Intel Quad core X5550 Xeon Processors, 2.66GHz8M Cache and 70GB Memory. In total, we divided our dataset into 13 classes. Figure 3.6 shows the size distribution of positive duplex samples among the 13 classes.

In developing the k-mean based cluster algorithm, we represented each seed binding structure with an eight dimensional binary vector. Each dimension of such a vector is either zero or one, representing a canonical binding base pair or a mismatch in the corresponding

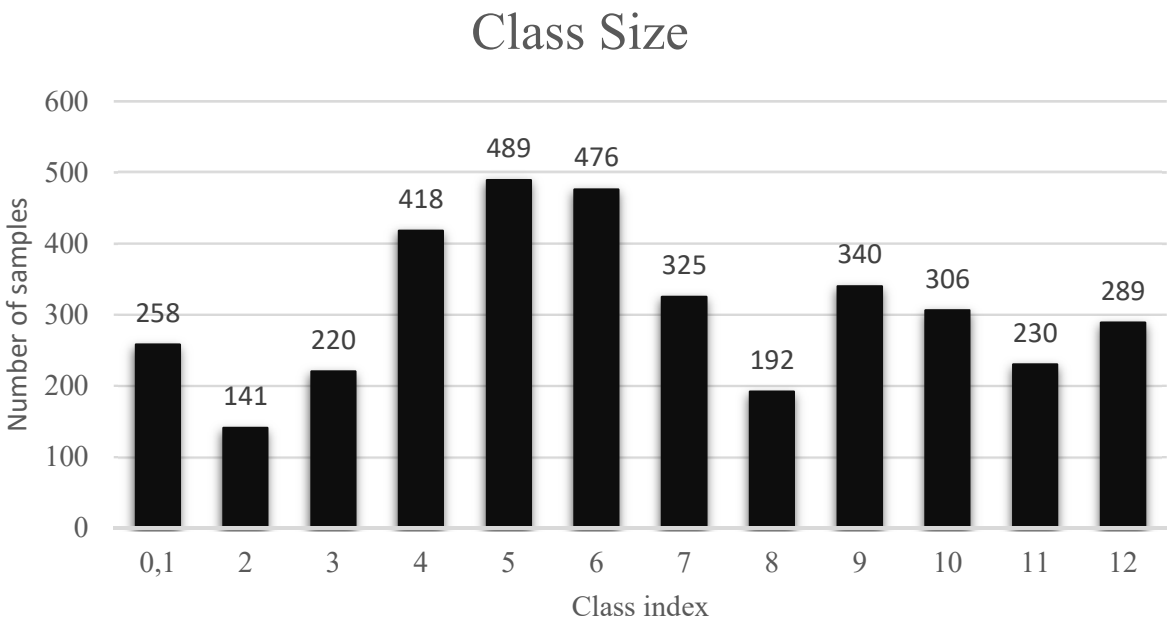

Figure 3.6: Distribution of class sizes of experimentally validated miRNA target sites in the 13 classes, where the classes indexed from 0 to 5 are of canonical seeds, class 6 is non-canonical with just one mismatch in the seed and the rest are non-canonicals with ≥ 2 mismatches. The classes are the result of a k-mean clustering algorithm, with total 3,684 of samples.

position. The distance function between two such vectors X and Y is defined as

$$Dist_{LP_z}(X,Y) = \left(\sum_{i=1}^{8}(x_i - y_i)^z\right)^{\frac{1}{z}} \tag{3.2}$$

For the k-mean clustering algorithm, the parameter z is often chosen with the value of 2 for the general distance function in LP-spaces (see formula 3.2); $z = 2$ does not necessarily work well in other dimensions than three, and surprisingly, some fractional values for z can work better for classification purposes, depending on the data distribution [2]. To identify the best parameter z for classifying our dataset, we examined values of z from 0.1 with step 0.1 up to 10. We defined the following heuristic scoring function (formula 3.3) between two

clusters C_0 and C_1 based on every examined value of z,

$$Clustering\_scr(C_0, C_1) = Dist_{LP_z}(c_0, c_1) - (r_0 + r_1) \tag{3.3}$$

where c_0 and c_1 are the centers of two clusters C_0 and C_1, respectively, and r_0 and r_1 are the radius of C_0 and C_1, respectively. Based on this objective function, the optimum value for z is identified as 0.4 for the merge of classes 10 and 12. Therefore, our classification uses the distance function $Dist_{LP_{0.4}}$.

3.3 Data and Tests

From mirTarBase, duplexes whose secondary structures have been provided in research articles were selected as positive samples for our machine learning method. In each class of miRNA-target duplexes, we partitioned positive samples into two sets; randomly we chose 80% of positive samples for training and the other 20% of positive samples for testing. We generated negative samples for training and for testing in different ways. Specifically, we generated negative samples for training according to the new technique we introduced in Section 3.2 to achieve more informative learning. Negative samples for testing were chosen randomly for performance evaluation.

To train an SVM model better, we used kernel functions to prepare data suitable to be separated by a hyperplane. For our data, the RBF kernel function has the best results in comparison to other kernel functions. For a kernel function to work the best, we identified parameters γ and C such that the pipeline of RBF and SVM has the best performance in training phase. To achieve this, we did a 10-fold cross validation on the training set.

We did a grid search for parameters γ and C in each class; Table 3.1 shows the optimum parameter values for all 13 classes of duplexes. These parameter values were used to train an SVR with all training data in each class. The reason for using SVR is to control sensitivity and specificity by a chosen threshold. We evaluated our test data on these trained SVRs.

Table 3.1: Optimal values of γ and C, and training accuracy for each of the 13 classes.

Class No.	γ	C	Training Accuracy
0	0.0000	32768	94.43
1	0.0000	32768	94.43
2	0.0000	2048	96.83
3	0.0000	8192	96.73
4	0.0000	32768	97.18
5	0.0000	2048	98.17
6	0.0000	8192	94.53
7	0.0020	8	93.09
8	0.0020	2	91.61
9	0.0020	8	91.49
10	0.0020	8	92.25
11	0.0000	2048	94.33
12	0.0020	2	92.31

Since the output of such an SVR is a single real value in the interval $[0, 1]$, a threshold value T is needed for it to decide if a sample is positive or negative based on the output value of the SVR which may be above or below the threshold T.

In our initial test, we chose the threshold T such that the false positive rate is below 20% and with the sensitivity percentage being maximized on test sets in each class. Figure 3.7 shows sensitivity, specificity and MCC for the 13 classes on chosen threshold for each class, and the averaged performance of all classes. In particular, the performance of our model for all classes, except classes 10 and 12, was very decent with the sensitivity (> 71%), the specificity (> 80%), and the MCC (> 35%).

Then we conducted the second test. For the two classes, 10 and 12, with a slightly low performance, we merged them and re-clustered them into two in LP space with parameter z=0.4, i.e. $LP_{0.4}$. We did parameter search and training in these two new classes and the performance was improved. In Figure 3.8 we see the performance of previous classification, i.e., LP_2 versus new classification with $LP_{0.4}$. There are improvements in both of classes 10 and 12. These two classes contain samples that could not be clustered correctly by

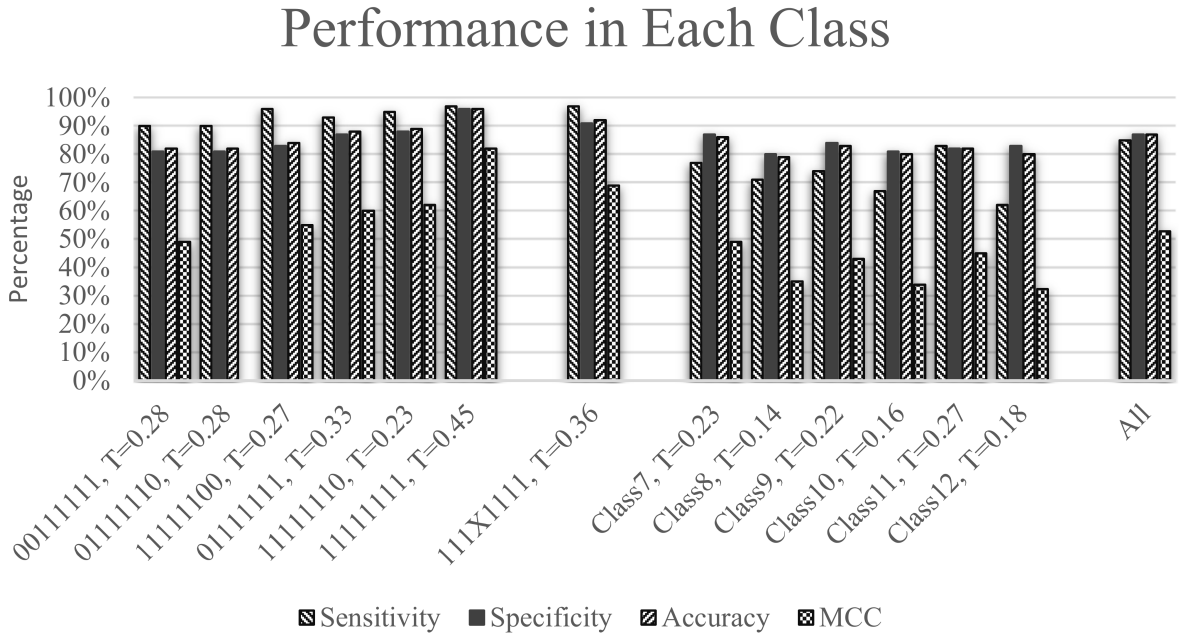

Figure 3.7: Initial test performance (sensitivity, specificity, accuracy and MCC values) for each of the 13 classes. From left to right are classes 0, 1, . . . , 12, where the patterns of seed binding are shown for the first classes 0-6, with class 6 having one mismatch (marked with X and it can be everywhere across the seed). Threshold T for every class is also shown. The rightmost is the average over all classes.

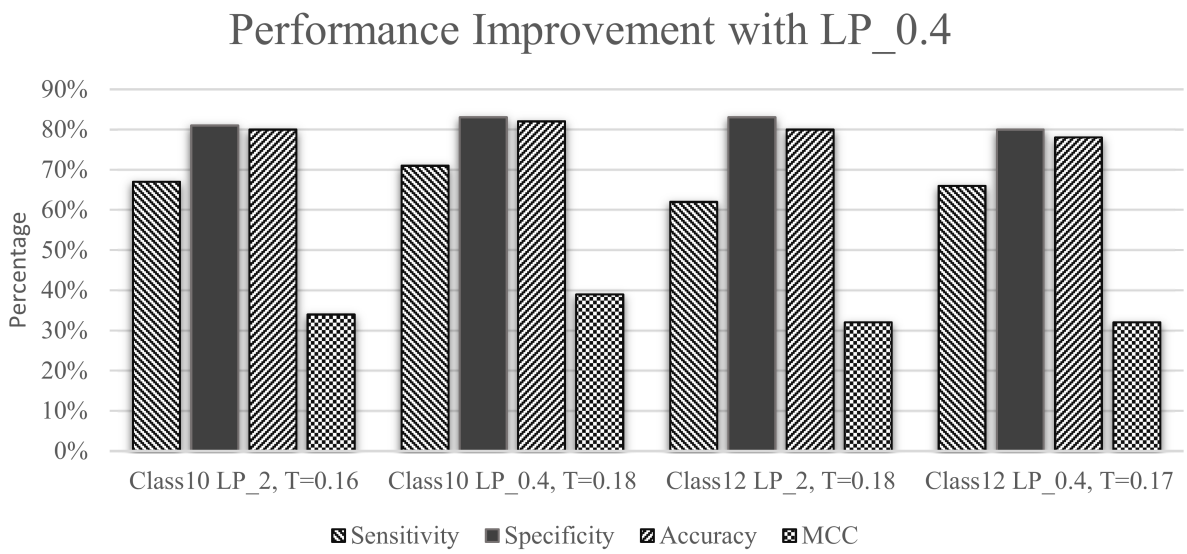

Figure 3.8: Performance improvement with $LP_{0.4}$ in the second test. Classes 10 and 12 include samples that k-mean algorithm could not classify them correctly with default distance function in LP_2 space. We reclassified these samples with distance function in $LP_{0.4}$ space. The improved results show that the new classification of classes 10 and 12 is more suitable.

LP_2. Mapping these samples to space $LP_{0.4}$ enables us to cluster them better based on our objective function formula 3.3. The improved performance confirms the correctness of our objective function.

We then replaced the previous classification of classes 10 and 12 with the new ones to compute the final overall performance shown in Figure 3.9. The improved performance on all classes has sensitivity 86%, specificity 87%, accuracy 87% and MCC 53%.

To compare our results with the state of the art methods, we ran miRanda and RNAhybrid [41] on the test set that we have drawn from [30]. The latter is a free energy based tool for target prediction in long mRNAs. These programs are renown packages that their binaries or source codes could be downloaded and run locally. Analyses of the test set by these methods shows some large overlaps between positive and negative samples for RNAhybrid MFE in Figure 3.12. The overlaps for miRanda energy in Figure 3.11 and miRanda score in Figure 3.10 are similar to those in figures 3.1 and 3.3 respectively. The magnitude of

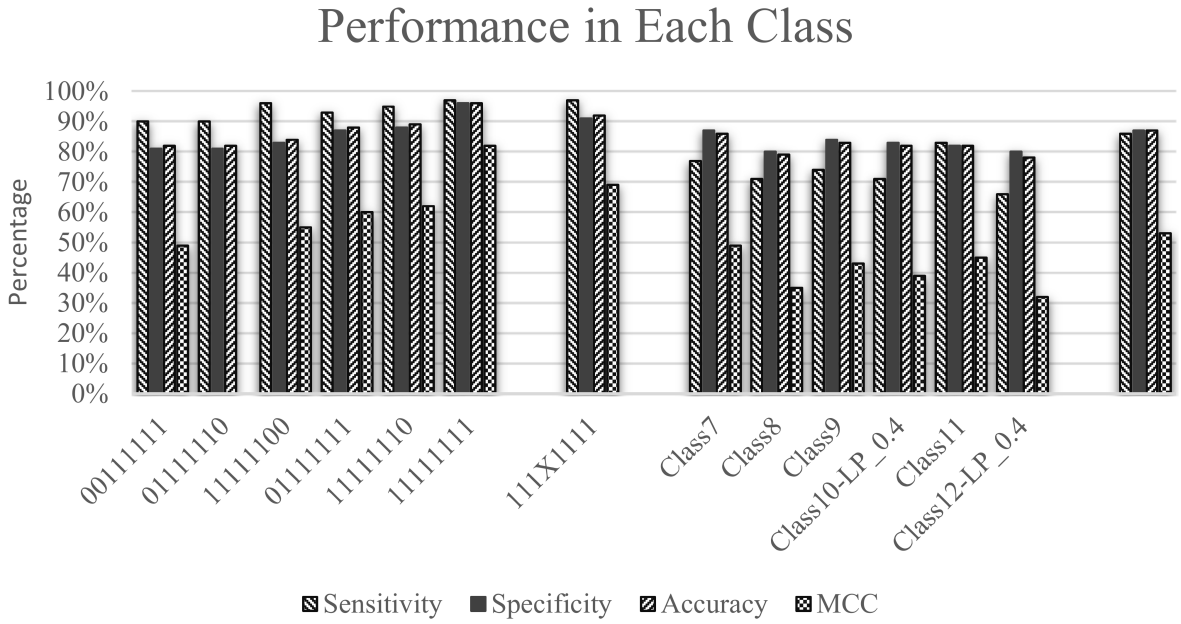

Figure 3.9: Performance improvement for each of the 13 classes, where classes 10 and 12 were obtained with distance function in $LP_{0.4}$ space. Notations are the same as in Figure 3.7.

these overlaps can result in low Area Under the Curve (AUC) in Table 3.2 for RNAhybrid MFE, miRanda MFE, and miRanda score. Our method surpasses these approaches with high margins, in terms of AUC. The reason for this significant improvement is, in addition to the novel ideas we used, that our method is data driven and the other two are rule based methods. Comparing data driven target prediction tools were not possible because either their entire training dataset and source codes were not available or they have been trained based on different sets of data and features, other than merely miRNA and target sequences.

3.4 Discussion

While most of miRNA target prediction tools have a false positive rate 30% to 40% and a recent work by [26] has a performance of 20% false positive rate and 70% sensitivity, our results on mirTarBase outperform current tools with a false positive rate of 13% and

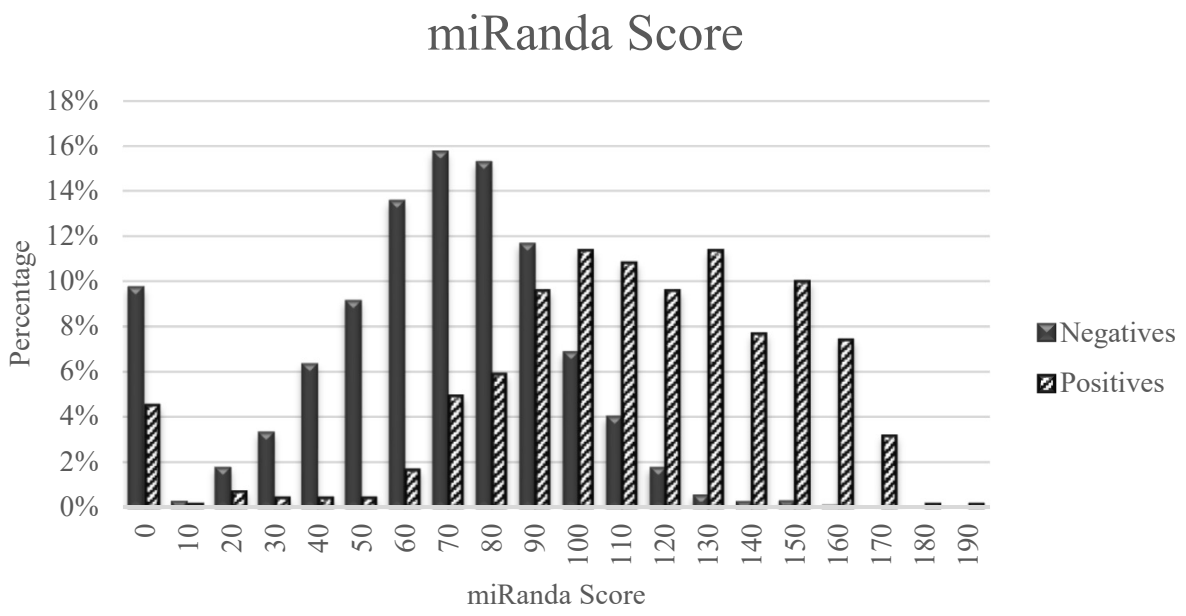

Figure 3.10: We evaluated the performance of miRanda on test sets. This graph shows the distribution of miRanda score for positive samples vs. negative samples. There is a significant overlap between positive and negative bars, showing that any miRanda score threshold chosen to separate positives from negatives, either suffers in specificity or sensitivity.

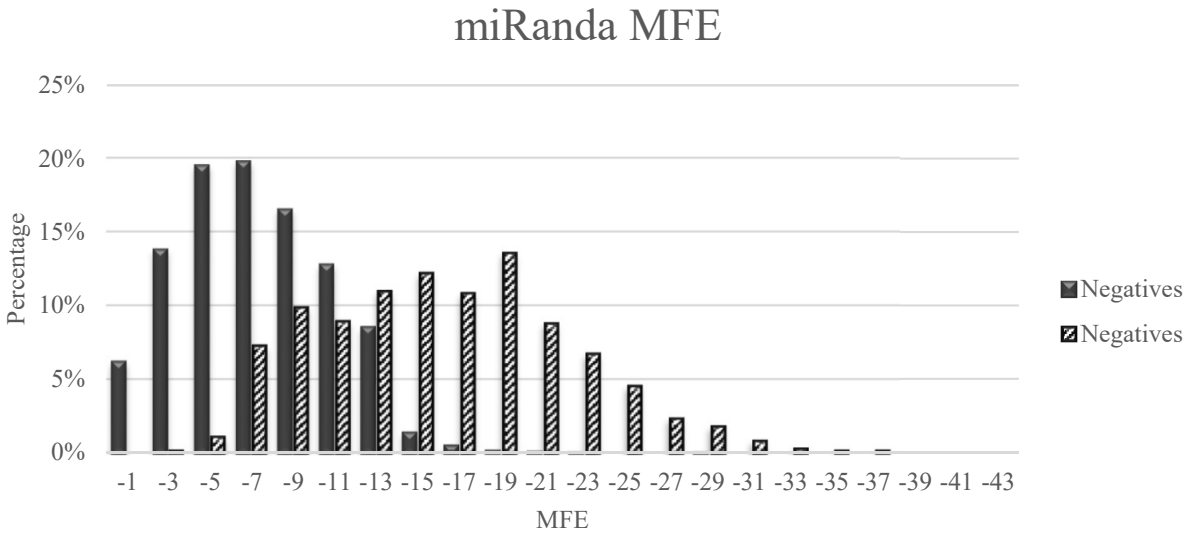

Figure 3.11: When using miRanda for miRNA target prediction the user needs to determine two thresholds; one for miRanda score and the other for its MFE. In Fig. S 2 we showed we can't have a threshold for miRanda score to maintain both high sensitivity and high specificity. This graph demonstrates that an MFE threshold can not improve sensitivity and specificity at the same time.

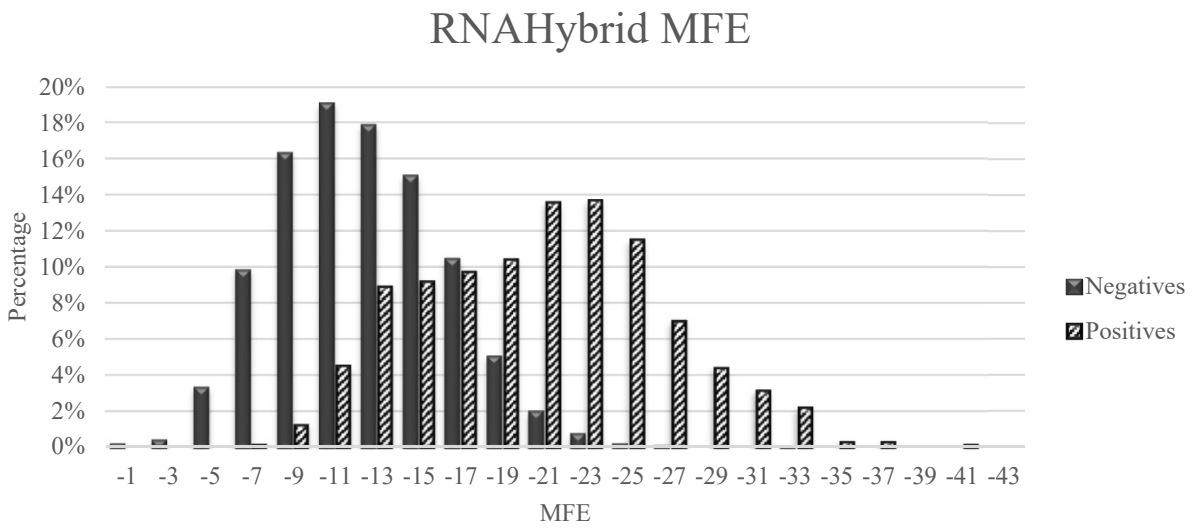

Figure 3.12: RNAHybrid MFE (Minimum Free Energy) distribution of positive samples vs. all possible negative samples in test sets. Positive samples are experimentally verified target sites while negative samples are all other locations on mRNA sequences which are not target sites. The MFE of binding duplex of miRNA sequence and a site were computed by RNAHybrid. The huge overlap of positive and negative samples demonstrates why MFE based target recognition methods have high false positive rate.

Table 3.2: Area Under the Curve (AUC) of our method vs. RNAhybrid MFE, miRanda MFE and miRanda (miR.) score per class, and on average. Total test set size is 7,370 samples.

Class No.	Our Method	RNAhyb. MFE	miRanda MFE	miR. score	Test set size
0	0.98	0.53	0.5	0.55	154
1	0.98	0.53	0.5	0.55	121
2	0.81	0.52	0.53	0.49	399
3	0.97	0.44	0.43	0.53	433
4	0.81	0.52	0.5	0.5	494
5	0.92	0.52	0.48	0.46	509
6	0.9	0.44	0.44	0.5	532
7	0.83	0.51	0.54	0.5	596
8	0.87	0.49	0.5	0.49	641
9	0.86	0.5	0.49	0.5	652
10	0.97	0.53	0.54	0.5	878
11	0.98	0.43	0.43	0.52	943
12	0.99	0.49	0.5	0.52	1018
Average AUC	0.92	0.5	0.5	0.5	

sensitivity of 86%. There are a couple of reasons for the better performance of our model; first, we did a fine grain classification, which divides the database into 13 classes. Having a good classification enables us to cluster and group samples which have the highest similarity to each other, therefore machine learning tools can learn the parameters for such a class with higher performance. Second, we developed a new way for selecting the most difficult negatives to maximize learning, while most of other data driven methods perform a random selection for choosing couple of negatives from all possibilities of negatives. Finally, our bipartite graph model, helps the SVM to learn more and converge faster by providing background information about the positive sample set in each class. Each sample has enough information to tell the SVM how far is this sample from the background distribution of positives in its class. Without our graph model and in most of other data driven works on target prediction, such information is not encoded in each sample; it would be difficult for a machine learning tool to figure out the distance of each sample from the background distribution except with hundreds of training iterations.

To predict targets for unseen given miRNA and mRNA sequences, the following steps need to be taken in the order; (1) Scan over mRNA with a window of length equal to miRNA length. (2) For each new pair of miRNA and the sequence under window, find out the class to which the pair belongs; To do so, one might use RNAcofold [27] to predict the secondary structure of duplex and use the seed binding pattern to determine its class and the corresponding model. (3) When the pair of sequences and its class are known, our program builds its bipartite weighted graph and prints out the vector of graph features. (4) We have a training model file for each class. We give the vector of features from step 3 and corresponding model file to LibSVM_predict in LibSVM package and it gives us an score. (5) By comparing the score in previous step and the class threshold in Figure 3.7 we can decide if the pair contains a target or not.

For the known miRNAs targets that we used to build our models, their secondary structures were predicted by RNAcofold. When a new and unknown sample is evaluated, we use

the same secondary structure prediction tool, therefore it will fall in the same class as if it were among the known and training samples.

3.5 Conclusion

miRNAs are small RNA sequences which regulate genes, and have important and diverse functions. As a result, miRNA studies and specifically miRNA target prediction received a lot of attention and many computational algorithms has been developed in the last two decades. One main issue with these methods has been the high false positive rate. We introduce the Correlation Graph, a bipartite graph to model any correlation between two sequences, including tertiary interactions.

To evaluate the Correlation Graph, we used mirTarBase, a widely used and experimentally verified database of miRNAs and their target genes. Our result shows a significant reduction in false positives due to the novel negative selection strategy, graph model, and machine learning approach.

Acknowledgment

Funding: This work was partly supported by NIH grant (award No: R01GM117596), as a part of Joint DMS/NIGMS Initiative to Support Research at the Interface of the Biological and Mathematical Sciences, and by NSF IIS grant (award No: 0916250). The authors would like to thank Dr. Robert W. Robinson for his feedbacks and comments on this work.

Chapter 4

A MULTI-HYPOTHESIS LEARNING ALGORITHM FOR HUMAN AND MOUSE MIRNA TARGET PREDICTION [1]

[1] M. Mohebbi, R. L. Malmberg, C. Momany, K. Rasheed and L. Cai. To be submitted to *Bioinformatics*.

Abstract

MicroRNAs (miRNAs) are small non-coding RNAs that play a key role in regulating genes and in many cell activities such as proliferation, differentiation, cell death and growth control. Dysfunction of cells in these tasks is correlated with the development of several kinds of cancer. The functionality of miRNAs depends on the location of their binding to the miRNA targets, therefore miRNA target prediction received a lot of attention for research in the last several years. Despite that, the underlying process of how they recognize their targets on mRNA genes is poorly understood. Computationally many machine learning methods have been applied on this problem and some might have high performance on recognizing the targets, yet none of them could give insight about different mechanisms for targeting. In this paper, we introduce an algorithm that simultaneously learns multiple hypotheses for miRNA binding and partitions the data accordingly. This multi-hypothesis learner not only improve the performance but also provides subsets of training data very related to each hypothesis, such that they could be used to discover new pathways in miRNA targeting. We exploited biologically meaningful features for predicting targets, to help our algorithm build hypotheses which can correlate with target recognition pathways.

Our results show that the algorithm can provide comparable performance to the state-of-the-art machine learning tools such as *RandomForest* in predicting miRNA binding sites. Moreover, feature selection on the partitions provided by our method, confirms that the partitioning mechanism is compatible with biological pathways of miRNA targeting. The resulting data partitions could be used for in vivo experiments to aid in discovery of the targeting mechanisms.

4.1 INTRODUCTION

MicroRNAs (miRNAs) are short RNA sequences of length 22 nucleotides that inhibit mRNA gene expressions. They perform as a guide to bind the RISC complex to the sequence specific locations on mRNA genes to silence them [1]. These specific locations are called target sites

and discovering the functionality of each miRNA depends on recognition of its target sites. MiRNAs can control many critical cell processes such as proliferation, differentiation, cell death, growth control and developmental timing [49]. Dysfunction of miRNAs could lead to tumor development and cancer in organs such as lung, brain, colon and breast in addition to causing hematopoietic cancers [33].

Despite the importance of miRNAs the detailed mechanism of miRNA target binding is poorly known. Lab experiments for finding targets are very slow and costly, therefore there is a huge demand for computational approaches. In the last decade dozens of algorithms, with a variety of approaches and techniques, have been developed. These methods are either specific for a few species or general for any kind. Methods for vertebrates are TargetScan and TargetScanS [48, 46], miRanda [21, 34], DIANA-microT [39] and for flies RNAhybrid [63]. Some general tools are miTarget [38] and MicroInspector [68].

The early computational approaches for target recognition were rule based, i.e., they had a set of discriminative rules derived from experimental and biological knowledge, such as MFE (Minimum Free Energy), duplex binding pattern, or target accessibility. Some popular rule based tools are RNAhybrid, TargetScan, and miRanda. In the last several years, considering the emerging relevant data sets, data driven methods have become popular. These methods use sophisticated machine learning and statistical models to learn more discriminative features for target identification [77]. Some of popular data driven tools are TargetSpy [72], miRanda-mirSVR [9] and Avishkar [26]. Almost all computational methods suffer from a high false positive rate. The innovation of more advanced sequencing techniques, and therefore more precise data sets, along with recent advances in machine learning methods, could lead to the development of more accurate algorithms.

The miRNA targeting process is not completely understood and biologists are especially interested in approaches that could give insights about the mechanisms of target recognition. Recent experimental studies of miRNA targeting reveal that there are multiple and different pathways and mechanisms for this process, while the earlier belief was merely based on seed

match of miRNA and target site sequences [14]. Currently, it is not even explicitly clear how many different and exclusive pathways guide miRNA targeting, therefore any research to compute models corresponding to biological hypothesis about targeting mechanism are highly demanded. Some machine learning techniques such as Bagging and Boosting or Random Forest aim to learn multiple hypotheses from the input data, but they do not provide any clue to check if these hypotheses are biologically meaningful or not.

In this work, we introduce a multi-hypothesis learning algorithm that not only improves the performance of a set of classifiers on miRNA targeting data by learning multiple hypotheses, but it also partitions the dataset per these learned models. These partitions could be analyzed to find miRNA target duplexes belong to the currently known targeting mechanisms. Moreover, data partitions not related to known mechanisms could be studied in vivo to discover new targeting pathways. We also expect that a meaningful partitioning of our dataset and learning a different hypothesis for each partition, could lead to a better overall performance. Our evaluations on human and mouse data shows that the multi-hypothesis learner can improve the performance of current state-of-the-art classifiers by a large margin.

4.2 DATA SETS

The success of data driven methods critically relies on the quality of the data. To build the most accurate models and the most realistic evaluations, we stablished our data based on mirTarBase [30], one of the most up-to-date data sets and the most referenced resource for miRNA target prediction research. mirTarBase contains more than 360,000 experimentally validated miRNA-target duplexes for 18 different species. We are interested in comparing this approach with both human and mouse records.

From mirTarBase, human and mouse miRNA-target duplexes were extracted whose secondary structures have been provided in research articles. Such duplexes were selected as positive samples for our machine learning method. However, negative samples are not di-

rectly available. Theoretically, any stretch of an appropriate length other than the real target in the 3'-UTR of a targeted mRNA gene can be considered a negative target of the corresponding miRNA. We randomly selected ten locations in the 3'-UTR of a targeted mRNA gene to pick up the negative samples for each positive sample with a ratio of ten to one. Each sample is a pair of miRNA sequence of length 22 and a site sequence of length 25 which is real target site for positive samples or a negative site that is not a target for the miRNA.

4.2.1 Test set and Training set

We have one training set and two tests. The human dataset is split 80% to 20% into the training set and human test set. All mouse data composes our second test set. In the human data extracted from mirTarBase, there are 322 unique miRNAs, 3651 target site sequences and 3722 pairs of miRNA and target sites. On average, each miRNA has > 10 targets sites. If we randomly select Test set samples from whole database, the odds of having many miRNAs in both test and training sets is high. To avoid such overlaps and to have the most reliable test set, we indexed pairs of miRNA and target sites by miRNA sequence. In addition, to make a test set having similar distribution to that of the whole dataset, we sort samples by miRNA sequences, put four consecutive (based on the sorting order) miRNA sequences and all their target and non-target sites in the training set, then one miRNA sequence and all its targets and non-targets into human test set and so on. In this way and in terms of miRNA sequence, not only do the human test set and training set have no overlaps but also the test set has very similar distribution to that of the whole database. Both test sets and training sets have ratio of 1:10 for positive vs. negative. The human test set is composed of 6127 samples (557 positives vs. 5570 negatives), and the total size of the mouse test set is 517.

4.3 THE MODEL AND METHOD

In this section, we explain a feature selection approach which not only is more efficient for miRNA targeting than data mining features selection methods, but also it is biologically meaningful too. Data mining algorithms could not be applied directly on this problem because each sample composed of sequences of miRNA and target, and miRNA sequence is identical among its positive(s) and its negative samples. Hence when we ran Weka [76], a data mining package, to extract features, all miRNA sequence nucleotides were excluded from selected features set. To cope with this problem, features must be defined based on correlations of miRNA and target nucleotides rather than merely on sequences of nucleotides. In addition, and to incorporate biological knowledge of miRNA targeting, we extract features from secondary structure of a duplex associated to a sample.

RNAfold [52] is a widely used secondary structure prediction tool for RNA sequences and molecules. It is a general tool and was customized in this work for miRNA and target duplexes. Based on the biology of miRNA targeting, sequences of miRNA and target sites should not make base-pairs with themselves but with the other sequence. In general, RNAfold could predict structures in which miRNA or target site sequences might bind to themselves. To avoid this problem, and include information about in vivo process of miRNA target binding, we tuned RNAfold to predict the structure of duplex based on rules we collected from biological literature explaining the actual mechanisms of miRNA targeting.

The seed of an miRNA consists of the nucleotides number 2 to 9 from the 5' end of the miRNA [48]. It is believed that the process of nucleotide binding between the miRNA and its mRNA target starts from this part [69]. When the binding in the seed region is continuous for 6 to 8 bps, it is called a canonical seed; otherwise it is called non-canonical [51]. Though the seed binding is considered the most important identifier for miRNA targets in mammals [62], a recent study shows it is not the only pathway for miRNA targeting [14]. To have a more comprehensive model, we consider correlations that occur not just in the seed region

but also in all other regions across a miRNA and its target.

4.3.1 RNAfold customization and feature selection

RNAfold for one given biomolecule sequence of nucleotides predicts the most stable secondary structure of the molecule. To use it for predicting miRNA and target site duplexes, we concatenate miRNA and target site sequences with a subsequence of length four 'X's in between. This sequence of length four is the shortest sequence that does not change the overall minimum free energy of the structure based on our experiments.

RNAfold can have a *constraints* file as an input parameter, to enforce the structure prediction process to occur based on a user's domain knowledge. Here we set these *constraints* for the miRNA targeting mechanism, to include rules for base-pairs which biologically expected to happen in seed, and rules prohibiting miRNA nucleotides from binding to miRNA itself. Similarly, there are rules avoiding target site sequence to bind over itself. We apply all these rules for duplexes with canonical seeds, while releasing seed base pairing constraints for non-canonical seeds.

Biological experiments and in vivo methods reveals several pathways for miRNA targeting [14]. The earliest discovered and the most dominant method of targeting is based on seed matching [47, 24]. In this mechanism, miRNA carried by Argonaute protein makes initial base pairs in the seed area. These bindings open the groove of Argonaut molecule to accommodate the target site [69]. To customize RNAfold for predicting duplexes in a similar fashion, we align the seed part of miRNA with nucleotides 2 to 8 from 3' side of target and pair these bases that can match to each other mutually; i.e. Adenine (A) to Uracil (U), Cytosine (C) to Guanine (G), Guanine to Uracil and vice versa.

The secondary structure predicted by RNAfold, is a list of base pairs between nucleotides in miRNA and target site. These base pairing features are nominal features, to convert them to numerical values while maintaining their independence, we code them with One-hot-encoding (OHE) approach [73]. Biologically there is no significance ordering among six

different base pairs, to keep this independence, we encode each matching base pair with one bit, totaling six bits, and one extra bit for mismatches or no base pairs. One and only one of these seven bits is *hot* or one at a time. In addition, we have two integer values indicating size of bulges on miRNA and on target site, adjacent to each nucleotide. These values are zero if there is no bulge in the structure pertaining to the current nucleotide. In total there are 9 features per each miRNA nucleotide.

Experimental studies on human miRNA targets show Adenine is a very frequent base at the far most 3' end of a target site, ie at t1 [69, 47]. To consider this domain knowledge preference into feature sets, we add four bits corresponding to A, C, G and U at t1. Another study on the structural basis of miRNA targeting, revealed that the nucleotide in t1 goes into a pocket inside AGO protein structure and does not pair to the corresponding nucleotide on miRNA, ie g1. Therefore, to reduce the size of feature set, we exclude g1 from being encoded. A factor indicating stability of a structural binding is MFE, we include it as the last feature. We fixed the length of miRNAs to 22 nucleotides, but g1 is not considered, therefore the total number of features for each sample is 194 or (1+4+21*9). This procedure has been illustrated inside dashed area of Fig. 4.1. Our MHL (Multi-Hypotheses Learner) algorithm which is explained in next section, treats each sample as a vector of these 194 features and learns several hypotheses pertaining to different miRNA targeting mechanisms. Figure 4.1 shows all sections of our bundle algorithm including the feature selection part and the MHL algorithm.

4.3.2 The Algorithm

The idea of the algorithm is to divide the dataset into two disjoint subsets sb_1 and sb_2 such that these two subsets have similar distributions of labels or classes. Learn the major pattern in sb_1 with classifier c_1 and store it as model m_1. Partition sb_2 based on m_1's performance in to two parts *can-decide* or *cannot-decide* samples. The partition *can-decide* contains instances where m_1 can predict their labels with confidence while the other partition

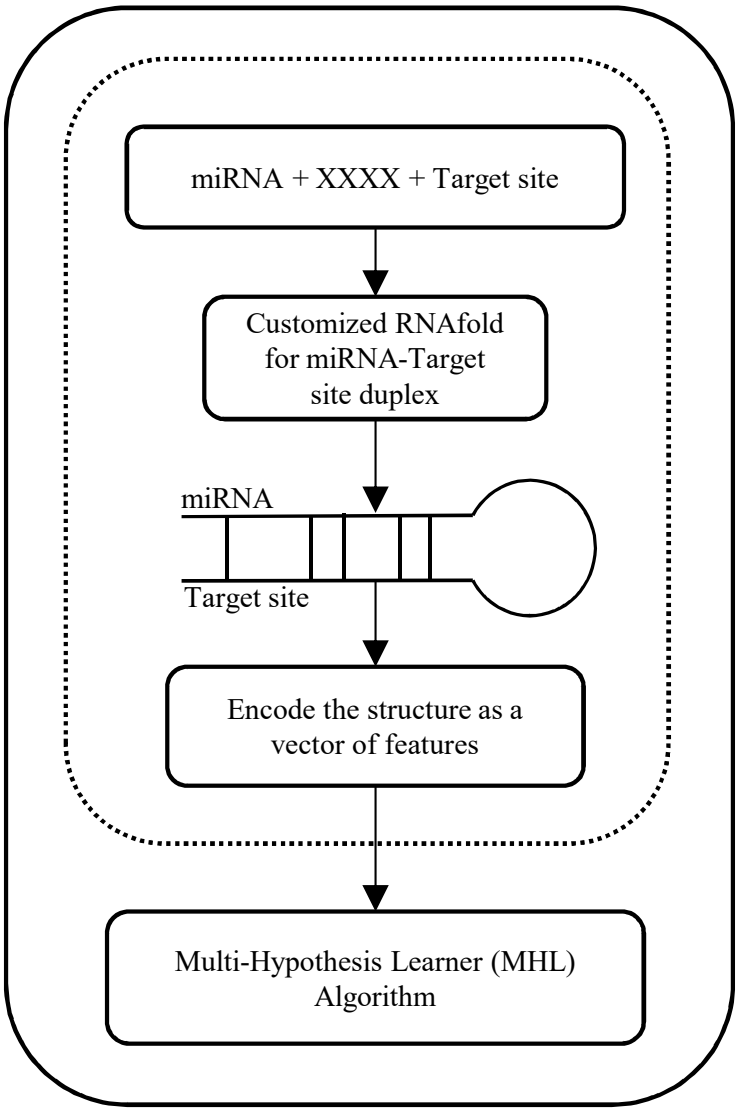

Figure 4.1: Our bundle algorithm; It gets two sequences of miRNA and target, predicts the secondary structure of their duplex by our customized version of RNAfold. The structure is encoded as a vector of features and passed down to MHL.

includes those that m_1 is not sure about their classification. Subsets *cannot-decide* and sb_1 are merged to yield a new training set. The process is repeated recursively on this new set until no further partitioning into *can-decide* and *cannot-decide* is possible.

The algorithm has two main parts; Trainer and Tester. Trainer gets the training set T_0, a classifier set C and a desired sensitivity and specificity, *_sen* and *_spec*. During a recursive procedure, Trainer builds regression models, i.e. hypotheses, specific for different patterns of data, which are observed in the input training set T_0. It also stores each produced model M along with two thresholds T_{up} and T_{down} for Tester part. The model M guarantees the desired sensitivity and specificity *_sen* and *_spec* for *can-decide* partition. For every sample evaluated by M with a value >= T_{up}, it would be classified as positive while labeled as negative if it's evaluation value is < T_{down}. Where evaluation value is between T_{down} and T_{up}, then the model does not classify the sample and it would be added to *cannot decide* set.

4.3.3 Trainer

Trainer is composed of three functions; $Splitter()$,
$Model\_Builder()$ and $Threshold\_Finder()$.

The function $Splitter(D, C)$ gets a dataset D and a set of classifiers C as input. It splits the input set D into two subsets A and B by the Stratification method [74] to maintain the same ratio of positive samples versus negatives in these subsets as it is in D. A and B are disjointing and complement of each other corresponding to D, i.e. $A \cup B = D$. By calling the function $Model\_Builder(c_i, A, B)$, the model m_i is built by classifier c_i on dataset A. In addition, the function splits B into *can-decide* and *cannot-decide* sets, merges A with *cannot-decide* and returns as D_{new1}. This is the new training set and the process recursively is repeated on this new set. To avoid any bias toward the way we split the data by the Stratification method, we swap the position of A and B then repeat the process.

Depending on how high the thresholds *_sen* and *_spec* are chosen, the $Model\_Builder()$ function may not be able to build such a model and might not return a new training set.

In such a case, it returns the same set as the input training set, indicating it failed to build the desired model. Given this condition, the *Splitter*() builds a model with c_i on the input training set and stops.

Algorithm 1: Splitter (D, C)

```
1  foreach classifier c_i ∈ C do
2  |   Split D into two subsets A and B by the Stratification method;
3  |   D_new1 = Model_Builder(c_i, A, B);
4  |   if |D_new1| < |D| then
5  |   |   Splitter (D_new1, C);
6  |   end
7  |   D_new2 = Model_Builder(c_i, B, A);
8  |   if |D_new2| < |D| then
9  |   |   Splitter (D_new2, C);
10 |   end
11 |   if |D_new1| == |D| OR |D_new2| == |D| then
12 |   |   train c_i with D, store and stop;
13 |   end
14 end
```

There are two thresholds associated with each trained model; T_{up} and T_{down}. We compute these thresholds such that the model m_i has a given and desired sensitivity and specificity *_sen* and *_spec*. The higher sensitivity and specificity results in larger *cannot-decide* subset in B. Let's call this subset as β.

Algorithm 2: Model_Builder(c_i, s_a, s_b)

```
1 Train c_i with set s_a;
2 Evaluate set s_b by model m_{i_a};
3 Store the evaluations as a list L_b of Pair(sample.label, sample.evaluation);
4 Pair (T_down, T_up) = Threshold_Finder(L_b, _sen, _spec) ;   /* find T's satisfying
    sensitivity and specificity */
5 β = subset of s_b that evaluated as >= T_down and < T_up;
6 Store the model m_{i_a} with (T_down, T_up) ;   /* s_b − β is can_decide subset */
7 Store (s_b − β) as an ARFF file ;   /* for further analysis */
8 Return s_a ∪ β.
```

Algorithm 3: Threshold_Finder(L_b, $\_sen$, $\_spec$)

```
1  T_up =1, T_down =0;
2  Votes [] = ∅;
3  do
4      foreach p_i ∈ L_b do
5          if p_i.evaluation >= T_up then
6              Votes[p_i] = positive;
7          end
8          if p_i.evaluation < T_down then
9              Votes[p_i] = negative;
10         end
11     end
12     Compute _sen_tmp and _spec_tmp for Votes[];
13     if _sen_tmp < _sen OR _spec_tmp < _spec then
14         stop and break;
15     end
16     T_down+ = ;                                   /*  = 0.05  */
17     T_up− = ;
18 while T_down < T_up;
19 Return Pair(T_down, T_up);
```

4.3.4 Tester

The Tester procedure loads all model files from the training step into the memory and when a new and unlabeled sample is given for evaluation, all models examine the sample. If a model evaluates the sample with a value between T_{down} and T_{up} then it does not vote, otherwise it votes with confidence as positive if the value is >= T_{up} and as negative for the value < T_{down}. Each vote associated with a weight, which is the size of the dataset used to build the model. The weighted average of all votes is returned as the final prediction.

4.4 DISCUSSION

In the training set, there might be several patterns of targeting, denoted by *circles*, *squares*, *triangles* and etc as shown in Fig. 4.2. *Circles* form the dominant pattern. Our MHL algorithm divides it to subsets A and B. Classifier c learns *Circles* pattern when it runs over subset A and create a model for circles, i.e. m_c. Partitioning B with m_c, removes *circles* from B and add the rest of B to A to form a new training set. Now, in the new training set *squares* are the dominant pattern and in next recursion step, a model is built for them. This recursion will continue until learning all patterns or there are no dominant pattern left. In later case, a model for the remaining samples is created by c and recursion stops.

4.4.1 Test of our Multi-Hypothesis Learning (MHL) bundle algorithm

To test the effectiveness of our algorithm, we compare the Area Under the Curve (AUC) of different Machine Learning (ML) models from Weka package [76] versus our Multi-Hypothesis Learning (MHL) bundle algorithm. Table 4.1 and Table 4.2 present these results on human (HSA) test set and mouse (MMU) test set respectively. Columns of these tables are classifier(s) name, desired sensitivity and specificity for MHL, AUC of ML models and AUC of our algorithm when the same ML classifier used as underlying model in the MHL.

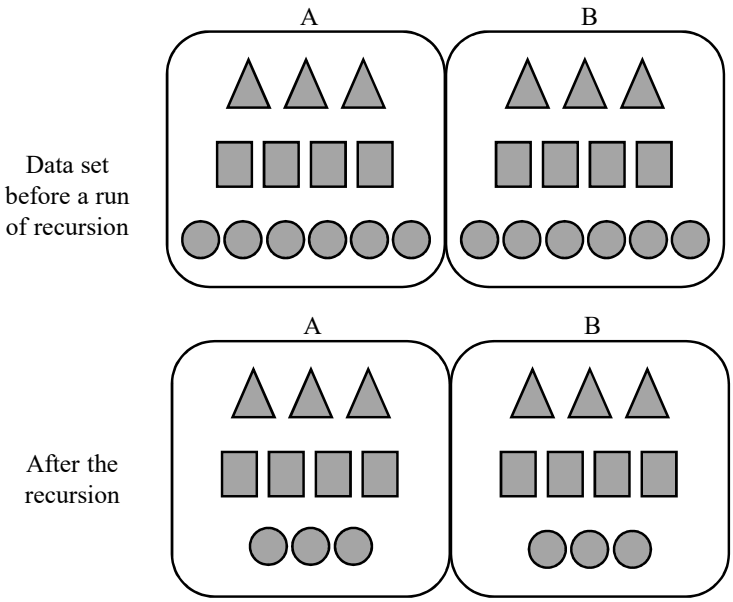

Figure 4.2: The MHL recursion algorithm to learn several hypotheses from a training dataset containing different patterns of data.

These tables show that our algorithm is effective, especially when the ML model does not perform well on a test set. It has the best performance with Linear Regression; For human test set, the AUC of this model for sequences of miRNA and target sites is 0.69 and when the same samples are given to our algorithm, it increases the AUC to 0.93. The algorithm improves the AUC by 0.24 which is the highest out performance over the ML model itself. It seems the highest achievable AUC for our training and human test sets is 0.93 and Random Forest is the only classifier that our algorithm does not have improvement over it as it already performed very well with AUC of 0.92. The effectiveness of MHL algorithm is more obvious when it enables mediocre-performed classifiers such as REPTree, Linear Regression and M5P to beat Random Forest. Classifiers performing weak on this test set such as Random Tree, Decision Stump and Artificial Neural Networks (ANN) can also be used in the MHL module to deliver a performance of 0.75 to 0.83 in AUC. The AUC of ML models has an average of 0.73 with a standard deviation 0.13 while our algorithm can perform with AUC average of 0.87 and standard deviation 0.08. Our algorithm improves the average performance by 14%.

Table 4.1: Test set: HSA (Human), $|HSA| = 6129$ samples

Classifier(s)	Sen./ Spec.	ML models	our MHL bundle
RandomTree	85/85	0.59	0.78
DecisionStump	85/85	0.58	0.75
REPTree	85/85	0.82	0.93
RandomForest	85/85	0.92	0.92
ANN	90/90	0.66	0.83
LinearRegression	85/85	0.69	0.93
M5P	90/90	0.85	0.93
Rand.Tree&DecisionStump	85/85	NA	0.84
REPTree&RandomForest	80/80	NA	0.93

Table 4.2: Test set: MMU (Mouse), $|MMU| = 517$ samples

Classifier(s)	Sen./ Spec.	ML models	our MHL bundle
RandomTree	85/85	0.74	0.95
DecisionStump	85/85	0.48	0.61
REPTree	85/85	0.93	1
RandomForest	85/85	0.98	0.99
ANN	90/90	0.75	0.97
LinearRegression	85/85	0.52	0.99
M5P	90/90	0.95	0.97
Rand.Tree&DecisionStump	85/85	NA	0.94
REPTree&RandomForest	80/80	NA	1

To compute the averages, we did not consider combined classifiers, i.e., the last two rows in both tables left blank.

We conjectured that there might be some subsets of data learn-able with a classifier better than the other, therefore we recursively searched all possible combinations of partitioning with different classifiers. To test this idea, we used combinations of RandomTree and Decision Stump in MHL and surprisingly, the performance increased by 0.06 in AUC versus using these classifiers individually in MHL. This combination outperformed these classifiers by 0.25 if they were used individually and without MHL.

Human and mouse branched from a common ancestor about 80 million years ago. They

have comparably similar genomes and virtually the same set of genes [25]. Therefore, it is useful to train a model by human genomic data and test it on mouse data sets. Similarly we ran the same model used for testing human data, for mouse dataset and the results are shown in table 2. Our algorithm improves over all ML classifiers and the maximum improvement is for Linear Regression again with increase of 0.47 in AUC. The average performance of our algorithm is 0.93 with standard deviation 0.14 while ML models has an average of 0.76 and standard deviation 0.14. The algorithm average performance surpasses over ML methods average by 0.17.

Contrasting table 4.1 and table 4.2 shows that the ML models and our algorithm performs slightly better on mouse than human data. The similar performance of these models on both mouse and human test sets shows miRNA target duplexes and targeting mechanism features are evolutionary conserved across both species. Some miRNAs are sequentially conserved among human and mouse, consequently there might be an small portion of similar samples in both the (human) training set and the mouse test set. This could be the reason for a tiny better performance on the mouse test set over the human test set. In the human test set, we reduced the chance of sequentially similar miRNAs in both training and test sets by sorting miRNAs and dividing them between training and test sets alternatively with the given ratio. In the mouse test set, samples belong to a different species and any correlation with human training set is due to biological relevance. Therefore, a small portion of miRNAs which are common between mouse and human, may lead to sample similarity between human training set and mouse test set.

While some machine learning packages such as $RandomForest$ can have comparable AUC performance, the main advantage of MHL over all ML methods is to provide clue into the dataset and partition the data into sets that are biologically meaningful clusters. For this study MHL printed out five subsets of training data and we ran CFS (Correlation based Feature Selection) subset feature selection [28] from Weka package over these subsets and over all the training set, then the results are shown in table 4.3. Row one shows features

from all training data and the rest of columns belong to the five resulting subsets.

Table 4.3: CFS feature selection on training data versus on subsets provided by MHL algorithm. It shows that MHL can help to extract biological details from subsets while they couldn't be seen by running CFS on complete training set. In each feature the number represents the miRNA nucleotide index.

Dataset	Selected Features
All Training Data	2_AU, 2_UA, 2_MisMatch, 3_UA, 4_AU, 4_UA, 4_LoopLen_miRNA, 5_AU, 5_UA, 6_AU, 6_LoopLen_miRNA, 7_AU, 8_AU, 8_UA, 10_LoopLen_target, 15_AU, 20_AU, 22_GU, MFE
Subset 1	2_AU, 2_UA, 2_GC, 3_UA, 4_AU, 4_UA, 4_LoopLen_miRNA, 5_AU, 5_UA, 5_LoopLen_miRNA, 6_AU, 6_LoopLen_miRNA, 7_AU, 7_LoopLen_miRNA, 8_AU, 8_UA, 13_AU, 13_GU, 15_AU, 19_UA, 20_LoopLen_target, 21_GU, 22_AU, MFE
Subset 2	t1_A, 2_AU, 2_UA, 2_GC, 2_MisMatch, 2_LoopLen_miRNA, 3_UA, 3_LoopLen_miRNA, 4_AU, 4_UA, 5_UA, 6_AU, 6_GU, 7_AU, 8_AU, 8_UA, 9_GU, 10_GU, 16_AU, 17_UA, 21_AU, 21_UA, 22_AU, MFE
Subset 3	2_AU, 2_UA, 3_UA, 4_AU, 4_LoopLen_miRNA, 5_AU, 6_AU, 6_UA, 7_AU, 8_AU, 8_UA, 9_LoopLen_target, 10_LoopLen_target, 15_AU, 21_UA, 22_AU, 22_GU, MFE
Subset 4	2_AU, 2_MisMatch, 4_AU, 7_GU, 8_LoopLen_target, 19_UG, 20_AU, 20_LoopLen_target, 22_AU, MFE
Subset 5	2_AU, 2_UA, 2_GC, 3_UA, 4_AU, 4_UA, 4_LoopLen_miRNA, 5_AU, 5_UA, 6_AU, 6_LoopLen_miRNA, 7_AU, 8_AU, 8_UA, 10_LoopLen_target, 16_AU, 20_AU, 20_LoopLen_target, 21_UA, MFE

By contrasting rows two to five versus row one, we can see that MHL algorithm could partition the data in a more biologically meaningful way. The feature 2_*mismatch* separates canonical seed samples from non-canonicals and MHL partitioned these two type of samples into subsets 1, 3 and 5 for canonicals, versus subsets 2 and 4 for non-canonicals. The appearance of Adenine in the first nucleotide position of target, i.e t1_A, is a major identifier of targets for many cases and all such samples put in subset 2. GC is a strong base-pair and one biological mechanism of targeting [69] claims base-pairs on positions 2, 3 and 4 make the groove inside AGO protein to open and to accommodate target. MHL has been able to cluster samples related to this mechanism into subsets 1, 2 and 5. Thermodynamically, continuing base-pairs on 3' end of miRNA, provide more stable duplexes [12], this has been

captured in subsets 1, 2, 3 and 4. These biologically interpretable details seen in table 4.3 within subsets 1 to 5 could not be extracted by the same feature selection algorithm on the complete dataset, i.e. the first row of the table. This shows that the MHL algorithm could provide subsets of the data which seems to have biologically correlated samples; because only in these subsets, the feature selection algorithm could discover the biologically validated target identifiers while not in the total training set. The subsets could be further studied to figure out targeting mechanism of each sample and therefore each miRNA. Some subsets or features may not have a known biologically interpretation based on the current understanding of miRNA targeting mechanisms, but they could be used for in vivo experiments to discover and verify new targeting mechanisms.

4.5 CONCLUSION

miRNAs are small endogenous non-coding RNA molecules that have a critical function in suppressing genes and they also correlate with many diseases and cancers. Due to the importance of their effects in several cell pathways, biologists are very interested to discover their functionality. Their function may be correlated with the way they recognize their targets. A lot of research has been going to develop algorithms for miRNAs' target prediction, and in this work, we present a multi hypotheses learner algorithm, MHL that could provide more accurate results by biologically meaningful partitioning of the miRNA target duplexes. In addition, these partitions could be used for better understanding of targeting mechanisms as well as providing sequences for in vivo experiments, to discover new pathways.

Our evaluations and results shows that the partitioning approach can significantly improve the performance of a classifier. Moreover, feature selection in the resulting partitions suggests that the partitioning mechanism is compatible with biological pathways of miRNA targeting.

Acknowledgement

This work was supported in part by NIH grant (award No: R01GM117596), as a part of Joint DMS/NIGMS Initiative to Support Research at the Interface of the Biological and Mathematical Sciences, and NSF IIS grant (award No: 0916250).

Chapter 5

Conclusion

miRNAs play a key role in regulating genes and in many cell activities such as proliferation, differentiation, cell death, and growth control. Dysfunction of cells in these tasks correlates with several cancers and tumors development. Functionality of miRNAs varies depending on the location of their bindings, therefore miRNA target prediction received a lot of attention in the last several years of research. Despite that, the underlying process of how they recognize their targets on mRNA genes is poorly understood. Accurate prediction of miRNA targets can assist efficient experimental investigations on the functional roles of miRNAs. The prediction, however, remains a challenge task due to the lack of knowledge about the actual mechanism of miRNA targeting by having not adequate experimental data about the tertiary structure of miRNA-target binding duplexes.

In this dissertation, we targeted two issues related to the most of computational methods on miRNA target recognition; high false positive rates and lack of insight regarding the actual mechanism of miRNA targeting. We proposed a graph model, called Correlation Graph, to capture co-appearance of different nucleotides across two sequences of miRNA and target. This graph enable us to get information beyond secondary structure bindings, which includes tertiary interactions. Upon a Correlation Graph built for every duplex of miRNA and target, an SVM machine learning model, learns graphs that contain a real target for a given miRNA sequence. mRNA sequence can be very large while having just a few targets per a miRNA, therefore there are many non-target locations versus each target. This makes our binary dataset to be very imbalanced. Imbalance datasets could significantly deteriorate

the performance of machine learning methods therefore we designed a novel method to re-sample our dataset to reduce imbalance ratio while improving the learning process. We evaluated our model versus other miRNA target prediction methods on human miRNAs and target data obtained from mirTarBase, a database of experimentally verified miRNA-target interactions. Our method achieved a sensitivity of 86% with a false positive rate below 13% in predicting targets which is a significant improvement in contrast to other published works with a 30% to 40% false positive rate. In addition, we compared our model with the state-of-the-art methods miRanda and RNAhybrid on the test data, our method outperformed both of them by a significant margin.

The second issue with the most of computational methods for miRNA target recognition is the lack of insight into the actual mechanisms of targeting. To address this problem, we introduce an algorithm that simultaneously learns multiple hypotheses and partitions the data accordingly. This multi-hypothesis learner not only improves the performance but also provides subsets of training data very pertinent to each hypothesis, such that they could be used to discover new pathways in miRNA targeting. We exploited biologically meaningful features for targeting mechanisms, to help our algorithm build hypotheses which could correlate with target recognition pathways. Our results show that the algorithm can provide comparable performance to the state-of-the-art machine learning tools such as RandomForest. In addition, our evaluations show that the partitioning approach can significantly improve the performance of a classifier.

Performing feature selection on the partitions provided by our method confirms that the partitioning mechanism is compatible with biological pathways of miRNA targeting. These partitions could be used for better understanding of the currently known mechanisms or to discover new pathways.

Bibliography

[1] Vikram Agarwal, George W Bell, Jin-Wu Nam, and David P Bartel. Predicting effective microrna target sites in mammalian mrnas. Elife, 4:e05005, 2015.

[2] Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. On The Surprising Behavior Of Distance Metrics In High Dimensional Space. Springer, 2001.

[3] Charu C Aggarwal and Chandan K Reddy. Data Clustering: Algorithms And Applications. CRC Press, 2013.

[4] Ines Alvarez-Garcia and Eric A Miska. Microrna functions in animal development and human disease. Development, 132(21):4653–4662, 2005.

[5] Victor Ambros. The functions of animal microrna. Nature, 431(7006):350–355, 2004.

[6] Stefan L Ameres and Phillip D Zamore. Diversifying microrna sequence and function. Nature Reviews Molecular Cell Biology, 14(8):475–488, 2013.

[7] David P Bartel. Micrornas: genomics, biogenesis, mechanism, and function. Cell, 116(2):281–297, 2004.

[8] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(Feb):281–305, 2012.

[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] Doron Betel, Manda Wilson, Aaron Gabow, Debora S Marks, and Chris Sander. The microrna. org resource: targets and expression. Nucleic Acids Research, 36(suppl 1):D149–D153, 2008.

[11] Béla Bollobás. Modern Graph Theory, volume 184. Springer Science & Business Media, 2013.

[12] Julius Brennecke, Alexander Stark, Robert B Russell, and Stephen M Cohen. Principles of microrna–target recognition. PLoS Biol, 3(3):e85, 2005.

[13] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2(3):27, 2011.

[14] Nicole Cloonan. Re-thinking mirna-mrna interactions: Intertwining issues confound target discovery. Bioessays, 37(4):379–388, 2015.

[15] Reinhard Diestel. Graph Theory {Graduate Texts In Mathematics; 173}. Springer-Verlag Berlin and Heidelberg GmbH & amp, 2000.

[16] C. Dieterich and P. F. Stadler. Computational biology of rna interactions wiley interdiscip. Rev. RNA, 4:107–120, 2013.

[17] John G Doench and Phillip A Sharp. Specificity of microrna target selection in translational repression. Genes & Development, 18(5):504–511, 2004.

[18] Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison. Biological Sequence Analysis: Probabilistic Models Of Proteins and Nucleic Acids. Cambridge university press, 1998.

[19] Daniel C Ellwanger, Florian A Büttner, Hans-Werner Mewes, and Volker Stümpflen. The sufficient minimal set of mirna seed types. Bioinformatics, 27(10):1346–1350, 2011.

[20] Anton J Enright, Bino John, Ulrike Gaul, Thomas Tuschl, Chris Sander, Debora S Marks, et al. Microrna targets in drosophila. Genome Biology, 5(1):R1–R1, 2004.

[21] Anton J Enright, Bino John, Ulrike Gaul, Thomas Tuschl, Chris Sander, Debora S Marks, et al. Microrna targets in drosophila. Genome Biology, 5(1):R1–R1, 2004.

[22] Marc R Fabian and Nahum Sonenberg. The mechanics of mirna-mediated gene silencing: a look under the hood of mirisc. Nature Structural and Molecular Biology, 19(6):586–593, 2012.

[23] Minimum free energy structure. Minimum free energy structure. http://eternawiki.org/wiki/index.php5/Minimum_Free_Energy_Structure. Accessed: 06-01-2017.

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

[25] Genome.gov. https://www.genome.gov/10001345/. Accessed: 06-01-2017.

[26] Asish Ghoshal, Ananth Grama, Saurabh Bagchi, and Somali Chaterji. An ensemble svm model for the accurate prediction of non-canonical microrna targets. In Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics, pages 403–412. ACM, 2015.

[27] Andreas R Gruber, Ronny Lorenz, Stephan H Bernhart, Richard Neuböck, and Ivo L Hofacker. The vienna rna websuite. Nucleic Acids Research, 36(suppl 2):W70–W74, 2008.

[28] Mark A Hall. Correlation-based feature selection for machine learning. PhD thesis, The University of Waikato, 1999.

[29] Sheng-Da Hsu, Feng-Mao Lin, Wei-Yun Wu, Chao Liang, Wei-Chih Huang, Wen-Ling Chan, Wen-Ting Tsai, Goun-Zhou Chen, Chia-Jung Lee, Chih-Min Chiu, et al. mirtarbase: a database curates experimentally validated microrna–target interactions. Nucleic Acids Research, page gkq1107, 2010.

[30] Sheng-Da Hsu, Yu-Ting Tseng, Sirjana Shrestha, Yu-Ling Lin, Anas Khaleel, Chih-Hung Chou, Chao-Fang Chu, Hsi-Yuan Huang, Ching-Min Lin, Shu-Yi Ho, et al. mirtarbase update 2014: an information resource for experimentally validated mirna-target interactions. Nucleic Acids Research, 42(D1):D78–D85, 2014.

[31] Yong Huang, Xing Jia Shen, Quan Zou, Sheng Peng Wang, Shun Ming Tang, and Guo Zheng Zhang. Biological functions of microrna: a review. Journal of Physiology and Biochemistry, 67(1):129–139, 2011.

[32] Richard J Jackson and Nancy Standart. How do microrna regulate gene expression. Sci Stke, 367(re1), 2007.

[33] Martin D Jansson and Anders H Lund. Microrna and cancer. Molecular Oncology, 6(6):590–610, 2012.

[34] Bino John, Anton J Enright, Alexei Aravin, Thomas Tuschl, Chris Sander, and Debora S Marks. Human microrna targets. PLoS Biol, 2(11):e363, 2004.

[35] Neil C Jones and Pavel Pevzner. An Introduction To Bioinformatics Algorithms. MIT press, 2004.

[36] Radial Basis Function Kernel. `http://pages.cs.wisc.edu/~matthewb/pages/notes/pdf/svms/RBFKernel.pdf`. Accessed: 06-01-2017.

[37] 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.

[38] 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(1):411, 2006.

[39] 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 & Development, 18(10):1165–1178, 2004.

[40] Jacek Krol, Inga Loedige, and Witold Filipowicz. The widespread regulation of microrna biogenesis, function and decay. Nature Reviews Genetics, 11(9):597–610, 2010.

[41] Jan Krüger and Marc Rehmsmeier. Rnahybrid: microrna target prediction easy, fast and flexible. Nucleic Acids Research, 34(suppl 2):W451–W454, 2006.

[42] Wojtek J Krzanowski and David J Hand. ROC Curves For Continuous Data. CRC Press, 2009.

[43] Machine Learning. What is machine learning? `https://www.coursera.org/learn/machine-learning/lecture/Ujm7v/what-is-machine-learning`. Accessed: 06-01-2017.

[44] Supervised learning. Supervised learning. `https://en.wikipedia.org/wiki/Supervised_learning#Approaches_and_algorithms`. Accessed: 06-01-2017.

[45] Unsupervised Learning. Unsupervised learning, machine learning technique for finding hidden patterns or intrinsic structures in data. `https://www.mathworks.com/discovery/unsupervised-learning.html`. Accessed: 06-01-2017.

[46] 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.

[47] 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.

[48] Benjamin P Lewis, I-hung Shih, Matthew W Jones-Rhoades, David P Bartel, and Christopher B Burge. Prediction of mammalian microrna targets. Cell, 115(7):787–798, 2003.

[49] Shuibin Lin and Richard I Gregory. Microrna biogenesis pathways in cancer. Nature Reviews Cancer, 15(6):321–333, 2015.

[50] Bin Liu, Longyun Fang, Fule Liu, Xiaolong Wang, Junjie Chen, and Kuo-Chen Chou. Identification of real microrna precursors with a pseudo structure status composition approach. PloS One, 10(3):e0121501, 2015.

[51] Gabriel B Loeb, Aly A Khan, David Canner, Joseph B Hiatt, Jay Shendure, Robert B Darnell, Christina S Leslie, and Alexander Y Rudensky. Transcriptome-wide mir-155 binding map reveals widespread noncanonical microrna targeting. Molecular Cell, 48(5):760–770, 2012.

[52] Ronny Lorenz, Stephan H Bernhart, Christian Hoener 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.

[53] T M Witkos, E Koscianska, and W J Krzyzosiak. Practical aspects of microrna target prediction. Current Molecular Medicine, 11(2):93–109, 2011.

[54] Manolis Maragkakis, Martin Reczko, Victor A Simossis, Panagiotis Alexiou, Giorgos L Papadopoulos, Theodore Dalamagas, Giorgos Giannopoulos, G Goumas, Evangelos Koukis, Kornilios Kourtis, et al. Diana-microt web server: elucidating microrna functions through target prediction. Nucleic Acids Research, page gkp292, 2009.

[55] Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure, 405(2):442–451, 1975.

[56] messenger RNA. messenger rna definition. http://www.dictionary.com/browse/messenger-rna. Accessed: 06-01-2017.

[57] MFE. Principle of minimum energy. https://en.wikipedia.org/wiki/Principle_of_minimum_energy. Accessed: 06-01-2017.

[58] microRNA. microrna definition. https://en.wikipedia.org/wiki/MicroRNA. Accessed: 06-01-2017.

[59] miRanda. miranda manual. http://cbio.mskcc.org/microrna_data/manual.html. Accessed: 06-01-2017.

[60] Kevin P Murphy. Machine Learning: A Probabilistic Perspective. MIT press, 2012.

[61] Nucleotides. Nucleotides and nucleic acids. http://www.bioinfo.org.cn/book/biochemistry/chapt12/bio1.htm. Accessed: 06-01-2017.

[62] Sarah M Peterson, Jeffrey A Thompson, Melanie L Ufkin, Pradeep Sathyanarayana, Lucy Liaw, and Clare Bates Congdon. Common features of microrna target prediction tools. Front Genet, 5:23, 2014.

[63] Marc Rehmsmeier, Peter Steffen, Matthias Höchsmann, and Robert Giegerich. Fast And Effective Prediction Of MicroRNA/Target Duplexes, volume 10. Cold Spring Harbor Lab, 2004.

[64] RNAcofold. Rnacofold manual. https://www.tbi.univie.ac.at/RNA/RNAcofold.1.html. Accessed: 06-01-2017.

[65] RNAfold. Rnafold manual. https://www.tbi.univie.ac.at/RNA/RNAfold.1.html. Accessed: 06-01-2017.

[66] RNAhybrid. Rnahybrid manual. https://bibiserv2.cebitec.uni-bielefeld.de/rnahybrid. Accessed: 06-01-2017.

[67] Roc. The area under roc curve. http://gim.unmc.edu/dxtests/roc3.htm. Accessed: 06-01-2017.

[68] Ventsislav Rusinov, Vesselin Baev, Ivan Nikiforov Minkov, and Martin Tabler. Microinspector: a web tool for detection of mirna binding sites in an rna sequence. Nucleic Acids Research, 33(suppl 2):W696–W700, 2005.

[69] Nicole T Schirle, Jessica Sheu-Gruttadauria, and Ian J MacRae. Structural basis for microrna targeting. Science, 346(6209):608–613, 2014.

[70] Bernhard Scholkopf and Alexander J Smola. Learning With Kernels: Support Vector Machines, Regularization, Optimization, And Beyond. MIT press, 2001.

[71] Sensitivity. Sensitivity definition. http://www.mathworks.com/help/phased/examples/detector-performance-analysis-using-roc-curves.html?requestedDomain=www.mathworks.com. Accessed: 06-01-2017.

[72] Martin Sturm, Michael Hackenberg, David Langenberger, and Dmitrij Frishman. Targetspy: a supervised machine learning approach for microrna target prediction. BMC Bioinformatics, 11(1):1, 2010.

[73] Antonio Gulli Sujit Pal. Deep Learning with Keras. Packt Publishing, 2017.

[74] Steven K. Thompson. Sampling. Wiley, 2012.

[75] Yuka Watanabe, Masaru Tomita, and Akio Kanai. Computational methods for microrna target prediction. Methods in Enzymology, 427:65–86, 2007.

[76] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. Data Mining: Practical Machine Learning Tools And Techniques. Morgan Kaufmann, 2016.

[77] Dong Yue, Hui Liu, and Yufei Huang. Survey of computational algorithms for microrna target prediction. Current Genomics, 10(7):478–492, 2009.