COMPUTATIONAL ESTABLISHMENT OF MICRORNA METABOLIC NETWORKS

A Dissertation Submitted to
the Graduate School of Engineering and Sciences of
İzmir Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in Molecular Biology and Genetics

by
Müşerref Duygu SAÇAR DEMİRCİ

June 2017
İZMİR

We approve the thesis of **Müşerref Duygu SAÇAR DEMİRCİ**

Examining Committee Members:

Assoc. Prof. Dr. Jens ALLMER
Department of Molecular Biology and Genetics, İzmir Institute of Technology

Prof. Dr. Anne FRARY
Department of Molecular Biology and Genetics, İzmir Institute of Technology

Prof. Dr. Şermin GENÇ
İzmir Biomedicine and Genome Center, Dokuz Eylül University

Assoc. Prof. Dr. Bünyamin AKGÜL
Department of Molecular Biology and Genetics, İzmir Institute of Technology

Assist. Prof. Dr. Zerrin IŞIK
Department of Computer Engineering, Dokuz Eylül University

12 June 2017

Assoc. Prof. Dr. Jens ALLMER
Supervisor, Department of Molecular Biology and Genetics
İzmir Institute of Technology

Prof. Dr. Volkan SEYRANTEPE
Head of the Department of
Molecular Biology and Genetics

Prof. Dr. Aysun SOFUOĞLU
Dean of the Graduate School of
Engineering and Sciences

ACKNOWLEDGMENTS

I would like to express my thanks to Assoc. Prof. Dr. Jens ALLMER, for his guidance throughout my graduate education including the preparation of this thesis. I feel that it is a great privilege to work with him.

I am grateful to my thesis committee members Prof. Dr. Anne FRARY and Assist. Prof. Dr. Zerrin IŞIK for their advices.

I am also thankful to TÜBİTAK for providing financial support through the project 113E326.

Finally, I would like to offer my special thanks to Yılmaz Mehmet DEMİRCİ, who has always been my source of strength and inspiration, for his encouragement, understanding and patience even during hard times of this study.

ABSTRACT

COMPUTATIONAL ESTABLISHMENT OF MICRORNA METABOLIC NETWORKS

MicroRNAs (miRNAs) are single-stranded, small, non-coding RNAs, that control gene expression at the post transcriptional level through various mechanisms such as translational inhibition, degradation and destabilisation of their target mRNAs. Despite the fact that thousands of miRNAs have been reported in various species, most still remain unknown. Due to this, the identification of new miRNAs is an essential process for analysing miRNA mediated post transcriptional regulation mechanisms. Moreover, many biological approaches suffer from limitations in their capacity to reveal rare miRNAs, and are further restricted to the state of the organism under examination. Such limitations have resulted in the construction of sophisticated computational tools for identification of possible miRNAs *in silico*. However, these programs suffer from low sensitivity and/or accuracy and as a result they do not provide enough confidence for validating all their predictions experimentally. In this study, the aim is overcoming these challenges by creating a new and adaptable machine learning based method to predict potential miRNAs in any given sequence. The efficiency of proposed method is shown by comparison with available tools on various data sets. By using this approach, miRNAs from the genomes of various organisms like human (*Homo sapiens*), fly (*Drosophila melanogaster*) and tomato (*Solanum lycopersicum*) are identified. Moreover, networks between the possible miRNAs of virus and human genes as well as the communications among nuclear and organelle genomes of *Solanum lycopersicum* through miRNAs are investigated.

ÖZET

MİKRORNA METABOLİK AĞLARININ BİLİŞİMSEL KURULUMU

MikroRNAlar (miRNAlar) tek diziden oluşan, küçük, kodlayıcı olmayan, hedef mRNAlarının translasyonel inhibisyonu, bozunması ve kararsızlaşması gibi çeşitli mekanizmalar aracılığıyla transkripsiyon sonrası seviyesinde gen ekspresyonunu kontrol edebilen RNAlardır. Farklı türlerde binlerce miRNA rapor edilmesine rağmen çoğu hala bilinmemektedir. Bu nedenle, yeni miRNAların belirlenmesi, miRNA aracılı transkripsiyon sonrası düzenleme mekanizmalarını analiz etmek için önemli bir işlemdir. Ayrıca, birçok biyolojik yaklaşım nadir miRNAları ortaya çıkarma kapasitesindeki sınırlamalardan muzdariptir ve inceleme altındaki organizmanın durumuyla daha da kısıtlıdır. Bu tür sınırlamalar olası miRNAların *in silico* olarak tanımlanması için karmaşık bilişimsel araçların yapımıyla sonuçlanmıştır. Ancak, bu programlar düşük duyarlılık ve/veya doğruluktan muzdariptir ve bunun sonucu olarak da tüm tahminlerin deneysel olarak doğrulaması için yeterince güven vermemektedir. Bu çalışmada amaç, verilen herhangi bir dizideki potansiyel miRNAları tahmin etmek için yeni ve uyarlanabilir makine öğrenme temelli bir yöntem oluşturarak bu zorlukların üstesinden gelmektir. Önerilen yöntemin verimliliği çeşitli veri kümeleri üzerinde uygun araçlar ile karşılaştırılarak gösterilmektedir. Bu yaklaşımı kullanılarak insan (*Homo sapiens*), meyve sineği (*Drosophila melanogaster*) ve domates (*Solanum lycopersicum*) gibi çeşitli organizmaların genomlarından miRNAlar tanımlanmıştır. Ayrıca, hem olası virüs miRNAları ve insan genleri arasındaki ağlar hem de *Solanum lycopersicum* nükleer ve organel genomları arasındaki miRNA vasıtalı iletişim incelenmiştir.

TABLE OF CONTENTS

LIST OF FIGURES

Figure **Page**

LIST OF TABLES

LIST OF ABBREVIATIONS

miRNA . . . . . microRNA
nt . . . . . nucleotides
TU . . . . . transcription unit
pri-miRNA . . . . . primary miRNA
pre-miRNA . . . . . precursor miRNA
RISC . . . . . RNA inducing silencing complex
NGS . . . . . Next Generation Sequencing
ML . . . . . Machine Learning
mfe . . . . . minimum free energy
SVM . . . . . Support Vector Machine
NB . . . . . Naïve Bayes
MLP . . . . . Multi Layered Perceptron
RF . . . . . Random Forest
APLSC . . . . . Asymmetric Partial Least Squares Classification
G2DE . . . . . Generalized Gaussian Density Estimator
FS . . . . . Feature subset selection
DR . . . . . Dimensionality reduction
PCA . . . . . Principal Component Analysis
GA . . . . . Genetic Algorithm
MCCV . . . . . Monte Carlo Cross Validation
TPR . . . . . True positive rates
TNR . . . . . True negative rates

CHAPTER 1

INTRODUCTION

MicroRNAs (miRNA) are single-stranded RNAs of approximately 22 nucleotides (nt) in length, that control gene expression through post transcriptional regulation by using either translational inhibition or destabilization of the target mRNAs (Piast et al., 2005; Filipowicz et al., 2008). The first example of miRNAs was discovered in *C. elegans*, as a regulator of the developmental timing (Lee et al., 1993).

In diverse organisms ranging from viruses to higher eukaryotes, important and various processes are regulated through miRNAs' action. many links have been established between miRNAs and human diseases such as cancer and neurodegenerative diseases (Table 1.1) and still many more are under investigation (Bushati and Cohen, 2007; Hébert et al., 2009; Wang et al., 2008).

It has been estimated that miRNAs control activities of about 30% of all protein-coding genes in mammals (Filipowicz et al., 2008). This situation is valid not only in higher eukaryotes, but also in simple multicellular organisms like poriferans (sponges) and cnidarians (starlet sea anemone) since they also possess miRNAs (Kim et al., 2009). In addition, most of the animal miRNAs appear to be phylogenetically conserved suggesting that these miRNAs are carried throughout evolution due to their important actions, e.g. around 55% of *C. elegans* miRNAs have homologues in humans (Ibáñez-Ventoso et al., 2008). On the other hand, there are certain differences between biogenesis pathways of animal and plant miRNAs suggesting that there might be alternative evolutionary paths (Chapman and Carrington, 2007; Millar and Waterhouse, 2005).

Around half of the mammalian miRNAs seem to have a tendency to locate close to other miRNAs forming clusters and be transcribed as a single polycistronic transcription unit (TU) (Lee et al., 2004). Nevertheless, it has also been observed that specific miRNAs might be originated from distinct genomic locations (Figure 1.1) (Kim et al., 2009). In addition, depending on alternative splicing events, it is likely to find some mixed miRNA genes that can be classified into either intronic or exonic miRNA types.

Table 1.1. Human diseases related with miRNAs. Numbers in column "Unique" indicate number of all unique diseases associated with related miRNAs. (Data obtained from HMDD v2.0 (Li et al., 2014).

miRNA	**Disease**	**Unique**
hsa-mir-21	Carcinoma (Hepatocellular), Diabetes Mellitus (Type 2), Lupus Erythematosus (Systemic), Fibrosis, Multiple Sclerosis	125
hsa-mir-155	Leukemia (B-Cell), Tuberculosis (Pulmonary), Down Syndrome, Behcet Syndrome	89
hsa-mir-146a	Influenza (Human), Alzheimer Disease, Creutzfeldt-Jakob Syndrome, Gerstmann-Straussler-Scheinker Disease	71
hsa-mir-17	Lymphoma (T-Cell), Toxoplasma, SARS Virus, Obesity, Schizophrenia, Hypertension	65
hsa-mir-125b-1	Dermatitis (Atopic), Breast Neoplasms, Huntington Disease, Glioblastoma, Myocardial Ischemia	61
hsa-mir-20a	Pulmonary Disease (Chronic Obstructive), Hodgkin Disease, Pre-Eclampsia, Polycystic Kidney Diseases	60
hsa-mir-34a	Muscular Disorders (Atrophic), Lymphoma (B-Cell), Fatty Liver (Alcoholic), Arthritis (Rheumatoid), Cardiovascular Disease	59
hsa-mir-145	Hepatitis (Chronic), Sarcoma (Ewing's), Adrenocortical Carcinoma, Stroke, Heart Failure	57
hsa-mir-221	Sarcoma (Kaposi), Lymphoma (Primary Effusion), Hyperglycemia, Marek Disease, Carotid Artery Diseases, Asthma	54
hsa-mir-125b-2	Dermatitis (Atopic), Leukemia (Myeloid), Nevus (Pigmented), Lung Diseases (Interstitial), Myotonic Dystrophy	53
hsa-mir-126	Hepatitis (Chronic), Myeloproliferative Disorders, Ischemia, Medulloblastoma, Parkinson Disease, Crohn Disease, Inflammation	51
hsa-mir-16-1	Polycythemia Vera, Spinal Cord Injuries, Patau Syndrome, Liver Failure, Odontogenic Tumors, Sepsis, Acute Lung Injury	50
hsa-mir-92a-1	Scleroderma (Systemic), Autistic Disorder, Burkitt Lymphoma, Eosinophilic Esophagitis, Polycystic Kidney Diseases	50

1.1. MicroRNA Biogenesis

The current version of miRBase (Release 21, http://www.mirbase.org) lists miRNAs from more than 200 organisms. Although there are certain differences in terms of proteins or paths followed for canonical miRNA biogenesis, three steps are usually essential; transcription of primary miRNAs (pri-miRNAs) from the miRNA genes (Lee et al., 2002), initial processing of precursor miRNAs (pre-miRNAs) in nucleus (Hutvágner et al., 2001) and the further processing and generation of mature miRNAs in the cytoplasm (Figure 1.2). Human miRNA biogenesis pathway can be summarized in five main steps:

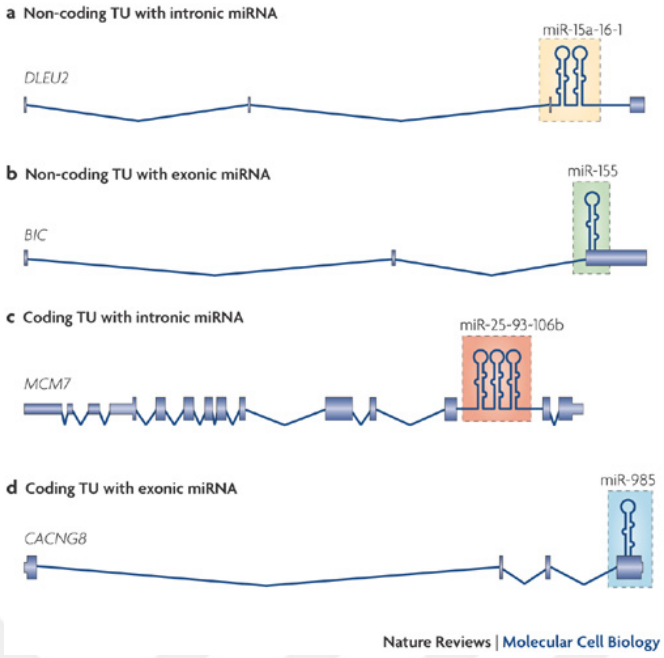

Figure 1.1. Possible genomic locations of miRNAs. a) Intronic miRNA cluster in non-coding transcripts, b) Exonic miRNAs in non-coding transcripts, c) Intronic miRNAs in protein-coding transcripts, d) Exonic miRNAs in protein-coding transcripts. The blue boxes indicate the protein-coding regions.(Source: (Kim et al., 2009))

1. Transcription: Most of the known miRNAs are transcribed by RNA polymerase II from various genomic locations (Figure 1.1). This pri-miRNA structure includes a double-stranded region with a hairpin loop and longer sequence extensions from the 5' and 3' ends of the hairpin in which other double-stranded regions might be found.

2. Microprocessor complex processing: The RNA-binding protein DGCR8 and DROSHA nuclease form the Microprocessor complex which results in removal of 5′ and 3′ ends of the pri-miRNA through endoribonucleolytic cleavage by the DROSHA (Lee et al., 2003). The cleaved RNA sequence also known as pre-miRNA has a structure of a short hairpin of about 60 to 70 nt.

3. Export from nucleus: Since miRNAs find their target mRNAs in cytoplasm, the pre-miRNA should be translocated from nucleus to cytoplasm in a complex with Exportin-5 and Ran-GTP. Binding of pre-miRNA to Exportin-5 needs at least 16 bp in the stem of the miRNA and the 3' overhang variations could affect the efficiency of this binding (Zeng and Cullen, 2004). It has also been demonstrated that pre-miRNA and Exportin-5 binding

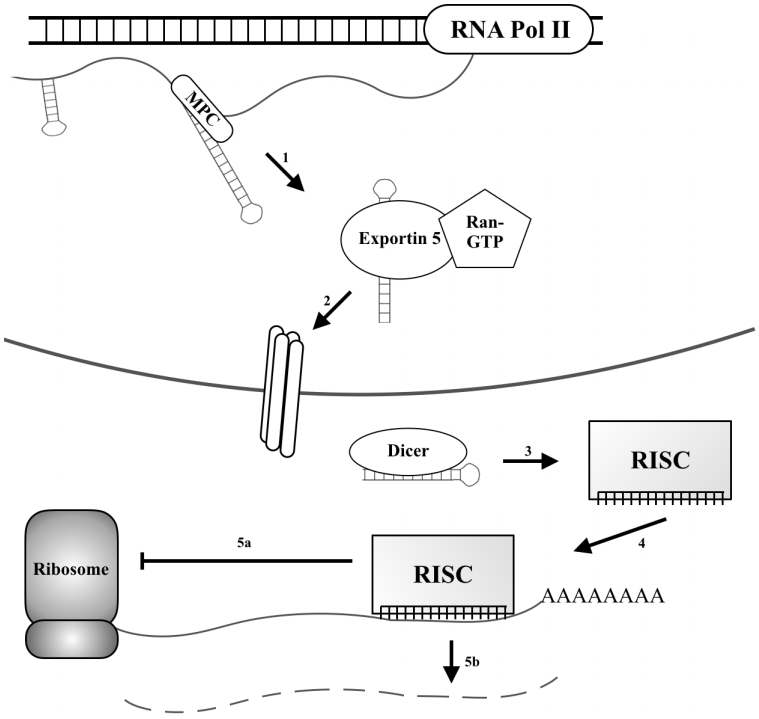

Figure 1.2. MiRNA biogenesis. RNA polymerase II transcribes pri-miRNA. The stem loop structure is recognised and cleaved by the microprocessor (MPC) resulting in pre-miRNA (1) which is then exported to cytoplasm by Exportin5 and Ran-GTP through nuclear pore complex (2). In cytoplasm, the pre-miRNA is further processed by Dicer and incorporated to RISC (3). RISC bound single stranded mature miRNA would be guided to target mRNA (4). The miRNA:mRNA interaction would lead either repression of protein expression (5a) or mRNA decay (5b).

is not only required for nuclear export but also for preventing the degradation of nuclear pre-microRNA (Zeng and Cullen, 2004).

4. Cleavage by DICER: Once in the cytoplasm the pre-miRNA is released from Exportin-5 through the hydrolysis of GTP and transfomed into mature miRNAs by RNase III enzyme, DICER cleavage (Lee et al., 2003). Different domains of DICER are responsible for the recognition and the cleavage of the pre-miRNA by removal of the loop structure and 3′ overhang in an ATP independent manner (Zhang et al., 2002).

5. Strand selection and targeting: After DICER processing, one strand of the miRNA duplex, the passenger strand, will be degraded while the other strand, the guide strand, will be guiding the Argonaute:miRNA complex, the RNA inducing silencing complex (RISC)

to target mRNAs. The selection of one strand over the other depends on thermodynamic properties of the duplex and in most of the cases the strand with the less thermodynamical stability at the 5′ end is selected (Khvorova et al., 2003).

1.2. MicroRNA Detection

There are many experimental and computational approaches to study and analyze miRNAs, ranging from classical molecular biology laboratory techniques like forward genetic screening to newer and high technology methods like Next Generation Sequencing (NGS).

1.2.1. Experimental Identification

The researchers who worked on a *C. elegans* mutant by application of forward genetics encountered intriguing findings. The mutation was in lin-4 which is a small noncoding RNA gene and its RNA had very interesting characteristics such as having a larger form with a stem-loop secondary structure (hairpin) and a smaller form originated from the stem-loop (mature sequence) (Lee et al., 1993). After this initial study, another similar RNA, let-7, was detected also in *C. elegans* which had a huge influence on miRNA area, since it was conserved among various organisms indicating that miRNA based post-transcriptional gene regulation is a general method (Reinhart et al., 2000).

Although a few other miRNAs have also been identified by forward genetics approaches, they are quite ineffective for miRNA gene detection due to numerous reasons (Berezikov et al., 2006):

- the small size of miRNAs
- further challenges in spontaneous or induced mutagenesis methods because of miRNAs' ability to tolerate mutations provided that the "seed sequence" is not involved
- it is still possible to fail to spot a miRNA mutant even though the miRNA gene is successfully hit and knocked out, since researchers usually focus on protein-coding regions to map a mutation but - miRNAs might be located in non-coding regions too (Figure 1.1)
-a phenotype-driven investigation will most likely fail to identify mutants due to redundancy (a mRNA can be targeted by many different miRNAs).

Other experimental techniques used for detection and/or validation of miRNAs include; Northern blotting, microarray and qRT-PCR. Each of these methods have some advantages and disadvantages (Table 1.2). Most of the experimentally detected miRNAs are identified through sequencing of size-fractionated cDNA libraries (Berezikov et al., 2006). There are many protocols developed by different groups and shown to be successful. Essentially the same principle is followed with differences in the details (Berezikov et al., 2006). Cloning approach for miRNA identification also suffers from limitations such as:

- tissue, stage, time specific and low level expressed miRNAs
- post-transcriptional modifications increasing the difficulty in cloning
- sophisticated bioinformatics methods required to analyze data.

Table 1.2. *In vitro* miRNA detection strategies. (Source: (Dong et al., 2013; Chugh and Dittmer, 2012; Baker, 2010))

Method	Good for	Suffers from
Northern blotting	used for novel and known miRNAs gold standard for validation	requires a large amount of total RNA detecting miRNAs with low abundance low-throughput and low-sensitivity relatively time consuming
Microarray	less expensive screening tool	lower sensitivity and dynamic range not for quantitative assays
qRT-PCR	the widest dynamic range and highest accuracy provide absolute miRNA quantification	normalization and specificity throughput issues
NGS	can easily detect sequence variation of miRNAs used for novel miRNAs high sensitivity	very complicated bioinformatics time consuming more expensive

1.2.2. Computational MicroRNA Prediction

Due to high cost and laborious steps of experimental detection as well as mentioned limitations, computational miRNA prediction has become an essential part of miRNA studies. One of the main and straightforward approaches is using already known miRNAs to search for their homologs in other organisms also known as “homology based search”. Although the system works well for widely conserved miRNAs, since it is not possible to predict non-conserved and/or species specific miRNAs by using this approach, *ab initio* methods become the only choice for a comprehensive search.

Independent from the choice of computational method used for miRNA identification, all predicted candidate miRNAs require experimental validation to be considered as a true miRNA. This means, we need a sensitive, highly accurate and specific method so that we will not have many false positives or false negatives.

1.3. Machine Learning Approaches for MicroRNA Analysis

Machine Learning (ML) is based on the idea that an algorithm can mimic human learning processes and extract rules to generate models. ML has become a popular method in various bioinformatics applications (Saçar and Allmer, 2013) since there are numerous biological fields such as genomics, systems biology, evolution, microarray and proteomics where ML can provide knowledge extracted from data (Larranaga, 2006).

While there are cases where unsupervised approaches on miRNA target prediction are applied (Heikkinen et al., 2011), ML for miRNA gene prediction is almost exclusively based on supervised learning in which a classification algorithm is trained for learning (Zhang and Nam, 2008).

In general, ML methods begins with obtaining data involving the sequence, structure and thermodynamic features characterising miRNAs such as minimum free energy (mfe) required for the secondary structure formation, number of A nt, length of hairpin sequence. Next, a classifier is trained by a set of known input data so it can generate rules based on these examples (input data; positive (known miRNAs) and negative (known non-miRNA examples)) (Lindow and Gorodkin, 2007). Then, the model generated by the classifier will be used on unknown samples to label them as miRNA or not.

There are many factors influencing the accuracy of the system but the most important ones include proper data analysis and the efficient selection of features, since data quality has a big impact on the overall process (Saçar et al., 2013) and calculation of features for the sequences is not an automatic process (Ding et al., 2010; Lindow and Gorodkin, 2007). At any rate, there are two main obstacles with the available machine learning based miRNA gene identification processes. The first one is the imbalance between positive and negative examples (Saçar and Allmer, 2013). Since the total number of actual miRNAs in a genome is not exactly known so far, it is assumed that there are only a few miRNA hairpins in any randomly chosen groups of hairpins obtained from the genome (Ding et al., 2010). In addition, the amount of positive examples is usually

smaller than that of negative ones. For example, one of the most common negative data sets used in miRNA classification analysis include around 9000 pseudo hairpins (Table 1.3) while miRBase lists less than 2000 human miRNA hairpins (Ng and Mishra, 2007). In our previous works, we showed that the imbalance problem between data sets can notably lead to reduction in the performance of machine learning approaches (Saçar and Allmer, 2013). The second major obstacle is that majority of the machine learning based methods makes assumptions about the features defining the data sets e.g., the length of the stem, the loop size and mfe. Consequently, if a sequence is found to be outside of these fixed limits, it is not treated as a possible miRNA thus cannot be identified by those methods which would cause an increased false negative ratio (Ding et al., 2010).

Table 1.3. List of published studies for *ab initio* miRNA prediction. Listed are the number of features that were effectively used, the training data that was employed and whether an implementation is available. Table is sorted by number of citations in Google Scholar. Use: + means it exists, - means there is no implementation, * means we experienced problems with the implementation.

Study	ML Algorithm	Feature Number	Positive Data	Negative Data	Use	Year	Cited
Xue	SVM	32	MiRBase 5.0	pseudo	*	2005	354
Jiang	RF, SVM	34	MiRBase 8.2	pseudo	*	2007	314
Ng	SVM	29	MiRBase 8.2	pseudo	*	2007	174
Batuwita	SVM	21	MiRBase 12	pseudo and other human ncRNAs	+	2009	142
Xu	SVM	35	MiRBase 2007	fragments from human genome	*	2009	70
Ding	SVM	32	Known miRNAs	UTRdb and ncRNA from Rfam 9.1	-	2010	44
Burgt	L score classifier	18	MiRBase 9.0	-	*	2009	25
Bentwich	-	26	hairpins from human genome	non-coding regions	-	2008	19
Gudys	NB, MLP, SVM, RF, APLSC	28	MiRBase 17	genomes and mRNAs	+	2013	19
Ritchie	SVM	36	MiRBase 17	non-Dicer transcripts	-	2012	17
Lopes	SVM, RF, G2DE	13	MiRBase 19	pseudo	*	2014	11
Gao	SVM	57	MiRBase 20	exones, ncRNAs (rFam)	*	2013	4
Chen	LibSVM	99	MiRBase 2013	pseudo and Zou	+	2016	-

There are various supervised machine learning algorithms such as Support Vector Machine, Naïve Bayes (NB), Multi Layered Perceptron (MLP), Random Forest (RF), Asymmetric Partial Least Squares Classification (APLSC), Generalized Gaussian Density Estimator (G2DE) and most of them have been used in miRNA analysis (Table 1.3).

1.3.1. Data Sets

While using classification algorithms for miRNA gene prediction, selecting positive examples seems considerably easier since almost exclusively known miRNAs are used as positive data, while finding or creating negative samples are highly challenging (Lindow and Gorodkin, 2007).

In general, positive data are obtained from miRBase (Kozomara and Griffiths-Jones, 2011), despite the fact that some of the entries in miRBase which are claimed to be miRNAs, do not have the required characteristic properties (e.g., having one terminal loop) to be labelled as true miRNAs. In various studies, it has been shown that when positive controls are taken from miRBase, further steps are necessary to produce efficient high-confidence positive controls (Xue et al., 2005). We also analysed this problem and showed that removing arguable miRNAs from miRBase increases prediction accuracy (Saçar and Allmer, 2013). Furthermore, in this study we found out that using filtered miRBase hairpins for mouse would produce higher true prediction scores (Figure 3.4).

A well-designed negative data set is one of the crucial prerequisites for an efficient classifier generation. If negative data are very artificial and do not have any similarity to positive data, there is an increased probability that the classifier will not be trained adequately to differentiate among real biological unknown sequences (Wu et al., 2011). Contrarily, in the case of negative data set being very similar to the positive data set, the classifier will have trouble discriminating between negative and positive (Wu et al., 2011).

For any RNA sequence to be classified as miRNA, it should be recognized and processed by the enzymes Drosha and Dicer (Figure 1.2). During construction of high quality negative samples, it is important to select sequences that are expressed in the same or similar manner as true miRNAs but are not recognized by Dicer. Since this is a very complex method to create negative samples, many tools use random genomic sequences and/or intronic, exonic sequences (Brameier and Wiuf, 2007; Xue et al., 2005). Nonetheless, these approaches produce weak negative data because there is no guarantee

that these randomly selected small RNAs would not be turned into functional mature miRNAs (Xue et al., 2005).

There are various negative data sets used in many studies (Table 1.3) and all available ones are included in this work (Section 2.1) in addition to new negative data sets designed by us. Their performances are compared and analysed showing that some of the negative data sets are more challenging than the others (Figure 3.5)

1.3.2. Feature Selection

While using machine learning for pre-miRNA detection, it is required to define features describing a miRNA and examples of such features have been suggested in the literature (Table 1.3). The common features that have been used for pre-miRNA analysis can be grouped into four major categories; sequence-based, structural, probability-based, and thermodynamic (Figure 1.3). All of the features that are used in this work can be found and calculated by using the website (http://jlab.iyte.edu.tr/software/mirna). Some examples for these features are:
Sequence based features; 16 dinucleotide frequencies %NN (%AA, %AC, %AG, %AU, %CA, %CC, %CG, %CU, %GA, %GC, %GG, %GU, %UA, %UC, %UG, %UU), regular internal repeat (dr), GC content (%GC), etc.
Structural features; 32 triplet elements i.e. A(((, U(((, U(.(, U.((, G(((, C(((, C(.(, hairpin length (hpl), hairpin loop length (hll), maximal bulge size (mbs), etc.
Thermodynamics based features; ensemble free energy (efe), ensemble frequency (efq), melting temperature (Tm), enthalpie (dH), entropy (dS), etc.
Probability based features derived from dinucleotide shuffling (dns); adjusted base pairing propensity (dP), adjusted minimum free energy of folding (dG), MFE index 1 (MFEI1), degree of compactness (dF), etc.

As it has been argued and demonstrated in various cases, one of the biggest hurdles that makes classification analysis drastically difficult is the increase in the dimensionality of the data. In addition, the data might not only be big in size but also sparse in the space it occupies. Such situations could cause huge troubles for ML, this phenomenon is also known as the curse of dimensionality (Powell, 2011). Consequently, using a lot of features can result in lower classification accuracy with a high computational cost.

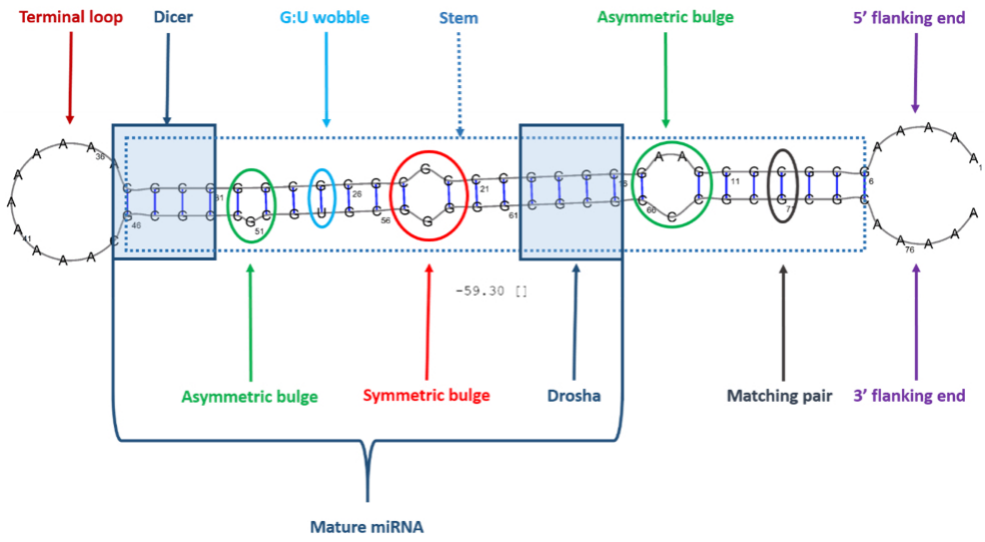

Figure 1.3. MiRNA hairpin structure. Some of the features of hairpin structure: G-U matches, symmetric and asymmetric bulges, stem length, mfe etc. Secondary RNA structure and mfe value are obtained by using RNAShapes (Steffen et al., 2006).(Source: (Saçar et al., 2014)

When dealing with real biological data sets, important features are mostly unknown a priori. Hence, to properly cover and represent the data sets, a very high number of features are designed and calculated. However most of these features tend to be not informative/beneficial. In cases like miRNA precursor analysis where data set size can be large, it is essential to remove irrelevant/redundant features to have a better learning in less time.

Two main methods are usually preferred for reducing the feature size: Feature subset selection (FS) approaches like filtering and dimensionality reduction (DR) strategy like Principal Component Analysis (PCA). Since DR based algorithms achieve dimensionality reduction through creating new features as combinations of the original ones, it is not possible to assess the value of a single feature with these methods, so for the rest of the section we will focus on FS approaches.

There are various definitions by many authors for FS based on the content of the analysis (Dash and Liu, 1997):

1. Idealized FS: it is based on searching for the minimum number of adequate feature subset.

2. Classical FS: where N is greater than M, selecting a subset of M features from a set of N features, so that the value of a criterion function is optimized over all subsets of size M (Dash and Liu, 1997).
3. Increasing prediction performance: it is based on selecting a subset of features either for improving prediction performance (accuracy, sensitivity etc.) or without any significant decrease in the classification performance decreasing the amount of the features.
4. Approximation based FS: selecting a small feature subset which keeps the class distribution as close to the initial value as possible.

The main categories of FS algorithms include filtering, wrapper, and embedded approaches. Unlike the first two, embedded FS is a process performed as a part of a ML algorithm. Filtering techniques are classifier independent and comparatively computationally simpler and faster (Saeys et al., 2007). Moreover, a feature ranking score such as Fisher score, Pearson correlation and information gain is calculated, and features with lower scores from the set thresholds are removed (Janecek et al., 2008).

Wrapper algorithms benefit from including a ML method for FS through a feedback mechanism. In other words, they depend on the classifier to make a discrimination among sets of features (Janecek et al., 2008). Feature subsets' space is searched and for each feature subset a measure of classifier performance like accuracy is calculated (Kohavi and John, 1997). Differently from filtering techniques wrapper approaches take feature dependencies into consideration (Saeys et al., 2007). However, the disadvantage of wrapper approach against filtering is that it has a higher probability of overfitting and more computational cost (Janecek et al., 2008; Saeys et al., 2007).

When a FS method is designed to find the very best subset of features, it must perform an exhaustive search meaning that all possible 2^N candidate subsets need to be considered and evaluated based on some evaluation function (Dash and Liu, 1997). The problem is that although this is the only way to make sure the best feature subset is selected, such exhaustive search, also known as NP-hard, is too time consuming, expensive and complex (Dash and Liu, 1997; Guyon and Elisseeff, 2003). There are many search strategies including best-first, branch-and-bound, simulated annealing, genetic algorithms and it seems greedy search methods; forward selection and backward elimination seem to be especially beneficial with less computational cost and robustness against overfitting (Guyon and Elisseeff, 2003).

Methods using heuristic or random search applications are designed for reducing computational complexity (Dash and Liu, 1997). To avoid eventual exhaustive search of

subsets, these algorithms require defined stop criteria. Nevertheless, there are at least four essential steps in a regular FS method: (1) a generation method to create the following candidate subset; it can start with (i) no features, (ii) with all features, or (iii) with a random subset of features; (2) an evaluation function/measure to evaluate the current feature subset; (3) a stopping condition to figure out when to stop searching and (4) a validation procedure to analyse if the obtained subset is valid (Dash and Liu, 1997).

Genetic Algorithm (GA) can be considered as computational model of evolution. The algorithm searches a set of possible solutions; each candidate solution for the given problem is named as a "chromosome", and the complete set of solutions is known as a "population". GA algorithm works in an iterative movement, passing from one population of chromosomes to the next one and each of these iterations is called a "generation" (Chtioui et al., 1998). In each generation, the population is ranked based on the fitness score of each chromosome and two well-adapted chromosomes are survived/selected as parents for reproduction (Chtioui et al., 1998). The chromosomes are traditionally represented as n-bit binary vectors (Vafaie and Jong, 1992; Shahamat and Pouyan, 2015). If there are 1000 features then each chromosome is a binary vector of dimension 1000; a bit value of 0 means that the corresponding feature is not selected, and if the bit is 1, the feature is selected (Shahamat and Pouyan, 2015).

The main parts of GA construction are listed (Figure 1.4):

1. Randomly generated initial population: The setting of number of chromosomes present in the initial population is an essential part for GA performance; while a large population produces more genetic diversity, it has slower convergence and a small population has the risk of converging to a local extreme (Xuan et al., 2011).

2. Fitness function calculation: The fitness scores are required for ranking, deciding quality of chromosomes and making selection for further steps.

3. Genetic operators: for producing the next population, a set of well-adapted feature subsets should be selected for crossover and mutation.

3a. Selection: In general, roulette wheel selection is applied to select individuals since it is known to decrease the probability of the reaching local optimal resolution (Shahamat and Pouyan, 2015; Xuan et al., 2011). While higher fitness score indicates greater chance of survival it does not eliminate the chance of the weaker ones to survive (Xuan et al., 2011).

3b. Crossover: A crossover operator such as single-point crossover is applied randomly to parents. The offspring will be created using parts from one parent and the remaining

parts from the other parent based on the location of crossover. The defined crossover rate will define how many individuals would be affected e.g., crossover rate of 20% means that 20% of individuals would be taking part in the crossover (Xuan et al., 2011).

3c. Mutation: For maintaining the diversity in a population a mutation operation can be applied. At a randomly chosen point (P), the value of bit is reversed; if it is 1 it would become 0 and the vice versa (Shahamat and Pouyan, 2015; Xuan et al., 2011).

4. Stop criteria: For ending the genetic iteration process a stop condition must be defined. There are many different criteria that can be used for this purpose such as defining maximum number of iterations, analysing fitness score trend; e.g. if the average fitness of the population remains stable for consecutive N iteration or if the difference is smaller than a threshold (Shahamat and Pouyan, 2015; Xuan et al., 2011).

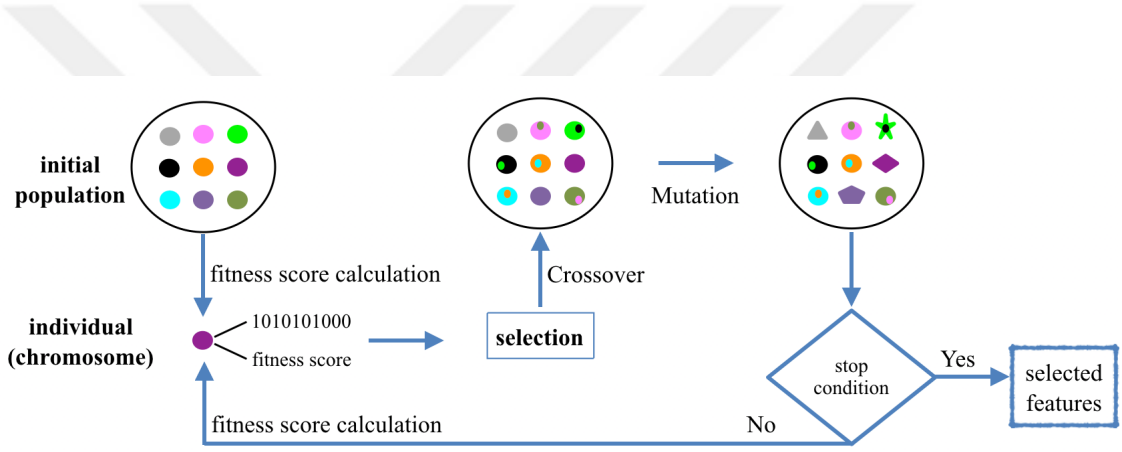

Figure 1.4. Basic GA workflow. The main constituents of GA are represented.

GA is mostly acknowledged for its ability to successfully search large spaces for finding an optimal or near optimal feature subset when not much information is known *a priori* (Vafaie and Jong, 1992; Shahamat and Pouyan, 2015). Moreover, as GA is comparably insensitive to noise in the data, it becomes a good choice for robust feature selection to increase the performance of classification applications (Vafaie and Jong, 1992).

1.4. MicroRNA Metabolic Networks

As already mentioned in previous sections, miRNAs take part in a wide range of networks due to their post-transcriptional regulatory function. In addition, their own biogenesis pathway is regulated at many steps (sometimes by miRNA itself). Moreover, in some cases while a miRNA can target hundreds of distinct targets (one to many), it is also observed that a miRNA-target interaction might be an one to one relation. Furthermore, a mRNA might be targeted by numerous different miRNAs as well. All of these possibilities combined with the tissue and time specific expression of both miRNA and target make a complete miRNA network analysis a very challenging task. Consequently, many of the published work regarding miRNA networks focus on a specific condition like conserved miRNA interactions in a disease phenotype (Zafari et al., 2015)

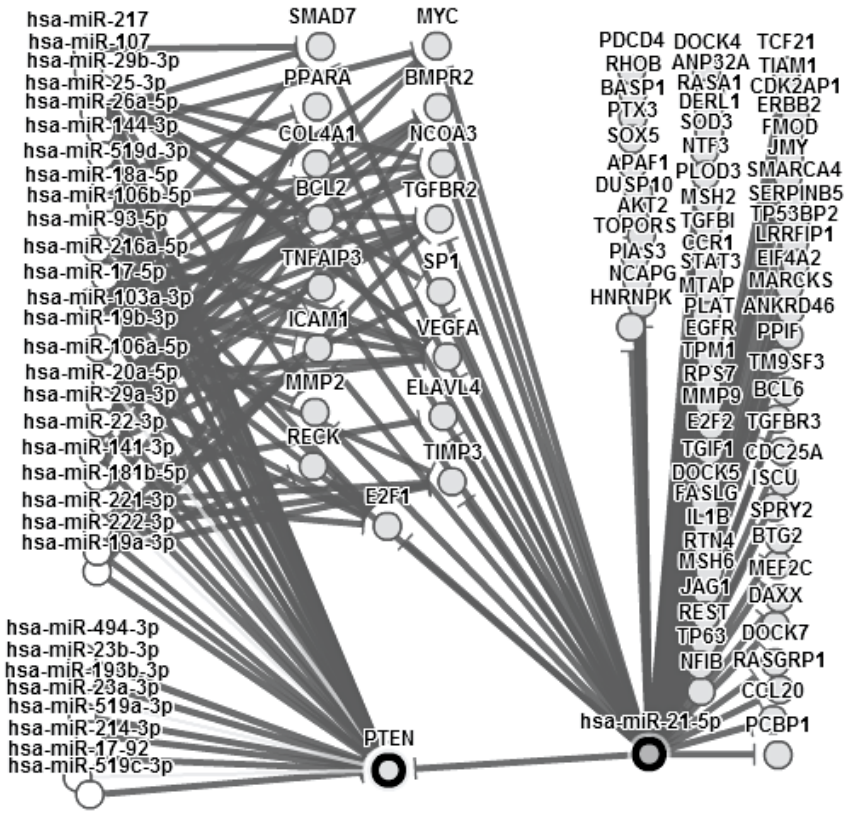

Figure 1.5. Simplified network for hsa-mir-21-5p. Data and image are obtained from miRTarBase (Chou et al., 2016). Note: not all of the targets are included (>500), only interactions with strong experimental evidence are shown.

1.4.1. Viral MicroRNAs

Various research outcomes indicate that many disease phenotypes are linked to activities of miRNAs (Table 1.1). Most of such links are discovered in cases where a miRNA and its targets are co-expressed in the same cell from the same genome. MiRNAs have also been associated with the sophisticated cross-talk between host and pathogen (Saçar et al., 2014) and are shown to be play a major role in viral pathogenesis (Gottwein and Cullen, 2008). Despite the fact that more research is needed to comprehend overall host-miRNA communications, recent studies indicate that viral miRNAs might have effects on the host cell (Grundhoff and Sullivan, 2011; Skalsky and Cullen, 2010). Nevertheless, given the fact that by encoding miRNAs, viruses can modulate the type and quantity of host genes to generate an environment suitable for viral replication makes miRNAs powerful and beneficial tools. In addition, high evolution rates of miRNAs provide the opportunity of easier adaption to new host targets. Finally, the most important advantage of using miRNAs for viruses would be escaping immunogenic response since the host itself produces miRNAs through similar pathways (Skalsky and Cullen, 2010).

Most of the viral miRNAs listed in miRBase are based on DNA viruses. Whether RNA viruses can generate functional miRNAs is a hot discussion topic. Some researchers find the idea not plausible due to various facts; if the RNA viruses under question are the ones replicating exclusively in the cytoplasm they would not be able to access to nuclear Drosha and more importantly, for a nuclear RNA virus excision of miRNA would eventually mean the cleavage and destruction of the viral genomic RNA which is not a favourable outcome for the virus (Skalsky and Cullen, 2010). Due to these hypotheses, one would expect for nuclear RNA viruses to skip the evolution step leading to miRNA production but growing evidence suggests that viruses with RNA genomes like Human Immunodeficiency Virus (HIV) can express miRNAs (Ouellet et al., 2008).

1.4.2. MicroRNAs of Mitochondria and Chloroplasts

Due to canonical miRNA biogenesis (Figure 1.1), almost all of the effort for identifying and analysing miRNAs is spent on nuclear DNA originated miRNAs. There are only a few works dealing with mitochondrial miRNAs. Interestingly one of these studies confirmed the localisation of pre-miRNA and mature miRNA in the mitochondria but the

authors could not determine the origin of these miRNAs; whether they are imported from the cytosol through a translocation system, and/or they could be processed in the mitochondria (Barrey et al., 2011). The presence of pre-miRNAs might suggest that at least some of the miRNA biogenesis machinery could be found in mitochondria so that these pre-miRNAs would be transformed into mature miRNAs. This hypothesis later proven to be true by two papers; the first one showing the presence of Ago2 in mitochondria (Dasgupta et al., 2015) and the second one explaining how miRNAs enhance mitochondrial translation during muscle differentiation with Ago2 (Zhang et al., 2014). Nevertheless, these results demonstrate the translocation of nuclear genome encoded miRNAs into the mitochondria but whether mitochondrial genome encoded miRNAs exist is still a question under investigation.

The presence of miRNAs in chloroplast is still far from being generally accepted. Wang et. al. reported functional and heat responsive chloroplast derived small RNAs in Chinese cabbage (*Brassica rapa*) which may suggest a miRNA-like regulation inside chloroplasts (Wang et al., 2011).

1.5. Aim

MiRNA mediated post-transcriptional regulation mechanisms have been an essential research topic for past years. Therefore, it is essential to have an effective approach for analysing miRNAs to be able to use them in various applications such as disease markers and treatment for human diseases. Such methods should be designed in a way that it would be able to detect the miRNAs in a given genome sequence. In this work, current literature is scanned and among all the studies using machine learning based *ab initio* miRNA identification, the most promising 13 are compared. In addition, new consensus models are generated to have a more accurate system which would make further experimental validation steps easier due to lower false positive and false negative ratios.

Here the developed workflow is constructed not only to increase the accuracy of prediction, but also to work well in other organisms. It is tested if the models created by using learning data specific for human miRNAs, will have high true prediction rates for almost all of the organisms' miRNAs listed in miRBase. Moreover, for miRNA precursor identification, a genome wide approach covering all candidate miRNA repository of any genome is designed and used in distinct organisms like viruses, *Drosophila*

melanogaster, *Homo sapiens* and *Solanum lycopersicum*. Further network analysis between the predicted miRNAs of retro-transcribing viruses and human genes and cross-talk search among the nuclear, mitochondrial and chloroplast genomes located in *Solanum lycopersicum* cells would provide new insights regarding miRNA actions.

CHAPTER 2

METHODOLOGY

In this study, classification for model generation, predictions on data sets and all the data analysis were performed using KNIME (Berthold et al., 2009) which is a workflow management and data analytics platform. The created workflows were named as izMiR with subparts such as izMiR prediction, izMiR learning and izMiR models. The data sets and izMiR workflows are further explained in detail on Nature Protocol Exchange (https://www.nature.com/protocolexchange/protocols/4919).

2.1. Data Acquisition

Positive examples were obtained from miRBase (release 21), the standard data source for positive data used in many *ab initio* pre-miRNA prediction (Table 1.3). For learning, human miRNAs from miRBase were used as positive data after certain filtering steps like excluding hairpins with identical sequences. At the end, 1828 human pre-miRNAs were used for the human training data set. However, for prediction analysis unfiltered miRBase data were used.

In the *Drosophila melanogaster* analysis additional models are generated by using 256 hairpins from miRBase as the positive data set.

For *Solanum lycopersicum* analysis miRNAs of *Solanum lycopersicum* (77 hairpins), *Nicotiana tabacum* (162 hairpins) and *Solanum tuberosum* (224 hairpins) were used as positive data. As negative data, 980 plant specific negative hairpin examples were obtained from http://nclab.hit.edu.cn/PlantMiRNAPred/.

As mentioned in Section 1.4.1, the quality of data sets have a huge impact on overall performance. Therefore, to be able to have a better understanding on the performance of our miRNA detection approach, various positive and negative data sets were retrieved and created to use in prediction. Apart from the data sets downloaded from miRBase, the remaining data sets are listed as:

Pseudo: data set was generated by Xue (Xue et al., 2005) but downloaded from Ng (Ng and Mishra, 2007), used for learning (as negative) and prediction (8492 hairpins)

Shuffled: created by shuffling sequences of human positive data from miRBase, used for prediction (1423 hairpins)

NotBestFold: created by not using the best fold proposed by RNAFold for human hairpins from miRBase, used for prediction (1881 hairpins)

NegHsa: previously published by Gudys (Gudyś et al., 2013), used for prediction (68048 hairpins), the original data set (http://adaa.polsl.pl/agudys/huntmi/huntmi.htm) had many duplicate identifiers and we filtered them by keeping only one and removing the rest, thereby reducing the amount of data from 87000 to 68000 examples

Zou: previously published by Zou (Wei et al., 2013), used for prediction (14246 hairpins)

Chen: previously published by Chen et al. (Chen et al., 2016), combination of examples from Pseudo and Zou, used for prediction (3054 hairpins)

hsaFR: created by generating random numbers between minimum and maximum values of each feature in human miRNA data set based on miRBase, used for prediction (5000)

hsaBQ: created by generating random numbers between lower quartile and upper quartile values of each feature in human miRNA data set based on miRBase, used for prediction (5000)

hsaAM: created by generating random numbers between 40 quantile and 60 quantile values of each feature in human miRNA data set based on miRBase, used for prediction (5000)

pseudoFR: created by generating random numbers between minimum and maximum values of each feature in pseudo data set, used for prediction (5000)

pseudoBQ: created by generating random numbers between lower quartile and upper quartile values of each feature in pseudo data set, used for prediction (5000)

pseudoAM: created by generating random numbers between 40 quantile and 60 quantile values of each feature in pseudo data set, used for prediction (5000)

2.2. Feature Selection

For our previous studies, more than 1000 features including the ones used or proposed in the 13 studies compared here (Xue et al., 2005; Jiang et al., 2007; Ng and Mishra, 2007; Batuwita and Palade, 2009; Xu et al., 2009; Ding et al., 2010; van der Burgt et al.,

2009; Bentwich, 2008; Gudyś et al., 2013; Lopes et al., 2014; Gao et al., 2013; Chen et al., 2016) (Table 1.3) and the new features designed by us have been implemented and tested. In the literature, some of the proposed features were not explained clearly so they were implemented to the best of our understanding. There are various ways to calculate features in pre-miRNA analyses (Yones et al., 2015) and our approach is available on our group web page: http://jlab.iyte.edu.tr/software/izmir.

Previous analysis in the feature sets revealed that while some of the features have high correlation with each other, some of them do not provide any information gain (Saçar and Allmer, 2013). Considering the time and computational power required to calculate features, an efficient feature selection methodology becomes an essential component of the overall analysis.

In order to have a good feature selection process in a reasonable amount of time, an in house distributed GA system was used in this study. The method developed by Mustafa Toprak (IYTE, Computer Engineering) depends on distributing the evaluation of feature subsets through HTCondor (Litzkow et al., 1988) and using KNIME platform (Berthold et al., 2009) to measure the classification accuracy of each feature subset group.

GA parameters were adjusted as listed:

1) Population size: 1000
2) Number of generation: not predefined
3) Probability of crossover: 1
4) Probability of random mutation: 0.01
5) Crossover strategy: Random single point
6) Stop condition: the score of the best individual was not improved for five generation

2.3. Learning Workflows

While performing classification it is vital to make a design carefully to avoid class imbalance problems that can affect the overall performance significantly (Saçar and Allmer, 2013). Unfortunately, well-known methods such as k-fold cross validation and leave-one-out have many problems (Kohavi, 1995; Varma and Simon, 2006) which lead us to sample positive and negative data separately. From both data sets equal amounts of data points were selected randomly and the selected samples were further divided into

training (70%) and testing sets (30%) through random sampling, a system also known as Monte Carlo Cross Validation (MCCV) (Figure 2.1). The training data were used to train three classifiers NB, DT and SVM. For each classifier, their performance scores and PMML models were obtained for each iteration. After 1000 iterations of the sampling and learning procedure, the best PMML models for NB and DT but not for SVM since Weka LibSVM (3.7) which did not provide PMML outputs, was used for the analysis because it was much faster than other SVM classifiers available in KNIME. Thus, SVM models were not saved, but for comparison with DT and NB, SVM scores produced during learning and testing were used. Model performances were evaluated by analysing measures such as recall, precision, sensitivity, specificity, F-measure, accuracy, Cohen's kappa, and Youden's index. The training workflow ensured that each study (feature set) and classifier receives exactly the same data in each iteration for making a fair comparison.

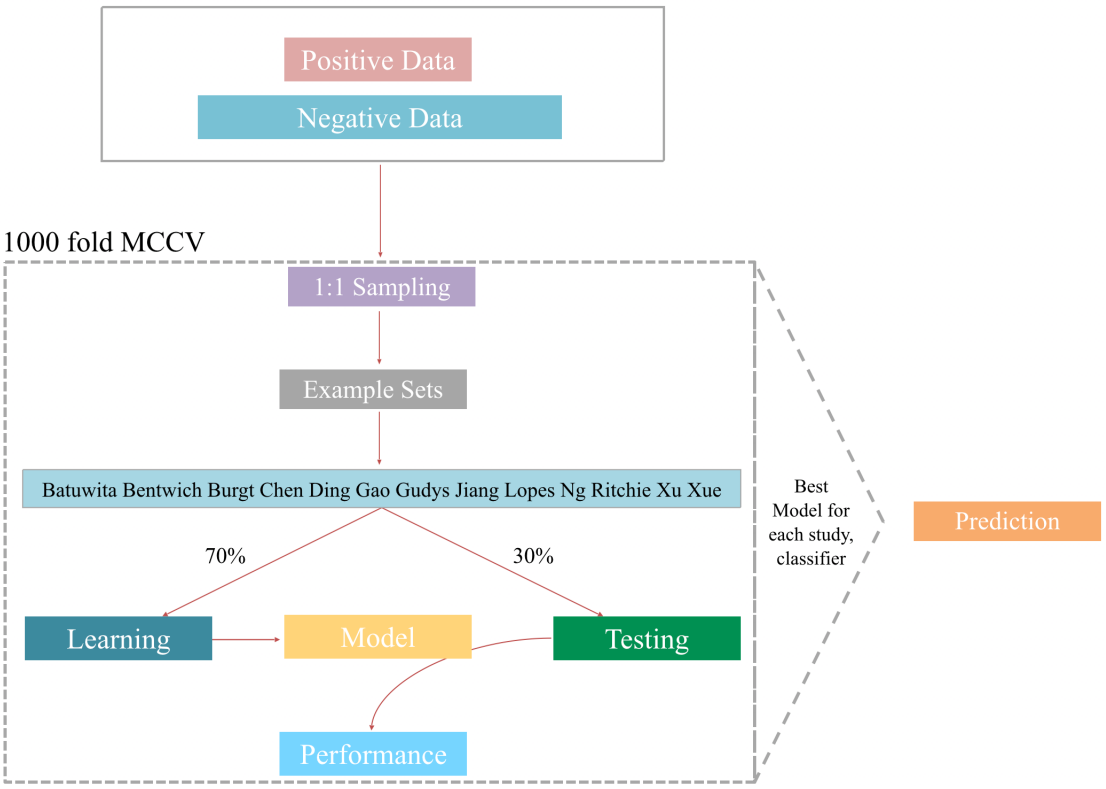

Figure 2.1. Learning workflow. Each study (feature groups) were trained and tested on exactly the same data sets with 1000 MCCV.

For finding mature sequences in the predicted hairpin sequences of *Solanum lycopersicum*, 612 mature sequences from *Solanum lycopersicum*, *Nicotiana tabacum* and *Solanum tuberosum* were used as positive data and negative data set was constructed by shifting the mature sequences by half of their length in the hairpin sequences and using the new extracted sequence. Various features were calculated such as: sequence length, number of matches and mismatches in the mature sequence region, single nucleotide counts (4), dinucleotide counts (16), trinucleotide counts (64), distances of start and end positions to 3' and 5' loop start, loop end etc. These data sets were applied to Random Forest learner through 1000 MCCV, 70% to 30% learning/testing ratio.

2.4. Prediction Workflows

For prediction, DT and NB models with the highest accuracy, and F-measure scores for each study were loaded into another workflow in KNIME (prediction workflow). With 13 studies and 2 classifiers there were 26 models that could be applied to input data for searching pre-miRNAs and providing scores to each predictions. In addition to these 26 individual models, some consensus approaches were generated to benefit from each studies' performance for improving the overall classification performance. The consensus models are also available in the izMiR framework provided (http://jlab.iyte.edu.tr/software/izmir).

In total six consensus models were designed based on majority vote, rule based prediction score evaluation, and a model generated from prediction scores:

(a) Majority vote: Equal weights were given to each model and a given sequence was predicted as miRNA by **ConsensusDT** and/or **ConsensusNB** models only if it was predicted as miRNA in at least 6 studies.

(b) **ConsensusRule**: if average DT score or average NB score was larger than 0.89 (lower quartile value of human data) for a given sample, then it was labeled as miRNA while average DT score or average NB score was less than 0.5, it was labeled as negative. The remaining samples were labeled as a candidate miRNA.

(c) Average of prediction scores: **AverageDT** and **AverageNB** were performed individually in the same manner to ConsensusRule with different thresholds; if the average value of prediction scores was smaller than 0.5, it was labeled as negative, otherwise it was predicted as miRNA.

(d) **ConsensusModel**: 26 models from studies were applied on the learning data (human miRNAs and pseudo data set) and the prediction scores, ranging between 0 and 1, were used to train and test a MLP classifier through the same approach shown in Figure 2.1. The model with the highest accuracy and F-measure was then used for predictions.

Any given input data for prediction were applied to all 32 described models and the numbers of hairpins predicted as miRNA, negative, and candidate were calculated for all of the models. True positive rates (TPR) and true negative rates (TNR) were provided as performance measures:
TPR = (number of hairpins correctly classified as miRNA / number of all hairpins)*100
TNR = (number of hairpins correctly classified as negative / number of all hairpins)*100

For constructing ROC curves, the models were applied to human pre-miRNAs from miRBase and pseudo negative data set.

For *Solanum lycopersicum* analysis, by using the mentioned data sets and selected features from GA, models from various classifiers; DT, LibSVM, MLP, NB and RF were trained and used for prediction.

2.5. Hairpin Extraction from Genomes

Before extracting hairpins, the genome was first divided into 500 nt fragments with 250 nt overlaps, the sequence was converted to RNA by changing T as U (T -> U) and by creating the reverse complementary for the template strand (Figure 2.2). RNAfold (Hofacker, 2003) was used for all secondary structure generation. Regular expressions were used to extract all structures that carry hairpin characteristics; e.g. a stem region with at least three consecutive matches and a terminal loop with at least three nucleotides. The extracted hairpins were filtered by removing duplicates and excluding hairpins which did not fit into length distribution of human hairpins listed in miRNA hairpin. All required features for remaining hairpins were calculated and analysed with the prediction workflow.

The human genome (GRCh38, DNA, primary assembly) contains 12,399,093 fragments from which 108,788,895 putative hairpins for one strand and 108,276,240 hairpins for the other were extracted. Filtering based on hairpin length (between 36 and 180; representing the smallest and the longest human stem loops in miRBase) was ap-

plied and after removing duplicate sequences from the 34,856,229 length-filtered hairpins, 27,932,492 putative pre-miRNA sequences remained. The same filtering approach resulted in 28,074,667 hairpins for the other strand.

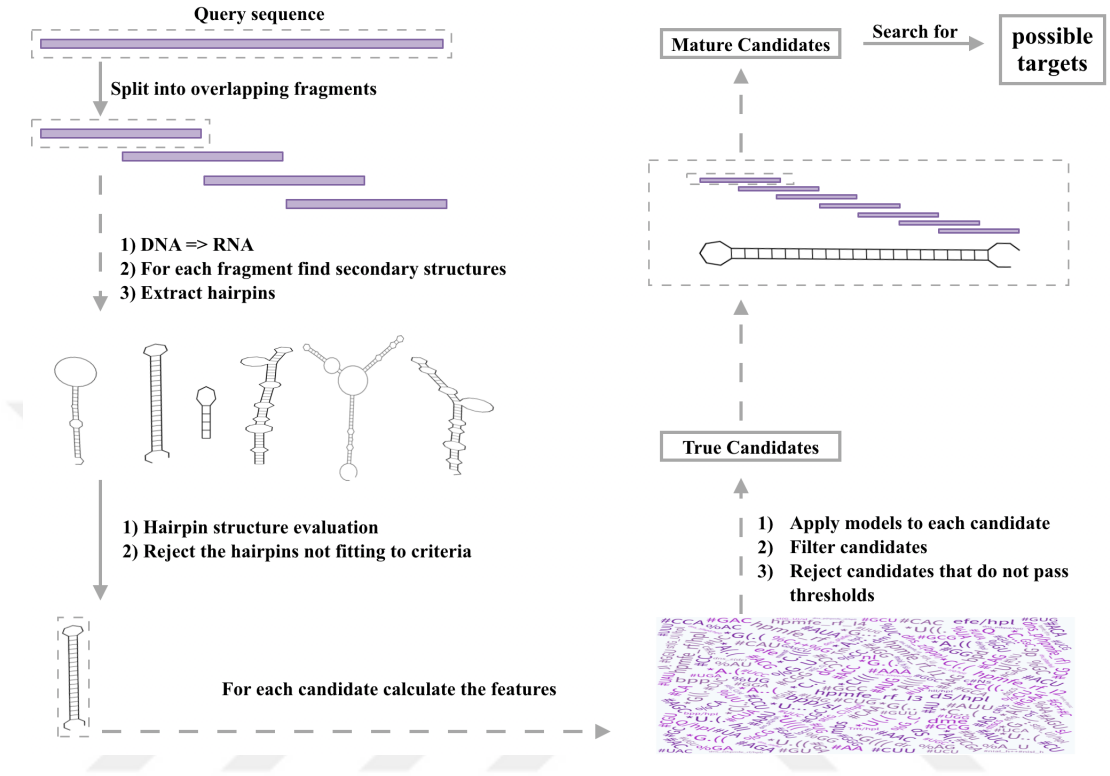

Figure 2.2. Genome wide search strategy for miRNAs.

On a personal computer, it would take several weeks calculating more than 800 features for all putative hairpins. Thus, instead of calculating 60000000 hairpins in human genome, the 2L chromosome of *Drosophila melanogaster* was used to test the capacity of proposed pre-miRNA detection method even for evolutionary distant species by employing both human data trained izMiR and drosophila trained izMiR model.

The *Drosophila melanogaster* (dme) genome (BDGP6 genome assembly) was fragmented into overlapping (250 nt) fragments of 500 nt length (575,896 fragments) and about 5 million hairpins per strand were extracted. Hairpins with less than 30 nucleotides were filtered leaving around 2 million hairpins per strand. The Chromosome 2L contained about 360,000 hairpins per strand and after removing duplicates, all hairpin features were calculated for this subset of putative pre-miRNAs (about 290,000 per strand). izMiR models generated using human (hsa model) and drosophila (dme model)

hairpins paired with pseudo were used. As a representative for human models AverageDT was used to analyse if the known hairpins for dme could be found.

Human T-lymphotropic virus 1 (NCBI Reference Sequence: NC_001436.1), Human T-lymphotropic virus 2 (NCBI Reference Sequence: NC_001488.1), Human immunodeficiency virus 2 (NCBI Reference Sequence: NC_001722.1), Human immunodeficiency virus 1 (NCBI Reference Sequence: NC_001802.1), Hepatitis B virus (strain ayw) (NCBI Reference Sequence: NC_003977.2) and Human endogenous retrovirus K113 (NCBI Reference Sequence: NC_022518.1) genomes contained 201 total fragments from which after performing filtering based on hairpin length (between 36 and 180; representing the smallest and the longest human stem loops in miRBase) and removing duplicate sequences, 412 hairpins from one strand and 480 hairpins from the other one were extracted. For these hairpins, features were calculated and prediction workflow was used.

The *Solanum lycopersicum* (https://solgenomics.net/organism/1/genome) genome was fragmented into overlapping (250 nt) fragments of 500 nt length (3126673 fragments). Genomes of chloroplast (NCBI Reference Sequence: NC_007898.3) and mitochondria (http://www.ncbi.nlm.nih.gov/assembly/GCA_000325825.1) produced 622 and 2405 fragments, respectively. About 26500000 hairpins per nuclear strand, 6000 hairpins per chloroplast strand, 21000 hairpins per mitochondrial strand were extracted. After removing duplicate sequences and hairpins below 40 or above 500 nucleotides long, around 6200000 hairpins per nuclear strand, 1200 hairpins per chloroplast strand and 4500 hairpins per mitochondrial strand were obtained.

For target prediction analysis, psRNATarget (Dai and Zhao, 2011) tool was used. To generate mature miRNA sequences, hairpins predicted as miRNAs were fragmented into 30 nt long sequences with 15 nt overlaps. Other than existing target pools in psRNATarget (*Homo sapiens* (human), transcript library and *Solanum lycopersicum* (tomato), unigene library) chloroplast and mitochondrial genes of *Solanum lycopersicum* from NCBI were also used.

CHAPTER 3

RESULTS

3.1. Comparison of Available Tools

An overview of the accuracy distribution using 1000 fold MCCV for the three classifiers' combined performance can be seen in Figure 3.1.

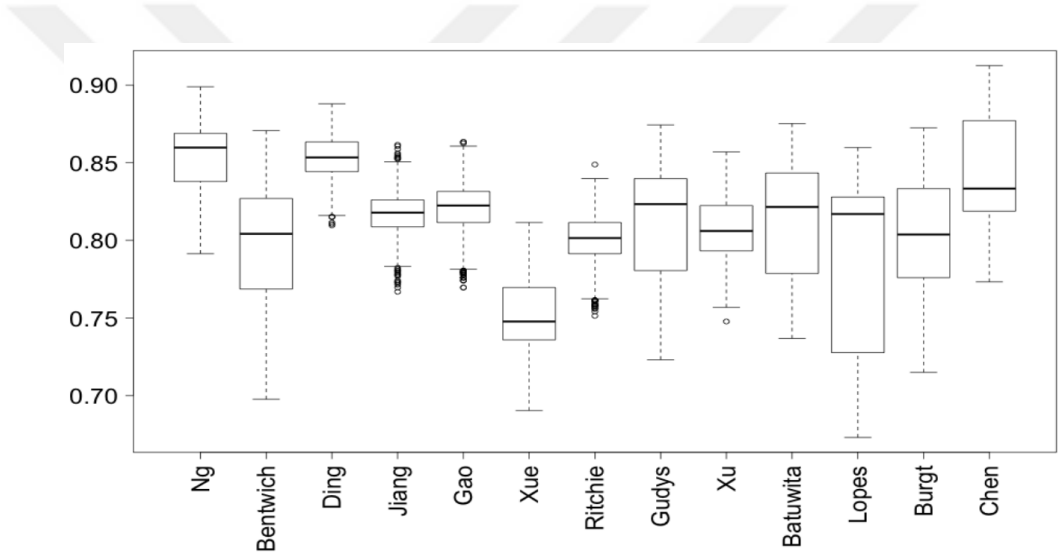

Figure 3.1. The accuracy distributions from three classifiers for each study.

According to Figure 3.1, Chen has the highest accuracy in terms of maximum value but Ng and Ding show a better overall performance since their accuracy distribution is much less data and/or classifier dependent (Figure 3.2). Also, the differences between their maximum values are not very large (ChenSVM: 0.913, NgDT: 0.899, and DingNB: 0.888, Figure 3.2). These results indicate that although all of the models perform better than random guessing, none of them significantly outperforms others. Hence, six consensus models are created by integrating the best models from all studies with equal weights.

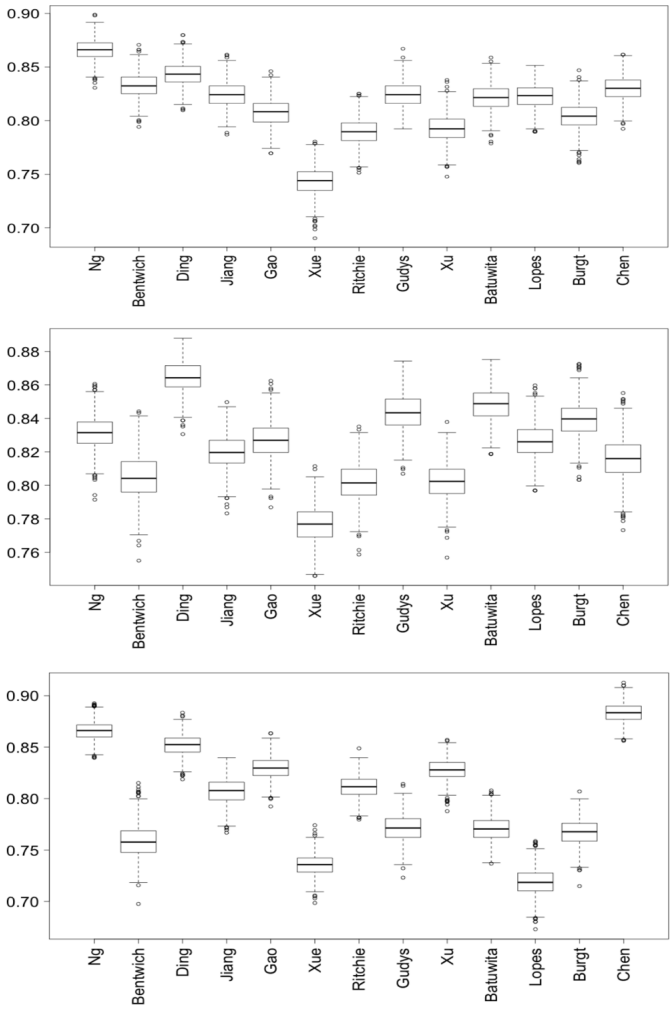

Figure 3.2. The accuracy distributions from each classifier for each study. DT (top), NB (middle), SVM (bottom).

Furthermore, the consensus models are compared to the individual studies by looking at area under curve values and their respective receiver operator characteristic curves (Figure 3.3). The AverageDT model performed best reaching an AUC value of 0.993. While DT models shows some distinctive curve patterns, all of the NB model curves followed similar lines.

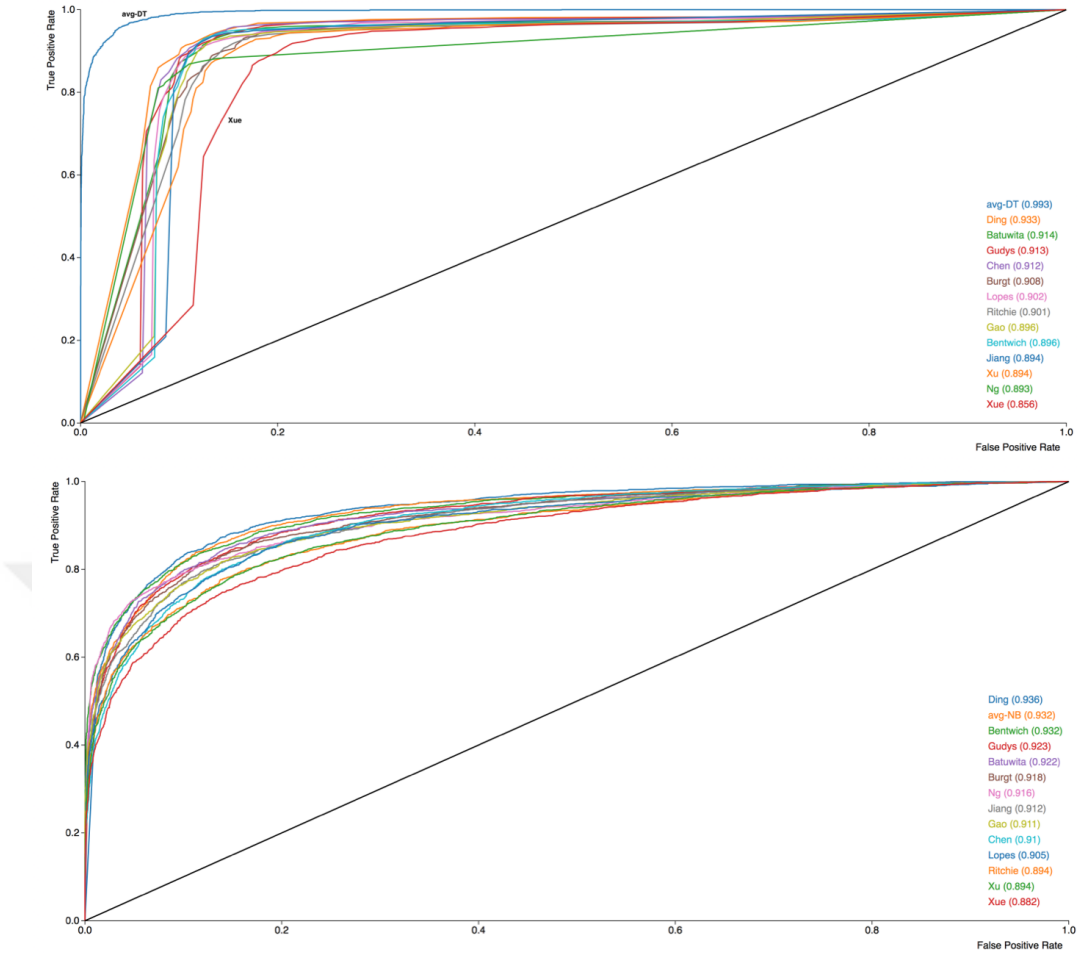

Figure 3.3. ROC graphs for DT and NB. For DT (top), AUC values vary among models but for NB (bottom) it seems like all of the models follow a similar line pattern.

3.2. Prediction Performance

The generated models are trained with human miRNA examples and considering species specific miRNAs it is possible that these models may not perform well for any other species listed in miRBase. In order to test this, the generated models and consensus approaches are used to predict all available pre-miRNAs from all 223 species in miRBase (Figure 3.4). Interestingly, the system works very well, even for plants like *Malus domestica* (mdm, apple) and *Arabidopsis lyrata* (aly). Another intriguing outcome of this analysis is that while TPR values of different models for some organisms are very similar,

e.g. *Malus domestica*, some of them are highly dispersed like *Canis familiaris* (cfa, dog) and *Eptesicus fuscus* (efu, big brown bat). Since this situation is also observed in filtered *Mus musculus* (mmu*) vs. unfiltered (mmu), this scattering pattern is most likely related to data quality.

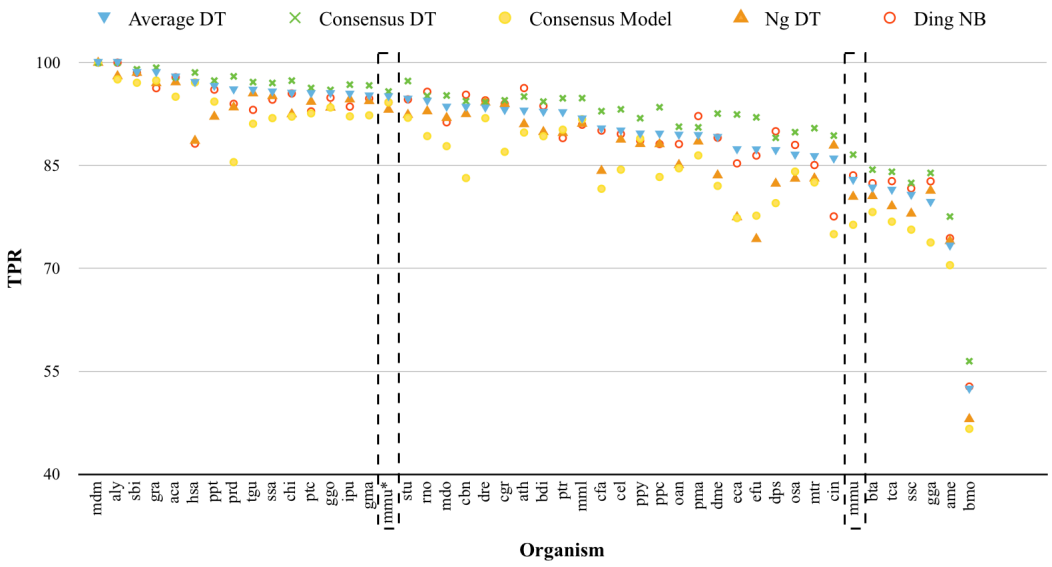

Figure 3.4. TPR of hairpins from different organisms. mmu* stands for filtered mouse hairpins from miRBase based on a minimum RPM value of 100 and mmu shows all mouse hairpins with no filtering. Only organisms with a minimum of 200 hairpins in miRBase are selected for this plot.

In most of the cases only TPR is considered as a performance measurement however finding true negative samples are equally essential. Therefore, TNR of various negative data sets explained in Section 2.1 are analysed (Figure 3.5). In general, models generated by NB classifier performed better than DT models and JiangNB performed best for the majority of all negative data sets, followed by AverageNB and Consensus-Rule. The most challenging data set for most of the models is NotBestFold (Figure 3.5). However since this data set is based on alternative structures of human miRNA sequences it is indeed difficult for sequence based features because the sequences of positive and negative data are the same, while it is comparatively easier for structural features based studies like Jiang to differentiate between data sets. Pseudo data set is used during training so it is expected that all of models will perform well for its prediction. Zou data set is created by using sequences from the coding region sequences (CDSs) (Wei et al., 2013)

so it might actually include real miRNA sequences since miRNAs can be originated from CDSs too (Figure 1.1) and this might be the reason why models showed decreased TNR. Chen data set is combination of samples from pseudo and Zou so its performance as expected placed in between those two data sets' performances. The performance of models on PseudoFR data set might be explained by the way this data set is generated. Since it is built by producing random numbers between minimum and maximum values of each feature in pseudo data set, it is expected that some of these minimum and maximum values are outliers.

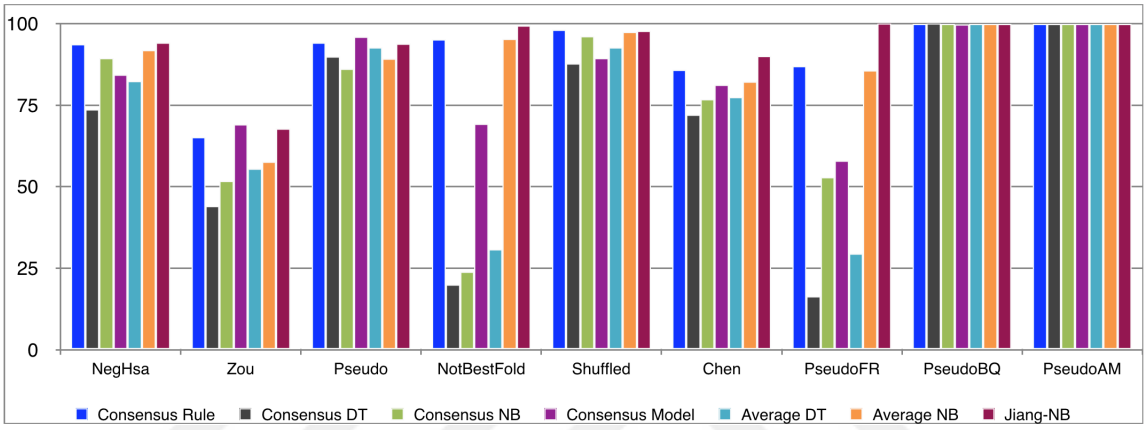

Figure 3.5. TNR of different negative data sets. (For further information about data sets see Section 2.1)

3.3. *Homo sapiens* Analysis

Among 60000000 hairpins remained after filtering in human data set, for around 200000 (3%) of them features are calculated and prediction is performed. Out of 913 (-) strand *Homo sapiens* hairpins listed in miRBase, 9 of them are found (Table 3.1) and only hsa-mir-3910-2 is identified with its suggested length while others are found either shorter or longer than miRBase entries. Predictions of each model are shown in Figure 3.6. When a cutoff value of 0.99 is applied to prediction scores of AverageDT for all calculated hairpins (3% of overall), 662 hairpins are found as miRNAs.

Table 3.1. Identified *Homo sapiens* hairpins. Hpl: length of extracted hairpin, HplM: length of miRBase hairpin, HplMStart: start position of miRBase hairpin in extracted hairpin, HplStart: start position of extracted hairpin in miRBase hairpin.

Accession	Chromosome	Strand	Hpl	HplM	HplMStart	HplStart	AverageDT	AverageNB
hsa-mir-4670	chr9	-	119	75	23		0,991	1,000
hsa-mir-5702	chr2	-	131	84	24		0,988	1,000
hsa-mir-3064	chr17	-	84	66	8		0,467	0,597
hsa-mir-4299	chr11	-	63	72		6	0,966	0,991
hsa-mir-4460	chr5	-	81	86		2	0,991	1,000
hsa-mir-4801	chr4	-	88	82	3		0,990	1,000
hsa-mir-4694	chr11	-	111	80	20		0,990	1,000
hsa-mir-3910-2	chr9	-	82	82	0	0	0,991	1,000
hsa-mir-8066	chr4	-	70	78		4	0,756	0,926

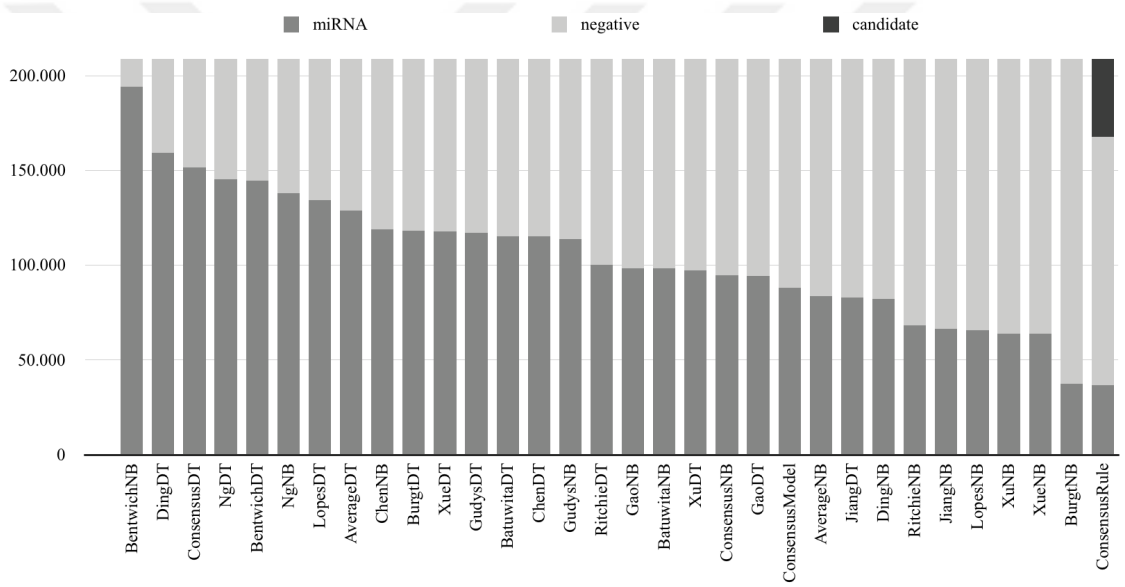

Figure 3.6. Predictions of models on *Homo sapiens* data.

3.4. *Drosophila melanogaster* Analysis

For analysing *Drosophila melanogaster* 2L hairpins, the models generated by two training workflows; one with human data and the other one with *Drosophila melanogaster* data are applied on the 56 2L hairpins from miRBase (Table 3.2).

Table 3.2. Scores for *Drosophila melanogaster* 2L hairpins.

Accession	AverageDT(dme)	AverageDT(hsa)	Accession	AverageDT(dme)	AverageDT(hsa)
dme-mir-4971	0,968	0,991	dme-mir-275	0,898	0,912
dme-mir-2490	0,980	0,991	dme-mir-1	0,984	0,911
dme-mir-2a-2	0,983	0,991	dme-mir-1006	0,978	0,900
dme-mir-2b-2	0,983	0,991	dme-mir-305	0,984	0,892
dme-mir-133	0,983	0,991	dme-mir-962	0,983	0,844
dme-mir-375	0,983	0,991	dme-mir-960	0,983	0,838
dme-mir-932	0,980	0,991	dme-mir-9382	0,966	0,831
dme-mir-2a-1	0,984	0,991	dme-mir-9378	0,919	0,825
dme-let-7	0,980	0,989	dme-mir-79	0,980	0,781
dme-mir-306	0,980	0,988	dme-mir-4972	0,983	0,776
dme-mir-964	0,983	0,988	dme-mir-4973	0,947	0,775
dme-mir-965	0,983	0,988	dme-mir-2489	0,843	0,772
dme-mir-967	0,977	0,987	dme-mir-9c	0,985	0,763
dme-mir-968	0,983	0,987	dme-mir-1002	0,977	0,755
dme-mir-959	0,983	0,986	dme-mir-4912	0,899	0,723
dme-mir-963	0,983	0,985	dme-mir-4910	0,979	0,721
dme-mir-4974	0,983	0,985	dme-mir-125	0,981	0,691
dme-mir-263a	0,979	0,985	dme-mir-287	0,901	0,688
dme-mir-87	0,980	0,978	dme-mir-1005	0,981	0,660
dme-mir-2495	0,976	0,976	dme-mir-2497	0,916	0,641
dme-mir-100	0,984	0,974	dme-mir-4984	0,973	0,637
dme-mir-4946	0,981	0,964	dme-mir-9374	0,978	0,634
dme-mir-966	0,886	0,957	dme-mir-4970	0,787	0,614
dme-mir-124	0,978	0,938	dme-mir-4987	0,947	0,483
dme-mir-9b	0,980	0,934	dme-mir-4943	0,903	0,328
dme-mir-961	0,979	0,914	dme-mir-288	0,974	0,242
dme-mir-2280	0,983	0,912	dme-mir-1004	0,710	0,182
dme-mir-2b-1	0,983	0,912	dme-mir-4914	0,739	0,017

Out of the 56 hairpins from 2L, the genome wide approach is not able to extract 16 of them (Table 3.3). Further analysis on these hairpins revealed that since the genome wide approach searches hairpins in a longer transcript (500 nt), instead of these 16 hairpins their shorter or longer versions with less mfe values are extracted. Moreover, it is possible that if there is a more likely hairpin closer and/or overlapping to these miRBase hairpins, they would be selected and extracted. In dme-mir-4912 case, prediction score is not calculated since the hairpin is 27 nt long which is smaller than defined threshold of minimum 30 nt.

For searching all potential miRNAs in 2L chromosome, the models hsa and dme are applied on all of the extracted hairpins. In order to define a threshold, prediction scores of 40 miRBase hairpins that are extracted from the chromosome (not listed in Table 3.3) are taken into consideration; thresholds are defined as the lower quartile values of the 40 hairpins' prediction scores from AverageDT, as 0.96 for dme model and 0.84 for the hsa model.

Table 3.3. Structures of 16 dme hairpins that are not found by GWA.

miRNA	hairpin fold(extracted fold from 500nt folding)	hsa score
dme-mir-100	)))))..((((((..(((((((..(((..(((((((.(.........).).)))))))..)))..)))))))..))))).....(((((((.(...	0,718
dme-mir-2489	((.....)))))).))).((((((.(((((..(((((((((....)))))))..).)))..)))))))))))))))))))).)))))))....)).	0,870
dme-mir-2495	))))))).....))).)..))))(((((.(((((((((((((.((((.((.....)).)))))))).)).)))))))).))))))))))))))))))..))))))))))	0,390
dme-mir-287	))).)))))))))...)))))))).)..)))))))..(((((..((((((((((((...))))))))))))..))))))..)))))))).))))))	0,870
dme-mir-4912	)))....))))))))))..((((((((((........)))..))))))))))))).((((((	na
dme-mir-4914	.((.((.....(((((..(((...)))..)))))..))))))))))).))))....(((((.((..((((........)))..)))))))...)))))	0,470
dme-mir-4973	))))..........((((((.((((((((((..(((.......(((((((....)))..))))).....))).........((((((.(((((.((....)).))))))).))))))).))))))))))).))))))((.((((((((((((...(	0,540
dme-mir-9382	(((((((((..((((.((((((.....................))))))))))..)))))))))).	0,410
dme-mir-4943	....)....)))))))))..))..)))((((((((((((..(.((((((((.(.....).)))..)))))))).).))))))))))).....((((((((((.((.((((	0,810
dme-mir-288	(((((.((((((..((.((((.((((((...((((((((((....)))).....)))))))..)))))))..))))).))....)))))).)))))	0,465
dme-mir-4910	..((...(((((((..(((((.((((..(((.....)))..)).)))))))..)))))))...)	0,140
dme-mir-4970	.....((.(((..((..((((.(((((((((.(((.(((.((((.....................)))).))).))))..))))))))).))))..)).)))....))....	0,243
dme-mir-964	(((((...)))))))))))))((.((((((((((((...(((((((.............)))))))))...))))))))))))).)).((((((...)))	0,990
dme-mir-9b	.))))))).....((((((((((...((((.(((((.((.....)).)))))).))))..))))))))).((((((((((..((.((.	0,940
dme-mir-4984	(((.....))))..)))...((((((((((((((((((.....(((((((.....))))))).)))))))))))))))))))))....))))))((((....))))	0,930
dme-mir-1004	((((((.((((((((....((((((..(.(((((.....)))))..)))))).....(((((((	0,190

Based on the threshold scores 25 of 2L miRNAs listed in miRBase are also found in the extracted hairpin pool in both models' predictions while in the hsa model dme-mir-275, dme-mir-9378, dme-mir-1006, dme-mir-966, dme-mir-967 hairpins passed the threshold whereas the dme model identified dme-mir-125, dme-mir-275, dme-mir-9374, dme-mir-960, dme-mir-962, dme-mir-9c. Overall 17455 candidate miRNAs with the dme model and 43103 candidates with the hsa model are predicted out of 581,883 hairpins in the 2L chromosome (Figure 3.7).

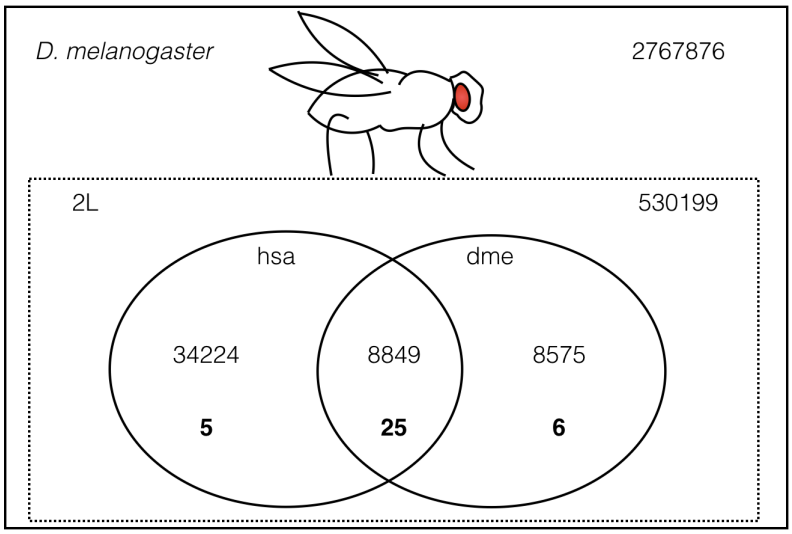

Figure 3.7. Predicted miRNAs in 2L chromosome of *Drosophila melanogaster*. Circles include number of hairpins passing confidence thresholds in hsa and dme models. Numbers in bold characters indicate miRBase hairpins extracted and passed the thresholds.

3.5. *Solanum lycopersicum* Analysis

Analysis of *Solanum lycopersicum* nuclear, mitochondrial and chloroplast genomes showed that all of these genomes could produce hairpin structured transcripts. All of the hairpins are filtered firstly by their length (as minimum 78 and maximum 240) and then based on the average of prediction scores of 5 classifiers (minimum average as 0.995). This filtration process results in 63516 unique hairpins from nuclear genome, 7 unique hairpins from chloroplast and 36 unique hairpins from mitochondria (Figure 3.8).

In order to find candidate mature sequences, the model with the highest accuracy score (0.935) is applied to predicted miRNA hairpins after fragmenting them to 24 nt long sequences with 6 nt overlaps. Mature candidates located in one of the arms and covering terminal loop are used for further analysis.

In the next step, we searched for interactions among source of miRNAs and their targets. For this purpose, miRNAs originated from each genome are tested to see whether they have targets in other genomes as well as in their own genome (Figure 3.8).

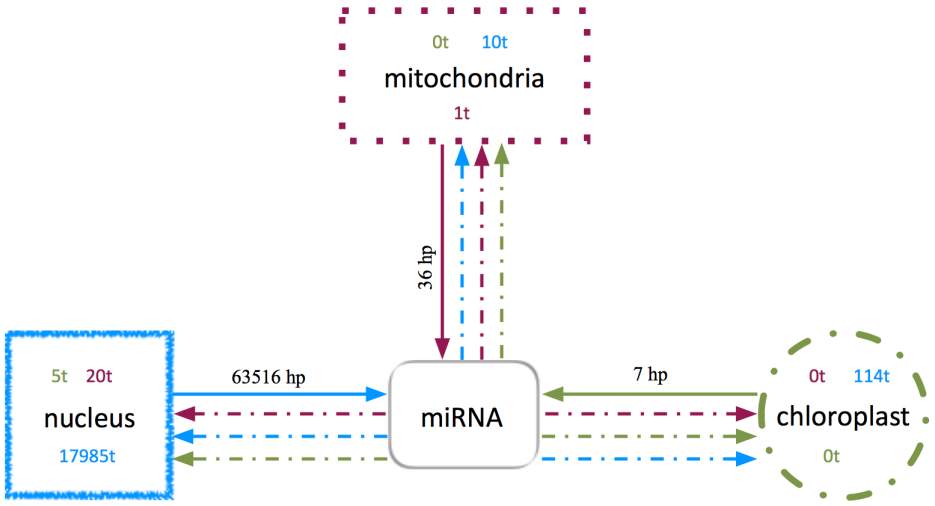

Figure 3.8. *Solanum lycopersicum* miRNA network. Colours indicate the source of miRNAs; blue nuclear, red mitochondrial, green chloroplast. #t indicates the number of predicted targets for miRNAs (origin colour coded) found in the genome.

As it can bee seen from Figure 3.8, some miRNAs that have targets in organelle genomes have been identified. For instance, some of the nuclear miRNAs seem to be

targeting chloroplast tRNA genes and ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (rbcL) gene which is translated in chloroplast. Also, some of mitochondrial targets of nuclear miRNAs include tRNA genes, mitochondrial gene for ABC-type heme transporter subunit (gi:1304246), mitochondrial rpl10 gene for ribosomal protein L10 (gi:304376250), mitochondrial gene for 18S ribosomal RNA (gi:658152043) and mitochondrial atp9 gene for ATPase subunit 9 (gi:13077). The only identified target of mitochondria originated miRNAs in mitochondria genes is 18S ribosomal RNA gene (gi:658152043)

For an organelle genome whether it is mitochondria or chloroplast, to produce miRNAs would mean that either transportation of miRNA machinery (Figure 1.2) through the membranes of organelles or export of pre-miRNA to cytosol. In the case of rbcL targeting, the nuclear originated miRNA should be transported into chloroplast. How such transportation occurs is still not a fully answered question not only for plants but also for human mitochondrial miRNAs (Shinde and Bhadra, 2015).

3.6. Virus Analysis

The possible miRNAs of retroviruses were analyzed by using the described genome wide approach (Figure 2.2) and applying the obtained data to izMiR workflows. At the end, out of 38 hairpins (3 from Human endogenous retrovirus K113 (HERK113), 3 from Hepatitis B virus (strain ayw) (HB), 10 from Human T-lymphotropic virus 1 (HTLV1), 10 from Human T-lymphotropic virus 2 (HTLV2), 3 from Human immunodeficiency virus 2 (HIV2) and 9 from Human immunodeficiency virus 1 (HIV1)) had prediction scores above 0.90 for AverageDT. One of the 9 predicted hairpins of Human immunodeficiency virus 1 matched hiv1-mir-TAR. From these 38 hairpin sequences, 52 mature sequences were extracted and tested to see if they could target human genes. At the end, 26 mature sequences seemed to have capacity to target 79 genes in human transcriptome (Table 3.4).

In a similar manner, known human miRNA mature sequences from miRBase were used to see if they could also target viral genes (Table 3.5). According to information presented in Table 3.5, none of the 2588 mature sequences of human would not target genes of HIV2 and HERK113.

Lastly, we did also checked if viral miRNAs could target their own viral genes as well. However, none of the 52 mature sequences targeted their own genes.

Table 3.4. List of human genes that might be targeted by viral miRNAs.

Virus	Target
HB	RALGPS2, DZIP1, ANGPT1, MTHFSD
HIV2	SENP8, SART3, PTBP3, NAAA, FCRL1, FCRL1, BATF2, KIF3A
HERK113	ZNF592, SMCR8, SMC2, SNRPB2, ANGEL1, ANGEL1, PURA, PURA, PERP
HTLV1	RAD18, ARHGEF7, ERAP2, AGPAT5, PTPN4, FAM114A1, SERINC5, XIAP
HTLV2	PSMA5, OTUD4, LCA5, MKLN1, KIAA1549L, ZBTB6, TMEM220, ZNF365, NEO1, PRKAA2, FOXI1, HBP1, ZNF212, ITPRIP, MME, G6PC, ACP2, MST1L, DHX8, CDV3, EIF4B
HIV1	TNPO3, UTP14C, FBXO21, CFHR3, SGOL1, L3MBTL4, MEDAG, HTRA2, PALM2-AKAP2, APAF1, SH3BP2, LMTK2, ERN1, INPP4A, METTL7A, ZNF483, CLDN10, ADAMTS5, ZNF100, PAPPA, CRISP3, GPHB5, MED21, ARNT2, POLD2

Table 3.5. List of viral genes that might be targeted by human miRNAs.

mirNA	Virus	Target Description
hsa-mir-3960	HTLV1	gene=HTLV1gp1, protein=Pr gag-pro-pol, location=join(450..1718,1718..2245,2245..4836)
hsa-mir-8077	HTLV1	gene=HTLV1gp1, protein=Pr gag-pro-pol, location=join(450..1718,1718..2245,2245..4836)
hsa-mir-3960	HTLV1	gene=HTLV1gp2, protein=Pr gag-pro, location=join(450..1718,1718..2404)
hsa-mir-8077	HTLV1	gene=HTLV1gp2, protein=Pr gag-pro, location=join(450..1718,1718..2404)
hsa-mir-3960	HTLV1	gene=gag, protein=Pr55, location=450..1739
hsa-mir-8077	HTLV1	gene=gag, protein=Pr55, location=450..1739
hsa-mir-6802-5p	HTLV1	gene=env, protein=gp46 SU, location=4829..6295
hsa-mir-6752-5p	HTLV2	gene=HTLV2gs1, protein=pol polyprotein, partial=5', location=<2239..5187
hsa-mir-2116-5p	HTLV2	gene=HTLV2gp6, protein=hypothetical protein, location=5180..6640
hsa-mir-6779-5p	HTLV2	gene=HTLV2gp6, protein=hypothetical protein, location=5180..6640
hsa-mir-4796-5p	HIV1	gene=vif, protein=Vif, location=4587..5165
hsa-mir-6873-3p	HIV1	gene=rev, protein=Rev, location=join(5516..5591,7925..8199)
hsa-mir-1913	HIV1	gene=env, protein=Envelope surface glycoprotein gp160, precursor, location=5771..8341
hsa-mir-6873-3p	HIV1	gene=env, protein=Envelope surface glycoprotein gp160, precursor, location=5771..8341
hsa-mir-557	HB	gene=P, protein=polymerase, location=join(2309..3182,1..1625)
hsa-mir-511-3p	HB	gene=P, protein=polymerase, location=join(2309..3182,1..1625)
hsa-mir-5193	HB	gene=X, protein=X protein, location=1376..1840
hsa-mir-511-3p	HB	gene=S, protein=large envelope protein, location=join(2850..3182,1..837)
hsa-mir-511-3p	HB	gene=S, protein=middle envelope protein, location=join(3174..3182,1..837)
hsa-mir-511-3p	HB	gene=S, protein=small envelope protein, location=157..837

CHAPTER 4

CONCLUSION

Early estimates indicated that there are about 11 million potential hairpins in the human genome (Bentwich et al., 2005) and these hairpins might come from any part of the genome (Lindow and Gorodkin, 2007). However, the results presented in this work show that the actual number of hairpins that can be transcribed from the human genome is much higher than anticipated ($>$ 60 million). Combined with the fact that there are hundreds of features defining pre-miRNAs, analysing such huge data sets becomes a challenging issue. Furthermore hairpin structure is not limited to miRNAs, meaning that identifying the hairpins which would become functional miRNAs is a very essential task requiring highly accurate systems.

One of the main aims of this study is creating an integrative data mining platform which would provide all the steps to accomplish an efficient classification process for searching candidate pre-miRNAs in a given data set. To achieve this, available studies in the area are analysed (Table 1.3) and their individual performances and the new consensus approaches are compared (Figure 3.1, Figure 3.2 and Figure 3.3). Overall, the presented framework includes 13 individual studies (feature groups) with two classifier models and six consensus methods. Although some of the studies have higher accuracy values none significantly outperforms others. For the consensus models, equal weights are given to all individual models even though they do not perform equally well. Also, while selecting the best model among 1000 iterations, the highest accuracy scores are taken into consideration, which may not be the best way. However, the workflows are available for users and other scores like the F-measure and the Youden index are also calculated for each model, so if desired alternative model selection strategies can be performed.

The influence of data quality is also assessed by applying the generated models on various positive and negative data sets. Based on the pre-miRNAs of 223 species in miRBase, TPR scores for all consensus models and NgDT as the representative of studies are presented (Table A). Although, the system achieves high scores even for organisms with big evolutionary distance to human (learning data), for mouse obtained results did not seem quite well. This situation is further analysed by filtering miRBase mouse hair-

pins through setting RPM value as minimum 100 (mmu*) and the TPR score of this data set is much closer to expected range (Figure 3.4). Based on the information in Figure 3.5 the most challenging negative data sets for models to solve are Zou, NotBestFold, and pseudoFR data sets.

GWA reveals that various organisms genomes have the capacity to produce certain number of pre-miRNAs. *Homo sapiens* analysis indicates that human genome has a higher capacity of pre-miRNAs than estimated. Moreover, *Drosophila melanogaster* analysis show that even though the models trained with human miRNA hairpins make predictions with high TPR scores, the models specially created by *Drosophila melanogaster* miRNAs reach to higher scores (Table 3.2). An interesting finding of this analysis is that some of the miRNAs in miRBase are not extracted from the genomes since their structures would not create proper hairpins for Drosha and Dicer and/or they have competing neighbours with better structures (Table 3.3).

To our knowledge this is the first study performing a genome wide miRNA search in chloroplast and mitochondria and looking for communications between three genomes in *Solanum lycopersicum* cells. Network analysis reveals that there are potential miRNAs that might be originated from these genomes and target mRNAs of others as well as their own (Figure 3.8). Although these results require experimental validation, they suggest that among various communication mechanisms described so far, miRNAs might be another player in the field in addition to their post-transcriptional regulation functions.

Analysis of potential miRNAs produced by retroviruses and their effects on human gene regulation is an interesting research area. In this study we showed that a form of miRNA based cross-kingdom regulation through viruses and their host could be possible. In the future, experimental validation of the provided results would help increasing our understanding about the issue.

In conclusion, the work explained in this study provides the comparison of existing *ab initio* pre-miRNA prediction tools, enables the opportunity to use the workflows designed, offers various data sets and shows that the method is applicable to many species, even to eukaryotes with large genomes.

REFERENCES

Baker, M. (2010, sep). MicroRNA profiling: separating signal from noise. *Nature methods 7*(9), 687–92.

Barrey, E., G. Saint-Auret, B. Bonnamy, D. Damas, O. Boyer, and X. Gidrol (2011). Pre-microRNA and mature microRNA in human mitochondria. *PLoS ONE 6*(5), e20220.

Batuwita, R. and V. Palade (2009, apr). microPred: effective classification of pre-miRNAs for human miRNA gene prediction. *Bioinformatics (Oxford, England) 25*(8), 989–95.

Bentwich, I. (2008, jan). Identifying human microRNAs. *Current topics in microbiology and immunology 320*, 257–69.

Bentwich, I., A. Avniel, Y. Karov, R. Aharonov, S. Gilad, O. Barad, A. Barzilai, P. Einat, U. Einav, E. Meiri, E. Sharon, Y. Spector, and Z. Bentwich (2005, jul). Identification of hundreds of conserved and nonconserved human microRNAs. *Nature genetics 37*(7), 766–70.

Berezikov, E., E. Cuppen, and R. H. A. Plasterk (2006, may). Approaches to microRNA discovery. *Nature genetics 38 Suppl*(May), S2–S7.

Berthold, M. R., N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, K. Thiel, and B. Wiswedel (2009, November). Knime - the konstanz information miner: Version 2.0 and beyond. *SIGKDD Explor. Newsl. 11*(1), 26–31.

Brameier, M. and C. Wiuf (2007, jan). Ab initio identification of human microRNAs based on structure motifs. *BMC bioinformatics 8*, 478.

Bushati, N. and S. M. Cohen (2007, jan). microRNA functions. *Annual review of cell and developmental biology 23*, 175–205.

Chapman, E. J. and J. C. Carrington (2007, nov). Specialization and evolution of endoge-

nous small RNA pathways. *Nat Rev Genet 8*(11), 884–896.

Chen, J., X. Wang, and B. Liu (2016, jan). iMiRNA-SSF: Improving the Identification of MicroRNA Precursors by Combining Negative Sets with Different Distributions. *Scientific reports 6*, 19062.

Chou, C. H., N. W. Chang, S. Shrestha, S. D. Hsu, Y. L. Lin, W. H. Lee, C. D. Yang, H. C. Hong, T. Y. Wei, S. J. Tu, T. R. Tsai, S. Y. Ho, T. Y. Jian, H. Y. Wu, P. R. Chen, N. C. Lin, H. T. Huang, T. L. Yang, C. Y. Pai, C. S. Tai, W. L. Chen, C. Y. Huang, C. C. Liu, S. L. Weng, K. W. Liao, W. L. Hsu, and H. D. Huang (2016, jan). miRTarBase 2016: Updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Research 44*(D1), D239–D247.

Chtioui, Y., D. Bertrand, and D. Barba (1998). Feature selection by a genetic algorithm. Application to seed discrimination by artificial vision. *Journal of the Science of Food and Agriculture 76*(1), 77–86.

Chugh, P. and D. P. Dittmer (2012, sep). Potential pitfalls in microRNA profiling. *Wiley Interdisciplinary Reviews: RNA 3*(5), 601–616.

Dai, X. and P. X. Zhao (2011, jul). PsRNATarget: A plant small RNA target analysis server. *Nucleic Acids Research 39*(SUPPL. 2), W155–9.

Dasgupta, N., Y. Peng, Z. Tan, G. Ciraolo, D. Wang, and R. Li (2015, jul). miRNAs in mtDNA-less cell mitochondria. *Cell Death Discovery 1*(June), 15004.

Dash, M. and H. Liu (1997). Feature selection for classification. *Intelligent Data Analysis 1*(3), 131–156.

Ding, J., S. Zhou, and J. Guan (2010, jan). MiRenSVM: towards better prediction of microRNA precursors using an ensemble SVM classifier with multi-loop features. *BMC bioinformatics 11 Suppl 1*(Suppl 11), S11.

Dong, H., J. Lei, L. Ding, Y. Wen, H. Ju, and X. Zhang (2013, aug). MicroRNA: function, detection, and bioanalysis. *Chemical reviews 113*(8), 6207–33.

Filipowicz, W., S. N. Bhattacharyya, and N. Sonenberg (2008, feb). Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nature reviews. Genetics 9*(2), 102–14.

Gao, D., R. Middleton, J. E. J. Rasko, and W. Ritchie (2013, dec). miREval 2.0: a web tool for simple microRNA prediction in genome sequences. *Bioinformatics (Oxford, England) 29*(24), 3225–6.

Gottwein, E. and B. R. Cullen (2008, 06). Viral and cellular micrornas as determinants of viral pathogenesis and immunity. *Cell host & microbe 3*(6), 375–387.

Grundhoff, A. and C. S. Sullivan (2011, mar). Virus-encoded microRNAs. *Virology 411*(2), 325–343.

Gudyś, A., M. W. Szcześniak, M. Sikora, and I. Makałowska (2013). Huntmi: an efficient and taxon-specific approach in pre-mirna identification. *BMC Bioinformatics 14*(1), 83.

Guyon, I. and A. Elisseeff (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research (JMLR) 3*(3), 1157–1182.

Hébert, S. S., K. Horré, L. Nicolaï, B. Bergmans, A. S. Papadopoulou, A. Delacourte, and B. De Strooper (2009, mar). MicroRNA regulation of Alzheimer's Amyloid precursor protein expression. *Neurobiology of disease 33*(3), 422–8.

Heikkinen, L., M. Kolehmainen, and G. Wong (2011, may). Prediction of microRNA targets in Caenorhabditis elegans using a self-organizing map. *Bioinformatics (Oxford, England) 27*(9), 1247–54.

Hofacker, I. L. (2003, jul). Vienna RNA secondary structure server. *Nucleic Acids Research 31*(13), 3429–3431.

Hutvágner, G., J. McLachlan, A. E. Pasquinelli, E. Bálint, T. Tuschl, and P. D. Zamore (2001, aug). A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science (New York, N.Y.) 293*(5531), 834–8.

Ibáñez-Ventoso, C., M. Vora, and M. Driscoll (2008, jan). Sequence relationships among C. elegans, D. melanogaster and human microRNAs highlight the extensive conservation of microRNAs in biology. *PloS one 3*(7), e2818.

Janecek, A., W. N. W. Gansterer, M. Demel, and G. Ecker (2008). On the Relationship Between Feature Selection and Classification Accuracy. *Fsdm 4*, 90–105.

Jiang, P., H. Wu, W. Wang, W. Ma, X. Sun, and Z. Lu (2007, jul). MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic acids research 35*(Web Server issue), W339–44.

Khvorova, A., A. Reynolds, and S. D. Jayasena (2003, oct). Functional siRNAs and miRNAs exhibit strand bias. *Cell 115*(2), 209–16.

Kim, V. N., J. Han, and M. C. Siomi (2009, feb). Biogenesis of small RNAs in animals. *Nature reviews. Molecular cell biology 10*(2), 126–39.

Kohavi, R. (1995, aug). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence 14*(12), 1137–1143.

Kohavi, R. and G. H. John (1997). Wrappers for feature subset selection. *Artificial Intelligence 97*(1), 273 – 324.

Kozomara, A. and S. Griffiths-Jones (2011, jan). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic acids research 39*(Database issue), D152–7.

Larranaga, P. (2006, feb). Machine learning in bioinformatics. *Briefings in Bioinformatics 7*(1), 86–112.

Lee, R. C., R. L. Feinbaum, and V. Ambros (1993, dec). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell 75*(5), 843–54.

Lee, Y., C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Rådmark, S. Kim,

and V. N. Kim (2003, sep). The nuclear RNase III Drosha initiates microRNA processing. *Nature 425*(6956), 415–9.

Lee, Y., K. Jeon, J.-T. Lee, S. Kim, and V. N. Kim (2002, sep). MicroRNA maturation: stepwise processing and subcellular localization. *The EMBO journal 21*(17), 4663–70.

Lee, Y., M. Kim, J. Han, K.-H. Yeom, S. Lee, S. H. Baek, and V. N. Kim (2004, oct). MicroRNA genes are transcribed by RNA polymerase II. *The EMBO journal 23*(20), 4051–60.

Li, Y., C. Qiu, J. Tu, B. Geng, J. Yang, T. Jiang, and Q. Cui (2014, jan). HMDD v2.0: A database for experimentally supported human microRNA and disease associations. *Nucleic Acids Research 42*(D1), D1070–4.

Lindow, M. and J. Gorodkin (2007, may). Principles and limitations of computational microRNA gene and target finding. *DNA and cell biology 26*(5), 339–51.

Litzkow, M., M. Livny, and M. Mutka (1988). Condor-a hunter of idle workstations. In *[1988] Proceedings. The 8th International Conference on Distributed*, pp. 104–111. IEEE Comput. Soc. Press.

Lopes, I. d. O., A. Schliep, and A. C. d. L. de Carvalho (2014, jan). The discriminant power of RNA features for pre-miRNA recognition. *BMC bioinformatics 15*(1), 124.

Millar, A. A. and P. M. Waterhouse (2005, jul). Plant and animal microRNAs: similarities and differences. *Functional & integrative genomics 5*(3), 129–35.

Ng, K. L. S. and S. K. Mishra (2007, jun). De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics (Oxford, England) 23*(11), 1321–30.

Ouellet, D. L., I. Plante, P. Landry, C. Barat, M. È. Janelle, L. Flamand, M. J. Tremblay, and P. Provost (2008, apr). Identification of functional microRNAs released through asymmetrical processing of HIV-1 TAR element. *Nucleic Acids Research 36*(7),

2353–2365.

Piast, M., I. Kustrzeba-Wójcicka, M. Matusiewicz, and T. Banaś (2005). Molecular evolution of enolase. *Acta Biochimica Polonica 52*(2), 507–513.

Powell, W. B. (2011). *The Challenges of Dynamic Programming*, pp. 1–23. John Wiley and Sons, Inc.

Reinhart, B. J., F. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, H. R. Horvitz, and G. Ruvkun (2000, feb). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature 403*(6772), 901–6.

Saçar, M. D. and J. Allmer (2013, sep). Data mining for microrna gene prediction: On the impact of class imbalance and feature number for microrna gene prediction. In *2013 8th International Symposium on Health Informatics and Bioinformatics*, pp. 1–6. IEEE.

Saçar, M. D., C. Bağcı, and J. Allmer (2014, oct). Computational Prediction of MicroRNAs from Toxoplasma gondii Potentially Regulating the Hosts' Gene Expression. *Genomics, proteomics & bioinformatics 12*(5), 228–238.

Saçar, M. D., H. Hamzeiy, and J. Allmer (2013). Can MiRBase provide positive data for machine learning for the detection of MiRNA hairpins? *Journal of integrative bioinformatics 10*, 215.

Saeys, Y., I. Inza, and P. Larrañaga (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics 23*(19), 2507–2517.

Shahamat, H. and A. A. Pouyan (2015). Feature selection using genetic algorithm for classification of schizophrenia using fMRI data. *Journal of Artificial Intelligence and Data Mining 3*(1), 30–37.

Shinde, S. and U. Bhadra (2015). A complex genome-MicroRNA interplay in human mitochondria. *BioMed Research International 2015*, 206382.

Skalsky, R. L. and B. R. Cullen (2010). Viruses, microRNAs, and host interactions. *Annual review of microbiology 64*(1), 123–141.

Steffen, P., B. Voss, M. Rehmsmeier, J. Reeder, and R. Giegerich (2006, feb). RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics (Oxford, England) 22*(4), 500–3.

Vafaie, H. and K. D. Jong (1992). Genetic Algorithms as a Tool for Feature Selection in Machine Learning. *Proceedings Fourth International Conference on Tools with Artificial Intelligence TAI '92 6*(July), 267–281.

van der Burgt, A., M. W. J. E. Fiers, J.-P. Nap, and R. C. H. J. van Ham (2009, jan). In silico miRNA prediction in metazoan genomes: balancing between sensitivity and specificity. *BMC genomics 10*, 204.

Varma, S. and R. Simon (2006, jan). Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics 7*, 91.

Wang, G., J. M. van der Walt, G. Mayhew, Y. J. Li, S. Züchner, W. K. Scott, E. R. Martin, and J. M. Vance (2008, feb). Variation in the miRNA-433 Binding Site of FGF20 Confers Risk for Parkinson Disease by Overexpression of Alpha-Synuclein. *American Journal of Human Genetics 82*(2), 283–289.

Wang, L., X. Yu, H. Wang, Y.-Z. Lu, M. de Ruiter, M. Prins, and Y.-K. He (2011, jun). A novel class of heat-responsive small RNAs derived from the chloroplast genome of Chinese cabbage (Brassica rapa). *BMC genomics 12*(1), 289.

Wei, L., M. Liao, Y. Gao, R. Ji, Z. He, and Q. Zou (2013, nov). Improved and Promising Identification of Human MicroRNAs by Incorporating a High-quality Negative Set. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM 11*(1), 192–201.

Wu, Y., B. Wei, H. Liu, T. Li, and S. Rayner (2011, jan). MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences. *BMC bioinformatics 12*(1), 107.

Xu, G., Y. Zhang, H. Jia, J. Li, X. Liu, J. F. Engelhardt, and Y. Wang (2009, dec). Cloning and identification of microRNAs in bovine alveolar macrophages. *Molecular and cellular biochemistry 332*(1-2), 9–16.

Xuan, P., M. Z. Guo, J. Wang, C. Y. Wang, X. Y. Liu, and Y. Liu (2011). Genetic algorithm-based efficient feature selection for classification of pre-miRNAs. *Genetics and Molecular Research 10*(2), 588–603.

Xue, C., F. Li, T. He, G.-P. Liu, Y. Li, and X. Zhang (2005, jan). Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC bioinformatics 6*, 310.

Yones, C. A., G. Stegmayer, L. Kamenetzky, and D. H. Milone (2015, dec). miRNAfe: A comprehensive tool for feature extraction in microRNA prediction. *Bio Systems 138*, 1–5.

Zafari, S., C. Backes, P. Leidinger, E. Meese, and A. Keller (2015). Regulatory MicroRNA Networks: Complex Patterns of Target Pathways for Disease-related and Housekeeping MicroRNAs. *Genomics, Proteomics and Bioinformatics 13*(3), 159–168.

Zeng, Y. and B. R. Cullen (2004, jan). Structural requirements for pre-microRNA binding and nuclear export by Exportin 5. *Nucleic acids research 32*(16), 4776–85.

Zhang, B.-T. and J.-W. Nam (2008). Supervised Learning Methods for MicroRNA Studies. In *Machine Learning in Bioinformatics*, pp. 339–365. John Wiley & Sons, Inc.

Zhang, H., F. A. Kolb, V. Brondani, E. Billy, and W. Filipowicz (2002, nov). Human Dicer preferentially cleaves dsRNAs at their termini without a requirement for ATP. *The EMBO journal 21*(21), 5875–85.

Zhang, X., X. Zuo, B. Yang, Z. Li, Y. Xue, Y. Zhou, J. Huang, X. Zhao, J. Zhou, Y. Yan, H. Zhang, P. Guo, H. Sun, L. Guo, Y. Zhang, and X. D. Fu (2014, jul). MicroRNA directly enhances mitochondrial translation during muscle differentiation. *Cell 158*(3), 607–619.

APPENDIX A

PERFORMANCE ON MIRBASE DATA

Table A.1. TPR scores for organisms in miRBase. hp: number of hairpins, CR: ConsensusRule, CDT: ConsensusDT, CNB: ConsensusNB, CR: ConsensusRule, AvgDT: AverageDT, AvgNB: AverageNB. Table sorted from the highest hp to the lowest.

Acronym	hp	CR	CDT	CNB	CM	AvgDT	AvgNB	NgDT
hsa	1881	76,40	98,56	86,12	97,13	97,08	82,62	88,62
mmu	1193	59,09	86,59	82,23	76,36	82,82	77,03	80,47
mmu*	380	83,95	95,79	93,42	94,21	95,00	91,05	93,16
bta	808	65,72	84,41	80,45	78,22	81,68	77,60	80,57
gga	740	54,86	83,92	77,43	73,78	79,59	72,84	81,35
eca	715	56,36	92,45	83,50	77,34	87,27	80,14	77,48
mtr	670	73,73	90,45	84,48	82,54	86,27	82,99	83,13
ptr	655	78,32	94,81	88,24	90,23	92,67	86,72	89,77
ppy	642	76,95	91,90	86,60	88,94	89,56	84,58	88,16
mml	619	80,29	94,83	88,85	91,76	91,76	86,75	91,11
osa	592	74,49	89,86	86,49	84,12	86,49	85,47	83,11
gma	573	84,82	96,68	94,76	92,32	95,11	93,02	94,42
efu	502	50,40	92,03	80,68	77,69	87,25	77,29	74,30
cfa	495	65,66	92,93	89,09	81,62	90,30	86,06	84,24
rno	495	77,78	95,15	93,54	89,29	94,34	91,72	92,93
bmo	487	35,32	56,47	54,41	46,61	52,36	51,33	48,05
mdo	460	76,30	95,22	91,52	87,83	93,48	89,35	91,96
oan	396	66,41	90,66	85,35	84,60	89,39	83,08	85,10
ssc	382	65,71	82,46	78,01	75,65	80,63	74,61	78,01
ssa	371	81,94	97,04	95,15	91,91	95,69	92,18	95,15
ppc	354	68,64	93,50	89,83	83,33	89,55	85,31	88,14
ggo	352	77,27	96,02	90,63	93,47	95,45	88,35	93,47

(cont. on next page)

Table A.1 (cont.).

Acronym	hp	CR	CDT	CNB	CM	AvgDT	AvgNB	NgDT
ptc	352	82,67	96,31	92,90	92,61	95,45	91,76	94,32
cin	348	59,48	89,37	81,03	75,00	85,92	77,30	87,93
dre	346	83,82	94,22	94,22	91,91	93,35	93,06	94,22
ath	324	77,78	95,06	93,83	89,81	92,90	92,59	91,05
bdi	317	77,60	94,32	91,80	89,27	92,74	90,22	89,91
aca	282	90,78	97,87	97,16	95,04	97,87	96,45	97,16
ipu	281	82,92	96,80	92,17	92,17	95,37	90,39	94,66
gra	269	85,50	99,26	94,80	97,40	98,51	92,57	97,03
chi	267	77,15	97,38	91,01	92,13	95,51	88,76	92,51
dme	256	64,84	92,58	87,11	82,03	89,06	82,42	83,59
ame	254	56,69	77,56	75,20	70,47	73,23	71,26	74,02
cel	250	75,20	93,20	89,20	84,40	90,00	87,60	88,80
tgu	247	77,73	97,17	92,71	91,09	95,95	88,66	95,55
pma	244	80,33	90,57	91,39	86,48	89,34	89,75	88,52
ppt	229	80,79	97,38	95,63	94,32	96,51	95,63	92,14
stu	224	86,61	97,32	95,09	91,96	94,64	94,20	92,41
tca	220	66,36	84,09	81,36	76,82	81,36	79,55	79,09
cbn	214	78,04	94,39	96,26	83,18	93,46	94,86	92,52
dps	210	64,29	89,05	89,05	79,52	87,14	84,29	82,38
mdm	206	96,60	100,00	99,51	100,00	100,00	99,51	100,00
aly	205	88,29	100,00	99,51	97,56	100,00	99,51	98,05
sbi	205	89,76	99,02	97,07	97,07	98,54	96,59	98,54
cgr	200	81,00	94,50	92,50	87,00	93,00	91,00	94,00
prd	200	76,50	98,00	91,50	85,50	96,00	90,00	93,50
oha	198	84,34	97,47	94,95	93,94	95,96	91,41	95,96
xtr	192	88,54	98,96	97,40	95,83	97,92	96,35	96,88
hco	188	73,40	89,89	84,04	85,64	88,83	81,91	88,30
ppe	180	90,00	97,22	96,67	96,67	96,67	95,56	96,67
tch	177	88,70	98,31	95,48	95,48	97,18	93,22	95,48
cbr	175	80,57	95,43	96,00	89,71	94,29	94,29	92,00
zma	172	86,05	97,67	96,51	95,93	97,09	95,93	96,51
ola	168	79,76	95,24	94,64	92,26	94,05	93,45	92,86
vvi	162	83,95	97,53	95,68	96,91	97,53	93,83	94,44

(cont. on next page)

Table A.1 (cont.).

Acronym	hp	CR	CDT	CNB	CM	AvgDT	AvgNB	NgDT
nta	161	90,68	94,41	94,41	94,41	94,41	94,41	95,03
crm	157	76,43	97,45	97,45	84,71	94,90	96,18	92,36
bfl	156	87,82	96,15	94,23	93,59	95,51	94,23	96,15
mes	153	92,16	100,00	98,04	100,00	100,00	97,39	99,35
sme	148	76,35	95,27	91,89	86,49	95,27	90,54	96,62
nve	141	62,41	82,98	78,01	74,47	78,01	73,76	79,43
dsi	135	71,11	90,37	89,63	79,26	85,93	87,41	82,96
ccr	134	91,04	98,51	97,76	94,78	98,51	97,01	97,01
dvi	134	74,63	97,01	92,54	92,54	96,27	91,79	94,03
pxy	133	85,71	96,24	94,74	91,73	95,49	92,48	96,24
tni	132	87,12	99,24	98,48	95,45	99,24	97,73	98,48
fru	131	87,79	100,00	98,47	96,95	99,24	98,47	100,00
cte	129	80,62	96,90	96,12	92,25	95,35	93,80	96,90
lus	124	83,06	100,00	97,58	98,39	100,00	96,77	99,19
api	123	79,67	95,12	91,87	87,80	95,12	90,24	96,75
cme	120	75,00	97,50	96,67	95,83	96,67	95,00	93,33
atr	119	85,71	99,16	99,16	98,32	99,16	99,16	98,32
bbe	118	85,59	96,61	93,22	94,07	95,76	93,22	95,76
tae	116	84,48	96,55	95,69	93,10	93,97	92,24	93,10
bma	115	79,13	95,65	92,17	92,17	94,78	89,57	93,04
sma	115	43,48	72,17	68,70	50,43	66,09	63,48	60,00
oar	106	78,30	97,17	86,79	91,51	94,34	85,85	90,57
str	106	82,08	98,11	96,23	90,57	98,11	95,28	94,34
aae	101	77,23	99,01	97,03	93,07	99,01	97,03	96,04
mse	98	83,67	94,90	91,84	92,86	92,86	91,84	93,88
asu	97	80,41	97,94	92,78	90,72	94,85	92,78	97,94
bra	96	95,83	100,00	100,00	100,00	100,00	100,00	100,00
hme	92	71,74	97,83	91,30	92,39	93,48	90,22	93,48
bna	90	91,11	100,00	98,89	97,78	100,00	98,89	97,78
sko	89	96,63	98,88	98,88	97,75	98,88	98,88	98,88
ppa	88	90,91	98,86	95,45	96,59	98,86	93,18	95,45
ata	88	88,64	100,00	98,86	98,86	100,00	97,73	98,86
dgr	82	84,15	100,00	98,78	92,68	100,00	98,78	100,00

(cont. on next page)

Table A.1 (cont.).

Acronym	hp	CR	CDT	CNB	CM	AvgDT	AvgNB	NgDT
tcc	82	91,46	100,00	100,00	100,00	100,00	100,00	100,00
der	81	77,78	95,06	97,53	86,42	92,59	95,06	91,36
cpa	79	81,01	92,41	87,34	86,08	88,61	87,34	92,41
dse	78	85,90	93,59	96,15	89,74	91,03	93,59	91,03
dwi	77	75,32	97,40	97,40	88,31	97,40	94,81	97,40
ghr	77	81,82	98,70	97,40	97,40	98,70	96,10	96,10
dan	76	76,32	94,74	96,05	90,79	94,74	92,11	94,74
dya	76	76,32	94,74	93,42	86,84	93,42	93,42	92,11
dpe	75	80,00	98,67	98,67	89,33	97,33	97,33	98,67
sly	75	94,67	98,67	98,67	97,33	98,67	98,67	98,67
mne	74	87,84	100,00	97,30	97,30	98,65	93,24	95,95
cqu	74	79,73	94,59	95,95	89,19	91,89	93,24	91,89
dmo	71	84,51	98,59	95,77	92,96	97,18	95,77	97,18
hvu	69	46,38	78,26	62,32	66,67	69,57	56,52	57,97
odi	66	77,27	87,88	83,33	84,85	86,36	81,82	86,36
aga	66	86,36	98,48	98,48	95,45	98,48	98,48	98,48
sha	64	40,63	89,06	76,56	85,94	87,50	75,00	81,25
spu	63	73,02	84,13	85,71	82,54	82,54	84,13	82,54
rco	63	92,06	100,00	100,00	100,00	100,00	100,00	100,00
lja	62	66,13	95,16	85,48	80,65	91,94	85,48	88,71
age	60	93,33	98,33	96,67	98,33	98,33	96,67	95,00
gsa	60	73,33	95,00	91,67	81,67	95,00	90,00	96,67
csi	60	56,67	86,67	76,67	80,00	85,00	73,33	83,33
lgi	59	93,22	100,00	98,31	100,00	98,31	98,31	98,31
smo	58	87,93	100,00	100,00	100,00	100,00	100,00	100,00
sja	56	50,00	62,50	64,29	57,14	60,71	62,50	64,29
nvi	53	84,91	98,11	94,34	94,34	96,23	90,57	94,34
tur	52	84,62	100,00	100,00	98,08	100,00	100,00	100,00
lva	50	92,00	96,00	96,00	94,00	96,00	96,00	96,00
cre	50	88,00	100,00	98,00	96,00	100,00	98,00	98,00
pmi	49	87,76	100,00	93,88	97,96	97,96	93,88	93,88
isc	49	79,59	97,96	97,96	93,88	95,92	97,96	95,92
lla	48	91,67	100,00	95,83	97,92	97,92	93,75	93,75

(cont. on next page)

Table A.1 (cont.).

Acronym	hp	CR	CDT	CNB	CM	AvgDT	AvgNB	NgDT
cca	48	93,75	100,00	97,92	100,00	100,00	97,92	97,92
esi	46	93,48	97,83	95,65	93,48	97,83	95,65	95,65
aqc	45	84,44	97,78	93,33	93,33	95,56	88,89	95,56
dpu	44	72,73	81,82	75,00	81,82	79,55	75,00	81,82
sla	42	88,10	97,62	95,24	100,00	97,62	95,24	90,48
hhi	40	67,50	92,50	85,00	85,00	87,50	82,50	85,00
pab	40	55,00	82,50	80,00	75,00	77,50	80,00	75,00
rlcv	36	88,89	97,22	97,22	97,22	97,22	94,44	97,22
pta	34	91,18	97,06	94,12	91,18	94,12	94,12	91,18
ngi	32	84,38	93,75	93,75	87,50	93,75	93,75	90,63
rgl	32	40,63	75,00	68,75	62,50	68,75	65,63	68,75
hbr	31	77,42	96,77	96,77	83,87	93,55	96,77	87,10
pde	29	93,10	96,55	96,55	96,55	96,55	96,55	96,55
pgi	29	100,00	100,00	100,00	100,00	100,00	100,00	100,00
nlo	28	92,86	96,43	96,43	92,86	96,43	96,43	96,43
csa	27	74,07	92,59	77,78	92,59	92,59	74,07	92,59
ebv	25	84,00	100,00	96,00	100,00	100,00	96,00	100,00
rmi	24	25,00	54,17	45,83	37,50	41,67	41,67	50,00
dev	24	45,83	75,00	79,17	62,50	70,83	79,17	66,67
egr	23	69,57	95,65	91,30	86,96	95,65	86,96	95,65
ahy	23	65,22	73,91	73,91	69,57	69,57	73,91	86,96
xla	22	90,91	95,45	90,91	86,36	90,91	90,91	81,82
emu	22	72,73	100,00	90,91	95,45	100,00	86,36	90,91
pol	20	85,00	95,00	95,00	90,00	95,00	95,00	95,00
aja	19	89,47	100,00	100,00	100,00	100,00	94,74	94,74
ssp	19	78,95	100,00	100,00	94,74	100,00	100,00	94,74
vun	18	94,44	100,00	100,00	94,44	100,00	100,00	94,44
ssl	18	88,89	100,00	100,00	100,00	100,00	94,44	100,00
hsv1	18	72,22	88,89	88,89	72,22	83,33	83,33	88,89
hsv2	18	83,33	88,89	88,89	88,89	88,89	88,89	88,89
mcmv	18	72,22	88,89	83,33	88,89	88,89	83,33	77,78
mdv2	18	72,22	88,89	94,44	83,33	83,33	94,44	83,33
hma	17	76,47	94,12	94,12	94,12	94,12	94,12	94,12

(cont. on next page)

Table A.1 (cont.).

Acronym	hp	CR	CDT	CNB	CM	AvgDT	AvgNB	NgDT
ddi	17	100,00	100,00	100,00	100,00	100,00	100,00	100,00
hvt	17	41,18	70,59	64,71	58,82	70,59	58,82	70,59
lca	16	87,50	93,75	87,50	93,75	93,75	87,50	93,75
htu	16	87,50	100,00	93,75	100,00	100,00	93,75	100,00
sof	16	50,00	93,75	100,00	87,50	87,50	100,00	87,50
far	15	13,33	80,00	53,33	53,33	73,33	53,33	46,67
hcmv	15	60,00	100,00	100,00	80,00	93,33	93,33	100,00
mghv	15	73,33	93,33	93,33	86,67	86,67	80,00	80,00
mdv1	14	71,43	100,00	92,86	92,86	100,00	92,86	100,00
pti	13	0,00	76,92	46,15	38,46	69,23	46,15	15,38
gso	13	92,31	100,00	100,00	100,00	100,00	92,31	100,00
dpr	13	92,31	100,00	100,00	100,00	100,00	100,00	100,00
kshv	13	61,54	92,31	84,62	69,23	84,62	84,62	69,23
prv	13	38,46	100,00	100,00	84,62	84,62	84,62	84,62
ocu	12	100,00	100,00	100,00	100,00	100,00	100,00	100,00
hbv	12	75,00	91,67	100,00	83,33	91,67	100,00	100,00
pbi	11	100,00	100,00	100,00	100,00	100,00	100,00	90,91
ssy	11	100,00	100,00	100,00	100,00	100,00	100,00	100,00
bol	10	100,00	100,00	100,00	100,00	100,00	100,00	100,00
bhv1	10	30,00	60,00	80,00	50,00	60,00	70,00	60,00
xbo	8	75,00	75,00	75,00	75,00	75,00	75,00	75,00
aqu	8	100,00	100,00	100,00	100,00	100,00	100,00	100,00
pvu	8	75,00	100,00	100,00	75,00	100,00	100,00	75,00
lmi	7	71,43	100,00	100,00	100,00	100,00	100,00	100,00
aau	7	71,43	71,43	85,71	71,43	71,43	85,71	100,00
iltv	7	85,71	85,71	85,71	85,71	85,71	85,71	85,71
rrv	7	85,71	100,00	100,00	100,00	100,00	100,00	100,00
han	6	66,67	66,67	66,67	66,67	66,67	66,67	66,67
ctr	6	33,33	100,00	100,00	83,33	100,00	100,00	83,33
egu	6	83,33	100,00	100,00	100,00	100,00	100,00	100,00
mja	5	40,00	80,00	60,00	40,00	60,00	60,00	80,00
hru	5	80,00	80,00	80,00	80,00	80,00	80,00	100,00
ccl	5	60,00	100,00	100,00	100,00	100,00	100,00	100,00

(cont. on next page)

Table A.1 (cont.).

Acronym	hp	CR	CDT	CNB	CM	AvgDT	AvgNB	NgDT
bhv5	5	60,00	100,00	100,00	60,00	100,00	100,00	100,00
blv	5	100,00	100,00	100,00	100,00	100,00	100,00	100,00
cln	4	75,00	100,00	100,00	75,00	100,00	100,00	100,00
bcy	4	75,00	100,00	100,00	100,00	100,00	100,00	75,00
bgy	4	100,00	100,00	100,00	100,00	100,00	100,00	100,00
crt	4	75,00	100,00	100,00	100,00	100,00	100,00	100,00
peu	4	0,00	100,00	50,00	75,00	100,00	50,00	50,00
hhv6b	4	75,00	75,00	75,00	100,00	75,00	75,00	75,00
meu	3	66,67	100,00	66,67	100,00	100,00	66,67	100,00
smr	3	100,00	100,00	100,00	100,00	100,00	100,00	100,00
har	3	33,33	66,67	66,67	66,67	66,67	66,67	66,67
hci	3	33,33	66,67	66,67	66,67	66,67	33,33	33,33
hpa	3	66,67	100,00	100,00	100,00	100,00	100,00	66,67
hpe	3	100,00	100,00	100,00	100,00	100,00	100,00	100,00
amg	3	66,67	66,67	100,00	66,67	66,67	100,00	100,00
hiv1	3	33,33	33,33	66,67	33,33	33,33	66,67	33,33
hvsa	3	100,00	100,00	100,00	100,00	100,00	100,00	100,00
pra	2	100,00	100,00	100,00	100,00	100,00	100,00	100,00
psj	2	100,00	100,00	100,00	100,00	100,00	100,00	100,00
cla	2	100,00	100,00	100,00	100,00	100,00	100,00	100,00
hex	2	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ama	2	100,00	100,00	100,00	100,00	100,00	100,00	100,00
bfv	2	50,00	100,00	100,00	100,00	100,00	100,00	100,00
pin	1	100,00	100,00	100,00	100,00	100,00	100,00	100,00
gpy	1	100,00	100,00	100,00	100,00	100,00	100,00	100,00
tre	1	100,00	100,00	100,00	100,00	100,00	100,00	100,00
lco	1	100,00	100,00	100,00	100,00	100,00	100,00	100,00
sci	1	100,00	100,00	100,00	100,00	100,00	100,00	100,00
gar	1	100,00	100,00	100,00	100,00	100,00	100,00	100,00
ghb	1	100,00	100,00	100,00	100,00	100,00	100,00	100,00
ttu	1	100,00	100,00	100,00	100,00	100,00	100,00	100,00
bkv	1	100,00	100,00	100,00	100,00	100,00	100,00	100,00
bpcv1	1	0,00	100,00	100,00	100,00	100,00	0,00	100,00

(cont. on next page)

Table A.1 (cont.).

Acronym	hp	CR	CDT	CNB	CM	AvgDT	AvgNB	NgDT
bpcv2	1	100,00	100,00	100,00	100,00	100,00	100,00	100,00
jcv	1	100,00	100,00	100,00	100,00	100,00	100,00	100,00
mcv	1	0,00	100,00	100,00	100,00	100,00	100,00	100,00
sv40	1	100,00	100,00	100,00	100,00	100,00	100,00	100,00

APPENDIX B

ACCURACY DISTRIBUTIONS

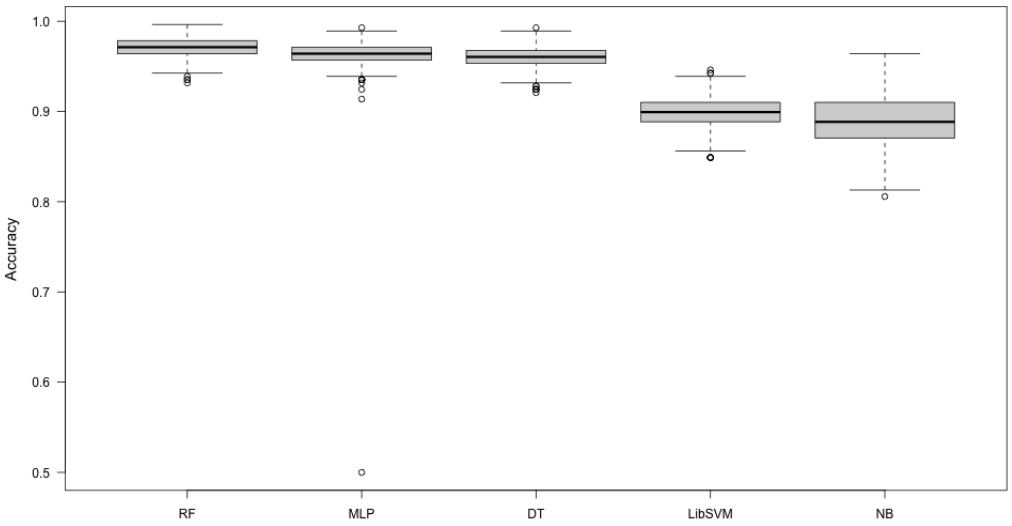

Figure B.1. The accuracy distributions of five classifiers for *Solanum lycopersicum* hairpin analysis.

VITA

Date and Place of Birth: 22.03.1988, İzmir - Turkey

EDUCATION

2014 - 2017 Doctor of Philosophy in Molecular Biology and Genetics

Graduate School of Engineering and Sciences, İzmir Institute of Technology

Thesis: Computational Establishment of MicroRNA Metabolic Networks

2011 - 2013 Master of Science in Molecular Biology and Genetics

Graduate School of Engineering and Sciences, İzmir Institute of Technology

Thesis: An Integrative Data Mining Approach for MicroRNA Detection in Human

Supervisor: Assoc. Prof. Dr. Jens Allmer

2006 - 2011 Bachelor of Molecular Biology and Genetics

Department of Molecular Biology and Genetics, İzmir Institute of Technology

PROFESSIONAL EXPERIENCE

2012 - 2017 Research and Teaching Assistant

Department of Molecular Biology and Genetics, İzmir Institute of Technology

SELECTED PUBLICATIONS

Saçar MD and Allmer J (2013) Data mining for miRNA gene prediction:on the impact of class imbalance and feature number for miRNA gene prediction, IEEE Xpl.

Saçar MD and Allmer J (2014) Machine learning methods for miRNA gene prediction, Methods in Molecular Biology, Springer, 1107:177 - 187.

Saçar MD, Bağçı C, and Allmer J (2014) Computational prediction of microRNAs from *Toxoplasma gondii* potentially regulating the hosts' gene expression, GPB.

Khalifa W, Yousef M, Saçar Demirci MD and Allmer J (2016) The impact of feature selection on one and two-class classification performance for plant microRNAs, PeerJ.

Saçar Demirci MD, Bağçı C, and Allmer J (2016) Differential expression of *Toxoplasma gondii* microRNAs in murine and human hosts, Springer, 143 - 159.

Yousef M, Saçar Demirci MD, Khalifa W and Allmer J (2016) Feature selection has a large impact on one-class classification accuracy for miRNAs in plants, Adv. in Bioinf.

Saçar Demirci MD, Toprak M, and Allmer J (2016) A Machine Learning Approach for MicroRNA Precursor Prediction in Retro-transcribing Virus Genomes, JIB.

Saçar Demirci MD and Allmer J (2017) Delineating the impact of machine learning elements in pre-microRNA detection, PeerJ.