AN INTEGRATIVE DATA MINING APPROACH FOR MICRORNA DETECTION IN HUMAN

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

MASTER OF SCIENCE

in Molecular Biology and Genetics

by
Müşerref Duygu SAÇAR

December 2013
İZMİR

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

Examining Committee Members:

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

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

_____
Prof. Dr. Bilge KARAÇALI
Department of Electrical and Electronics Engineering, İzmir Institute of Technology

16 December 2013

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

_____
Assoc. Prof. Dr. Ahmet KOÇ
Head of the Department of
Molecular Biology and Genetics

_____
Prof. Dr. R. Tuğrul SENGER
Dean of the Graduate School of
Engineering and Sciences

ACKNOWLEDGMENTS

I would like to express my deep and sincere gratitude to Assoc. Prof. Dr. Jens ALLMER, for his help, leadership, encouragement and patience during my graduate education including the preparation of this thesis. I see myself as one of the few people who had the chance to benefit from his wide knowledge and world view.

I want to thank my thesis committee members Prof. Dr. Bilge KARAÇALI and Assist. Prof. Dr. Bünyamin AKGÜL for their advices.

I also would like to thank to all my colleagues in İYTE – Bioinformatics for their patience and kindness.

I am especially thankful to my dear friends Sema TABAK, Şule YILMAZ, Mustafa TOPRAK, Mehmet GÖKTAY and Canan HAS for being with me in good times to have fun and in bad times to cry and complain together.

I am deeply grateful to my family for their support, love, understanding and encouragement in my whole life. My beloved brother, Sarper, you are the most caring and honest person I have ever known, thank you so much for being in my life.

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 support, understanding and patience even during hard times of this study.

ABSTRACT

AN INTEGRATIVE DATA MINING APPROACH FOR MICRORNA DETECTION IN HUMAN

MicroRNAs (miRNAs) are single-stranded, small, usually non-coding RNAs of about 22 nucleotides in length, that control gene expression at the posttranscriptional level through translational inhibition, degradation, adenylation, or destabilization of their target mRNAs. Although hundreds of miRNAs have been identified in various species, many more may still remain unknown. Therefore, the discovery of new miRNA genes is an important step for understanding miRNA mediated post transcriptional regulation mechanisms. First attempts for the identification of novel miRNA genes were almost exclusively based on directional cloning of endogenous small RNAs and high-throughput sequencing of large numbers of cDNA clones. However, conventional forward genetic screening is known to be biased towards abundantly and/or ubiquitously expressed miRNAs that can dominate the cloned products. Hence, such biological approaches might be limited in their ability to detect rare miRNAs, and restricted to the tissues and the developmental stage of the organism under examination. These limitations have led to the development of sophisticated computational approaches attempting to identify possible miRNAs *in silico*. Nevertheless, the programs designed to predict possible miRNAs in a genome are not sensitive or accurate enough to warrant sufficient confidence for validating all their predictions experimentally. With this study, we aim to solve these problems by developing a new and sensitive machine learning based approach to predict potential miRNAs in the human genome.

ÖZET

İNSAN MİKRORNA TESPİTİ İÇİN BÜTÜNLEŞTİRİCİ VERİ MADENCİLİĞİ YAKLAŞIMI

MikroRNAlar (miRNA) uzunluğu yaklaşık 22 nükleotid olan, tek diziden oluşan ve genellikle kodlama özelliği olmayan küçük RNA'lardır ve gen ekspresyonunu posttranskripsiyonel seviyede, hedefleri olan mRNAların translasyonel baskılanması ve istikrarsızlaştırılması yoluyla kontrol ederler. Çeşitli türlerde yüzlerce miRNA tespit edilmesine rağmen, miRNAların büyük bir çoğunluğu hala bilinmiyor olabilir. Bu nedenle, yeni miRNA genlerinin keşfi, miRNA aracılığıyla düzenlenen transkripsiyon sonrası regülasyon mekanizmalarının anlaşılması için önemli bir adımdır. MiRNA genlerin belirlenmesi için ilk girişimlerin neredeyse tamamı endojen küçük RNA'ların yönlü klonlanmasına ve çok sayıdaki cDNA klonlarının yüksek verimli sıralanmasına dayalıdır. Ancak konvansiyonel ileri genetik tarama, klonlanmış ürünleri domine eden, yüksek miktarda sentezlenen ve/veya her yerde görülen miRNAlara karşı yanlı bir yöntemdir. Fakat bu tarz biyolojik yöntemler nadir miRNAların saptanmasında etkisiz kalmaktadır. İncelenen doku ve organizmanın içinde bulunduğu gelişimsel dönemlerin farklılıkları da bu yöntemleri sınırlandıran faktörlerdendir. Bu sınırlamalar, olası miRNAları *in silico* olarak bulmak için sofistike bilgisayar programlarının geliştirilmesine yol açmıştır. Ancak bir genomdaki muhtemel miRNAları tahmin etme amacıyla oluşturulan bu programlar, tahminlerini deneysel olarak doğrulamak için yeterli güveni garanti edebilecek kadar hassas ya da kesin olmaktan çok uzaklardır. Bu çalışma ile yeni ve hassas, makine öğrenimine dayalı bir yaklaşım geliştirerek insan genomundaki olası miRNAların tahmini ve tüm bu sorunların çözülmesi amaçlanmıştır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1. MicroRNA Definition

MicroRNAs (miRNAs) are single-stranded and small RNAs of about 22 nt in length, that control gene expression at the posttranscriptional level through translational inhibition and destabilization of their target mRNAs (Ambros, Lee, Lavanway, Williams, & Jewell, 2003; Filipowicz, Bhattacharyya, & Sonenberg, 2008). They were initially discovered from *C. elegans*, as regulatory molecules modulating the developmental timing (R. C. Lee, Feinbaum, & Ambros, 1993). Genetic, biochemical, and computational studies have shown fundamental and diverse roles of miRNAs in multi-cellular organisms. Thousands of miRNAs have been discovered in various species, and their various roles are rapidly being elucidated (Bushati & Cohen, 2007; Jones-Rhoades, Bartel, & Bartel, 2006). Links between miRNA and human diseases including cancer and neurodegenerative diseases have also been established (Bushati & Cohen, 2007; Hébert et al., 2009; G. Wang et al., 2008).

In mammals, it has been estimated that the activity of approximately 30% of all protein-coding genes are controlled by miRNAs (Filipowicz et al., 2008). Not only higher eukaryotes, even simple multicellular organisms like poriferans (sponges) and cnidarians (starlet sea anemone) have miRNAs (V. N. Kim, Han, & Siomi, 2009). Moreover, many of the animal miRNAs seem to be phylogenetically conserved; ~55% of *C. elegans* miRNAs have homologues in humans, which implies that these miRNAs have had essential roles throughout evolution (Ibáñez-Ventoso, Vora, & Driscoll, 2008). However, animal and plant miRNAs appear to be evolved separately since their sequences, precursor structure and biogenesis mechanisms are very different from each other (Chapman & Carrington, 2007; Millar & Waterhouse, 2005).

Almost half of mammalian miRNA loci are found in nearby to other miRNAs and these clustered miRNAs are transcribed from a single polycistronic transcription unit (TU) (Y. Lee et al., 2004), while there may be cases in which particular miRNAs are

originated from distinct gene promoters (V. N. Kim et al., 2009) (Figure 1). While some miRNAs are encoded in non-coding TUs, others come from protein-coding TUs (V. N. Kim et al., 2009). About 40% of miRNA loci are located in the intronic region of non-coding transcripts, ~10% are found in the exonic region of non-coding TUs and the remaining ~40% of all miRNA loci are found in protein coding TUs; generally placed in intronic regions (V. N. Kim et al., 2009). Moreover, based on the alternative splicing events, it is possible to encounter some mixed miRNA genes which can be assigned to either intronic or exonic miRNA groups.

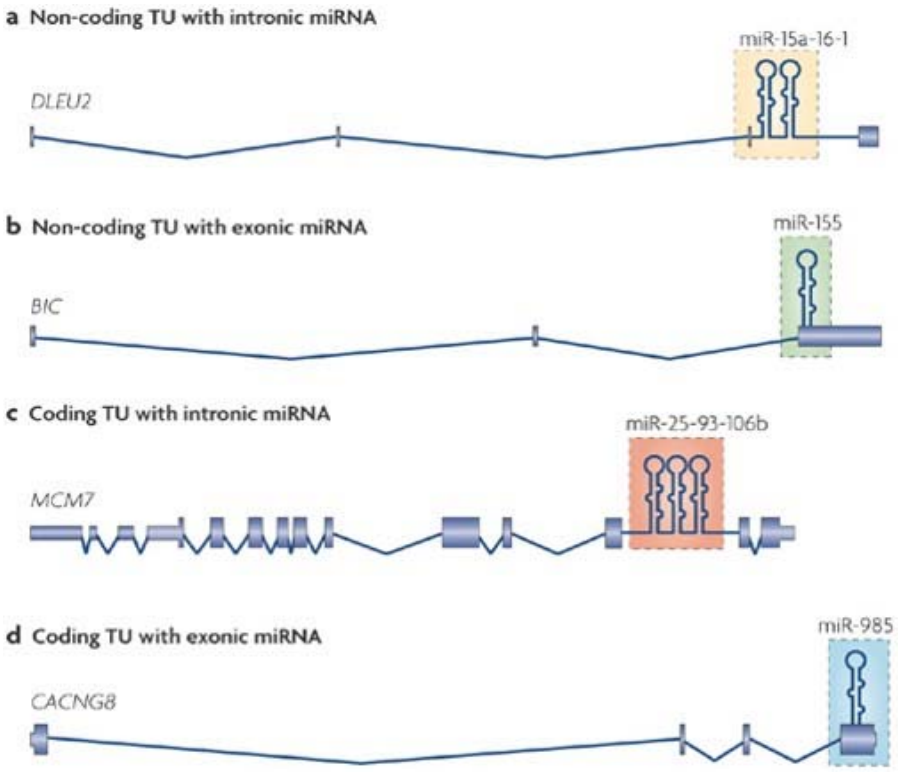

Figure 1. Genomic location and gene structure of miRNAs. MiRNAs can be classified into four groups according to their genomic locations relative to exon and intron positions. A) Intronic miRNAs in non-coding transcripts, e.g. the miR-15a~16-1 cluster found in the intron of a non-coding RNA gene, *DLEU2*. b) Exonic miRNAs in non-coding transcripts, e.g. miR-155 found in a previously defined non-coding RNA gene, *BIC*. c) Intronic miRNAs in protein-coding transcripts, e.g. the miR-25~93~106b cluster found in the intron of the DNA replication licensing factor *MCM7*. d) Exonic miRNAs in protein-coding transcripts, e.g. miR-985 hairpin found in the last exon of *CACNG8* mRNA. The hairpins represent miRNA stem-loops while blue boxes indicate the protein-coding regions. (Source: V. N. Kim et al., 2009)

1.2. MicroRNA Biogenesis

It has been stated that the biogenesis of miRNAs consist of three step-wise activities, firstly transcription of primary miRNAs (pri-miRNAs) from the miRNA genes (Y. Lee, Jeon, Lee, Kim, & Kim, 2002), partial processing of precursor miRNAs (pre-miRNAs) in nuclei (Hutvágner et al., 2001) and the formation of mature miRNAs in the cytoplasm (*Figure 2*). Pri-miRNAs are transcripts whose tertiary structure forms stem loop structures that are cleaved off by the microprocessor machinery in order to form ~60-100 nucleotide long pre-miRNA. Then, the pre-miRNA is transferred from the nucleus by exportin-5 in a ran-GTP dependent manner (Yi, Qin, Macara, & Cullen, 2003). Binding of pre-miRNA by exportin-5 requires a length of 16-18 base pairs in the stem of the miRNA, and its efficiency is affected by the variations in the 3' overhang (Zeng & Cullen, 2004). At the next step, these pre-miRNAs are transferred from nucleus to cytoplasm where they are released from exportin-5 after the hydrolysis of GTP and further processed into ~22 nucleotide long mature miRNAs by Dicer (Y. Lee et al., 2003). The RNase III enzyme, Dicer, removes the loop structure and 3' overhang in an ATP independent manner (H. Zhang, Kolb, Brondani, Billy, & Filipowicz, 2002). Different domains of Dicer take part in the recognition and the correct cleavage of the pre-miRNA. Following cleavage, one strand of the miRNA duplex is preferentially directed to the RNA inducing silencing complex (RISC). The thermodynamic properties of the duplex is the deciding factor for the selection of one strand over the other, and the strand with the less thermodynamical stability at the 5' end is mostly selected (Khvorova, Reynolds, & Jayasena, 2003). The mature miRNA - RISC complex then partners with an Argonaute protein, (commonly Ago2), and leads binding of the RISC complex to partly complementary positions usually located in the 3'-UTRs of targeting mRNAs (G. Wang et al., 2008).

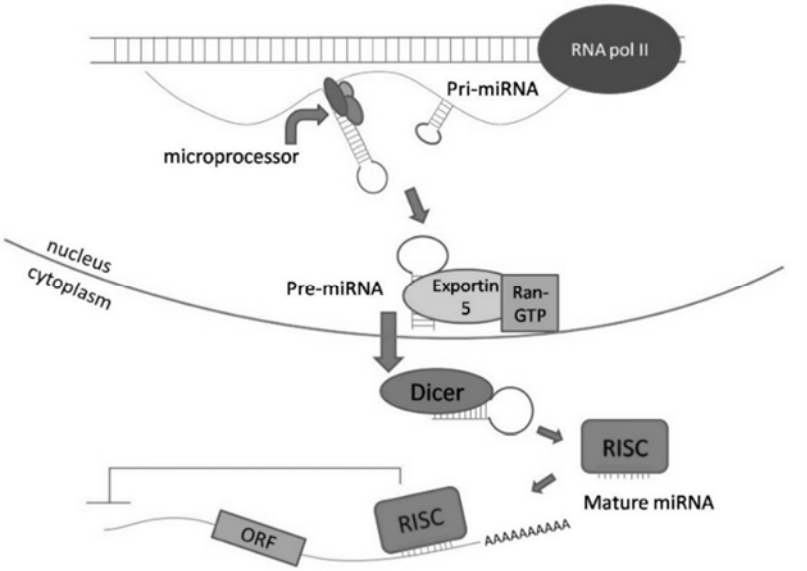

Figure 2. MiRNA biogenesis. RNA polymerase II performs the transcription of pri-miRNA. The stem loop structure is cleaved off by the microprocessor resulting in pre-miRNA which is then exported to cytoplasm by exportin5 in a ran-GTP dependent manner. In cytoplasm, the pre-miRNA is further processed by Dicer making a single stranded mature miRNA which would be bound by the RISC complex, guiding it to the target mRNAs and leading to repression of protein expression. (Source: Beezhold, Castranova, & Chen, 2010)

1.3. MicroRNA Gene Prediction

Previously, identification of novel miRNA genes were almost exclusively based on the usage of directional cloning of endogenous small RNAs and high-throughput sequencing of large numbers of cDNA clones (M Lagos-Quintana, Rauhut, Lendeckel, & Tuschl, 2001; Lau, Lim, Weinstein, & Bartel, 2001) (Table 1). Conventional forward genetic screening is biased toward abundantly and/or ubiquitously expressed miRNAs that dominate the cloned products (M Lagos-Quintana et al., 2001). Apparently, such biological approaches might be limited in their ability to detect rare miRNAs, and are of course limited to the tissues examined. This and the fact that miRNA precursors share a common secondary hairpin-shaped structure, has led to the development of sophisticated computational approaches attempting to identify possible miRNAs (Berezikov, Cuppen, & Plasterk, 2006).

Table 1. Experimental tools applied to discovery of miRNAs.
(Source: Gomes et al., 2013)

Technique	Study	Organism
Cloning	(R. C. Lee et al., 1993)	Nematode
	(Mariana Lagos-Quintana et al., 2002)	Mouse
	(Pfeffer et al., 2005)	Virus
	(Bentwich et al., 2005)	Human
	(He, Zhang, Liu, & Pan, 2007)	Rat
	(G. Xu et al., 2009)	Bovine
	(Long & Chen, 2009)	Cattle
Microarray	(Liu & Theil, 2004)	Human, Mouse
In situ hybridization	(Wienholds et al., 2005)	Zebrafish
	(Kloosterman, Wienholds, de Bruijn, Kauppinen, & Plasterk, 2006)	Zebrafish, Mouse
	(Nelson et al., 2006)	Human
	(Deo, Yu, Chung, Tippens, & Turner, 2006)	Mouse
Next-Generation Sequencing	(Bar et al., 2008)	Human
	(Morin et al., 2008)	Plant
	(Friedländer et al., 2008)	Nematode
	(Meng, Hackenberg, Li, Yan, & Chen, 2012)	Rat
	(Guzman, Almerão, Körbes, Loss-Morais, & Margis, 2012)	Plant
	(B. Kim et al., 2012)	Plant

Numerous approaches for the *in silico* prediction of miRNAs have been proposed so far (*Table* 2). These programs commonly regard the hairpin secondary structure of the miRNA precursor as the most important characteristic of a miRNA gene (van der Burgt, Fiers, Nap, & van Ham, 2009). RNA secondary structure prediction (RSSP) algorithms such as RNAfold (Hofacker, 2003) also estimate the thermodynamic stability of the RNA hairpin structures. Existing bioinformatics methods for the prediction of miRNA usually consist of: (1) genome-wide estimation of hairpin structures; (2) filtering (or scoring) of those hairpins based on their similarity in structure and sequence to known miRNA hairpins and (3) experimental confirmation of putative candidates (van der Burgt et al., 2009).

Table 2. Selected computational tools for miRNA prediction.
(Source: Gomes et al., 2013)

Tool	Conservation	Structure	Sequence	Machine learning	NGS application
miRscan (Lim et al., 2003)	+	+			
miRAlign (X. Wang et al., 2005)		+	+		
ProMiR (Nam et al., 2005)	+	+	+	+	
Triplet-SVM (Xue et al., 2005)		+	+	+	
miR-abela (Sewer et al., 2005)		+	+	+	
RNAmicro (Hertel & Stadler, 2006)		+	+	+	
miRFinder (Huang et al., 2007)	+	+		+	
miPred (Jiang et al., 2007)		+		+	
MiRRim (Terai, Komori, Asai, & Kin, 2007)	+	+		+	
miRDeep (Friedländer et al., 2008)		+			+
miRanalyzer (Hackenberg, Sturm, Langenberger, Falcón-Pérez, & Aransay, 2009)				+	+
SSCprofiler (Oulas et al., 2009)	+	+	+	+	+
HHMMiR (Kadri, Hinman, & Benos, 2009)		+	+	+	

Computational miRNA gene prediction approaches can be grouped into several categories. Generally, either homology modelling or *ab initio* methods are applied to extract possible miRNAs from a genome. However, there are at least two computational challenges: 1) the prediction of miRNAs in a genome and 2) the mapping of the miRNAs to likely targets.

1.3.1. Homology Based MicroRNA Gene Prediction

Homology-based mapping methods can build on available, experimentally validated, miRNAs and find similar structures and sequences in related species. The idea is that if a miRNA is identified in one genome then its homologs can be possibly found in other species (Lindow & Gorodkin, 2007). Since conservation indicates a function, it is assumed that conserved candidates are more likely to be miRNAs. Although, it has

been shown that for non-coding RNAs absence of conservation does not inevitably mean lack of function (Pang, Frith, & Mattick, 2006), searching for homologs especially in newly annotated genomes is a beneficial approach. Software facilitating mapping of known miRNAs to homologous genomes take both sequence similarity and miRNA secondary structure information into account. The theory is based on derivation of mature miRNAs from hairpin structure formed by folding its pre-miRNA. The approach taken by one of the most recent developments, MapMi (Guerra-Assunção & Enright, 2010), first scans the miRNA sequences against the target genome and then creates two potential pre-miRNAs from it. The ViennaRNA package (Hofacker, 2003) is used to compute the most likely folding of the extracted RNA sequences. In the end, the results are scored, ranked and displayed.

Although homology modeling can gather information from already successfully established miRNAs of a related organism's genome, it is also limited since completely novel miRNAs cannot be determined in this way. First attempts in this approach are mainly relied on identifying close homologs of published pre-miRNAs i.e. let-7 (Pasquinelli et al., 2000). This method might be seen as straightforward as aligning sequences through NCBI BlastN (McGinnis & Madden, 2004) but it can only reproduce results and cannot find new miRNA genes. Since many miRNAs are species specific (Mor & Shomron, 2013), they will always be missed by this method and therefore other strategies need to be used in tandem. Additionally, miRNA genes evolve very rapidly which further limits the applicability of homology-based methods (Liang & Li, 2009). A powerful approach developed for genome-wide screening of phylogenetically well conserved pre-miRNAs between closely related species is cross-species sequence conservation based on computationally intensive multiple genome alignments. However, this strategy also suffers lower sensitivity, especially in divergent evolutionary distance (Berezikov et al., 2006; Boffelli et al., 2003). Moreover, identifying pre-miRNAs that differ significantly or undergo rapid evolution at the sequence level while keeping their characteristic evolutionarily conserved hairpin structures, may also pose problems (Ng & Mishra, 2007). Another important issue is that non-conserved pre-miRNAs with genus-specific patterns are also likely to escape detection (Ng & Mishra, 2007).

There are various homology-based miRNA gene prediction softwares such as MirScan (Lim et al., 2003), MirFinder (Huang et al., 2007), miROrtho (Gerlach, Kriventseva, Rahman, Vejnar, & Zdobnov, 2009), miRNAminer (Artzi, Kiezun, &

Shomron, 2008), and ProMiR II (Nam, Kim, Kim, & Zhang, 2006). Also, some of these softwares use machine learning approaches such as ProMir (Nam et al., 2005).

1.3.2. Ab Initio MicroRNA Gene Prediction

While other methods mainly use comparative genomics information, *ab initio* miRNA gene prediction needs no other information than the primary sequence in order to determine whether it is a true miRNA (assuming there are already established negative and positive datasets). This is the only way for the identification of new miRNAs which have no close homologs (Brameier & Wiuf, 2007). Two different modes of operations are possible with one using multiple sequences and the other based on single sequences. The main difficulty in working on miRNAs by using *ab initio* methods is choosing proper rules that can reliably identify a given sequence as a true miRNA based on its properties. For instance, the hairpin structure (*Figure 3*) is one of the most commonly used features of miRNAs in prediction tools such as miPred (Ng & Mishra, 2007). The problem is that, although precursor miRNAs are believed to possess evolutionarily conserved RNA hairpin structures critical for the early stages of the mature miRNA biogenesis, the hairpin shape is not unique to miRNAs (Ding, Zhou, & Guan, 2010) and there are around 11 million hairpins estimated to be present in the human genome (Bentwich, 2008). Moreover, if the chosen discriminating parameter does not have a good sensitivity or accuracy potential, it would not be very useful and might have a potential to produce false positives and miss true miRNAs.

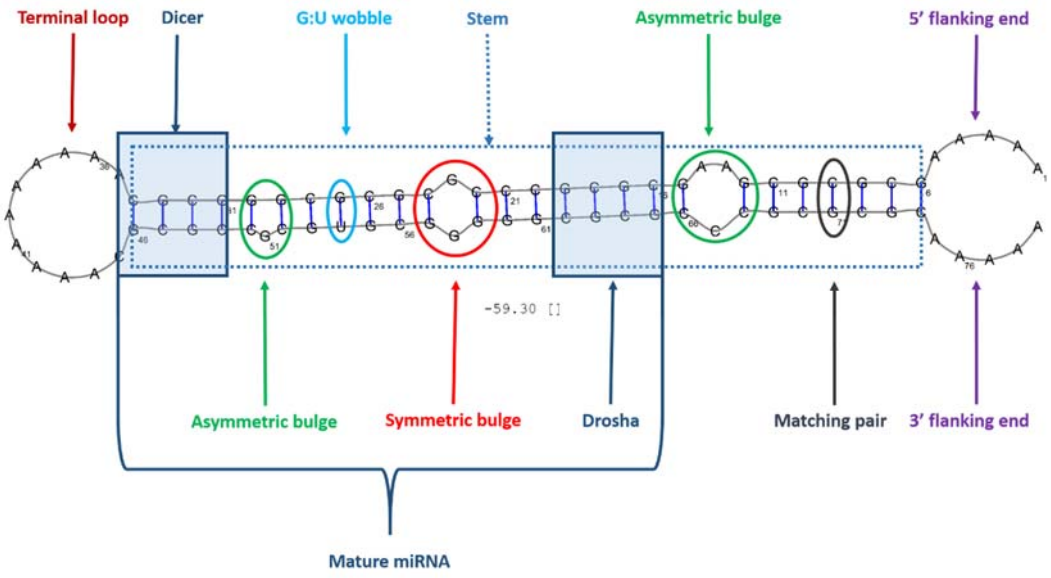

Figure 3. MiRNA hairpin structure. Some of the components of hairpin structure: G-U base pairing, symmetric and asymmetric bulges, stem etc. Secondary RNA structure and minimum free energy value is calculated by using RNAShapes (Steffen et al. 2006). (Source: Saçar & Allmer, 2014).

In a more systems-driven approach, the predicted mature miRNAs can be validated further by looking for targets, for instance 3'UTRs of mRNAs are potential targets for many miRNAs, and by evaluating the multiplicity of targets per miRNA and target sites per regulated mRNA. In order to have a better understanding of miRNAs and to be able to use them in prospective treatments of human diseases, we first need to establish a method to identify the human miRNAs in the genome and then, we should determine their potential targets and modes of action. Although there are many miRNA prediction algorithms designed by different groups, accuracy of such programs are not in the desirable range.

There are various advantages of using *ab initio* methods in miRNA prediction:

a) It is shown that the number of miRNAs present in the human genome is higher than previous estimates (Bentwich et al., 2005). Furthermore, it is estimated that the number of non-conserved miRNAs may be relatively large.

b) *Ab initio* prediction approaches are able to predict miRNAs in a particular genome without using comparative sequence analysis or demanding neither sequence nor structure conservation (Brameier & Wiuf, 2007). This allows the identification of entirely new miRNAs that have no known close homologs.

c) It has been shown that lowly expressed human miRNA genes evolve rapidly (Liang & Li, 2009). In this case, while the methods based on comparative genomic information may be limited in their use, using an *ab initio* approach would predict this kind of miRNAs seamlessly.

1.4. Machine Learning and MicroRNA Gene Prediction

1.4.1. Algorithms for Machine Learning

Machine learning has become a popular choice in many bioinformatics applications and studies (Figure 4). The range of speedily increasing data created by modern molecular biology techniques has intensified the need for accurate classification and prediction algorithms (Bhaskar, Hoyle, & Singh, 2006). There are numerous biological fields where machine learning methods are applied for knowledge extraction from data such as genomics, systems biology, evolution, microarray and proteomics (Larranaga, 2006).

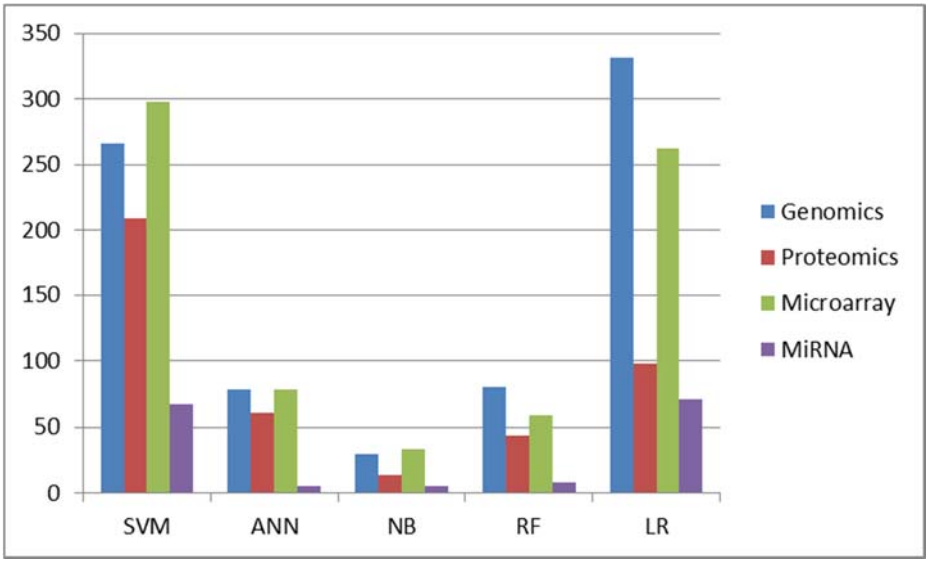

Figure 4. Fields in biology where machine learning methods are applied (The number of publications (y-axis) is calculated by searching PubMed with machine learning approach and the field name as key words). (Source: Saçar & Allmer, 2014).

Machine learning algorithms diverge from the rule-based miRNA prediction algorithms since the ‘‘rules’’ to determine whether a given sequence is a true miRNA are not manually created; instead these rules are learned from samples (Lindow & Gorodkin, 2007). Generally, machine learning-based methods start with the learning process involving the sequence, structure or thermodynamic features of miRNAs. Afterwards, a classifier is formed to decide whether unknown sequences are true miRNAs based on the information gained through positive and negative data sets. Normally, the parameters are a set of numerical features defining a candidate miRNAs such as the minimum free energy of folding and the results would be predictions towards the candidate being a miRNA or not.

Nevertheless, there are two major drawbacks with the current machine learning based miRNA gene identification processes. The main one is the imbalance on the number and quality of positive and negative examples used in classifier training. Due to the uncertainty about the exact number of real miRNAs in any genome, it is supposed that there are very few miRNA precursors in any randomly selected collections of hairpins extracted from the genome (Ding et al., 2010). Likewise, the number of positive examples is considerably smaller than that of negative examples. For instance, one of the commonly used negative dataset for miRNA prediction algorithms consists of approximately 9000 pseudo hairpins while the number of human miRNAs that can be obtained from miRBase is less than 2000 (Ng & Mishra, 2007). In this work, we have shown that the imbalance problem between positive and negative datasets can significantly reduce the performance of current machine learning approaches (see the Results and Discussion Section). Another problem is that most of the current machine learning based algorithms makes assumptions such as the length of the stem, the loop size and the minimum free energy (MFE) of the data, As a result, sequences outside of these predetermined borders are not considered as potential miRNA and cannot be predicted by those methods, inevitably reducing the prediction performance and accuracy (Ding et al., 2010).

To the best of our knowledge, there is no published study that uses unsupervised machine learning approaches for miRNA gene prediction while there are many studies using supervised machine learning algorithms such as support vector machine (SVM), neural networks (NN), hidden Markov models (HMM), and Naive Bayes classifiers (NB).

1.4.2. Supervised Approaches (Classification)

While there are some examples for the usage of unsupervised approaches on miRNA target prediction (Heikkinen, Kolehmainen, & Wong, 2011), machine learning for miRNA gene prediction is entirely based on supervised learning in which an algorithm is trained to learn; approximating a function that maps input data to expected outputs (B.-T. Zhang & Nam, 2008). Mostly, the inputs are a set of features characterizing sample (e.g., mfe, number of dinucleotides, length of stem etc.) and the output would be a decision for either miRNA or non-miRNA. While the anticipated output is unknown, the classifier is trained by a set of known inputs so it can generalize based on these examples (input data; positive and negative examples) and appropriately classify future samples (Lindow & Gorodkin, 2007). The most important element influencing the accuracy of the results is the selection of features, since parameterization of the samples into features is not performed automatically (Ding et al., 2010; Lindow & Gorodkin, 2007). In order to test the accuracy and precision of the machine learning process, a procedure called cross-validation is used. One round of cross-validation includes dividing the available data into two mutually exclusive subsets, training the classifier on one subset (the training or learning set), and validating it on the other subset (testing set) (Figure 5). The dataset can be divided in defined percentages (e.g. 70% of samples included in learning set, remaining 30% included in testing set. See Figure 5) but the crucial point is that these datasets must not have shared instances. After cross-validation the best performing classifier construction strategy is selected to train a classifier using the whole dataset and applied to make predictions on the future data.

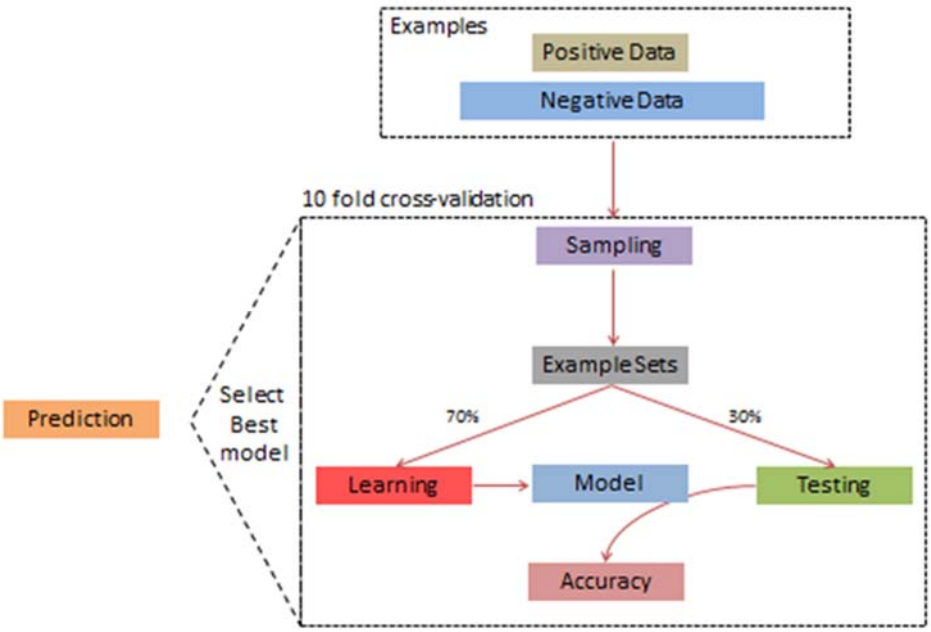

Figure 5. Machine learning based miRNA prediction. General work-flow of machine learning algorithms for miRNA gene prediction. (Source: Saçar & Allmer, 2014).

One of the initial works in the field by (Sewer et al., 2005) assembled 40 different sequence and structural signs to label a candidate as pre-miRNA. The SVM classifier model was trained using 178 known human pre-miRNAs as positive examples and 5395 random sequences obtained from tRNA, rRNA, and mRNA genes as negative examples although in reality, there is no guarantee that these RNAs would not include any functional miRNAs. As a result of huge difference between the number of positive and negative samples, their results have high specificity (91%) and low sensitivity (71%) for their dataset.

ProMiR was introduced in 2005 as an algorithm that uses a Hidden Markov Model and simultaneously takes into account structure and sequences of pre-miRNAs (Nam et al., 2005). A machine learning approach was used with positive examples from known human miRNAs and negative examples acquired randomly from the human genome. The predicted pre-miRNAs are further evaluated according to their minimum free energy and based on whether they are conserved among vertebrates. ProMiR II takes into account additional features such as miRNA gene clustering, G/C ratio conservation, and the entropy of the candidate sequences (Nam et al., 2006).

MatureBayes is a probabilistic algorithm developed by Gkirtzou et. al. using a Naive Bayes classifier to characterize potential mature miRNAs (Gkirtzou, Tsamardinos, Tsakalides, & Poirazi, 2010). Similar to previous approaches, it also performs classification based on the sequence and the secondary structure of the miRNA precursors.

1.4.3. One-Class Classification

The main challenge in classification is apportioning a new object to one of a set of classes which are defined in advance. This classification process is accomplished by using the learned rules based on a number of samples. Differing from other classification approaches, in one-class classification, it is assumed that information on only one of the classes, also known as the target class, is available. Consequently, due to the lack of samples representing the other class, the distinction between the two classes has to be assessed from data of only the real class (Tax, 2001).

Defining the negative class is another difficult challenge to overcome in developing machine learning algorithms for miRNA identification. Hence, machine learning methods have been offered for identifying miRNAs without the requirement of a negative class. Yousef and colleagues performed a study using one-class machine learning approach for miRNA gene prediction by using only positive data to construct the classifier (Yousef, Jung, Showe, & Showe, 2008). Even though the one-class method is less complex to implement and simpler to handle, the two-class systems mostly appear to be superior. Furthermore, there are additional problems due to some characteristic properties of miRNAs; e.g., pre-miRNAs must fold in a hairpin structure, but not all hairpins in the genome are miRNA sequences (Bentwich et al., 2005).

1.5. Learning and Test Data

In miRNA gene prediction studies, it is usually straightforward to select positive examples (e.g., using known miRNAs), while it is quite challenging to create negative samples (Lindow & Gorodkin, 2007). Usually, positive data for miRNA gene predictions are obtained from miRBase (Bentwich, 2008), although there are some entries in miRBase which are suggested as miRNAs but do not satisfy the required features to be

classified as miRNAs such as having more than one loop. It was shown that when a reference set of positive controls taken from miRBase, further improvements are needed to generate a proper high-confidence positive controls (Xue et al., 2005). We recently took this issue in consideration and found that filtering of questionable miRNAs from miRBase improves prediction accuracy (M. D. Saçar, Hamzeiy, & Allmer, 2013).

A proper negative dataset is one of the fundamental requirements for a well-trained classifier. In the presence of overly artificial negative data, there is a high chance that the machine learning method will not be trained sufficiently to discriminate among true miRNAs and non-miRNA sequences (Wu, Wei, Liu, Li, & Rayner, 2011). Conversely, if the negative dataset is very similar to the positive dataset, the machine learning approach will be unable discriminate between these two datasets (Wu et al., 2011).

A small RNA sequence should be recognized and processed by the enzyme Dicer to be categorized as a miRNA. This means that while constructing a negative control, samples should be selected among the transcripts that are expressed in the same cellular section as true miRNAs but are not recognized by Dicer. Since this is a very intricate system to produce negative samples, in most of the algorithms random genomic sequences or exonic sequences are used instead (Brameier & Wiuf, 2007; Xue et al., 2005). These sequences, however, are very inadequate negative controls because there is no confirmation that these transcribed small RNAs would not be recognized by Dicer and the other components of miRNA biogenesis pathway (i.e. Drosha, RISC) and processed into functional mature miRNAs (Xue et al., 2005).

A well-known negative dataset for miRNA gene prediction consists of 8494 pseudo hairpins from human RefSeq genes (Jiang et al., 2007) which have been selected such that they do not undergo any alternative splicing events (Ng & Mishra, 2007).

1.6. Aim of the study

In order to have a better understanding of miRNAs and to be able to use them in prospective treatments of human diseases, first, a method to detect the miRNAs in a genome needs to be established. In this study, our main objective is to design an approach which would allow us to identify miRNAs in human genome using machine learning

methods. To have a comprehensive view on the subject, we analyzed the current literature in the field, checked the sufficiency of the established databases for an accurate prediction and investigated the main problems in machine learning approaches such the impact of largely differing numbers of positive and negative data for *ab initio* miRNA prediction. Moreover, all the possible features (total of 740) that can potentially help discriminate for miRNA hairpins in human are considered and implemented.

Here we propose not only to increase the accuracy of prediction, but also to develop a method that can be used in other organisms. Although there are many miRNA prediction algorithms designed by different groups, their accuracies are not satisfactory and have not been properly validated in a comparative manner on the same reference data set. We therefore reviewed all previously described features for miRNA prediction (200+) and combined them with the features that we defined, finally selecting only the ones that are most informative without being correlated.

CHAPTER 2

MATERIALS AND METHODS

2.1. Data sets

Positive examples for human miRNAs were obtained from miRBase (http://www.mirbase.org/). Also, other positive data sets have been prepared by using available examples for human from miRTarBase and from Ensemble (Flicek et al., 2013). For negative examples we prepared different sets;

- Pseudo: the widely used pseudo hairpin sequences by Ng and Mishra (Ng & Mishra, 2007)
- ShuffledHuman: random sequences with the same length range with human miRNAs
- NegHSa: to be able to examine class imbalance effect, the largest available negative dataset for human was used (http://adaa.polsl.pl/agudys/huntmi/huntmi.htm) (Gudyś, Szcześniak, Sikora, & Makałowska, 2013).
- NotBestFold: we defined a completely new negative data constructed by choosing not best structures for known human miRNAs. Since the sequence is not altered this negative data would show the effect of highly used structural features on the classification accuracy.

2.2. Features

Selected features are of major importance for the accuracy and generalization of the model established through machine learning. In this case, we previously assessed 12 *ab initio* miRNA detection algorithms (Bentwich, 2008; Cakir & Allmer, 2010; Ding et al., 2010; Grundhoff, Sullivan, & Ganem, 2006; Jiang et al., 2007; Lai, Tomancak, Williams, & Rubin, 2003; Ng & Mishra, 2007; Pfeffer et al., 2005; Ritchie, Gao, & Rasko, 2012; van der Burgt et al., 2009; Y. Xu, Zhou, & Zhang, 2008; Xue et al., 2005) and determined the features that were used to describe a miRNA hairpin. More than 200

different features have been described and more than 100 have been used in machine learning for miRNA hairpin detection. In this study, in addition to implementing the features in the Java™ programming language, we generalized and normalized them to miRNA stem length or miRNA hairpin length where appropriate. The total number of features that we analyzed amounts to 740. We divided the features into four groups as sequence, structure, thermodynamic and probability based features. Some examples 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), inverted internal repeat (ir), free energy (hpmfe_rf), GC content (%GC) etc.

Structural features; 32 triplet elements i.e. A(((, A…, U(((, U(.(, U…, G(((, C(((, C(.(, hairpin length (hpl), loop length (hll), free energy per nucleotide (hpmfe_rf/hpl), matching base pairs (bpp), maximal bulge size (mbs) etc.

Thermodynamics based features; ensemble free energy (efe), ensemble frequency (efq), melting temperature (Tm), enthalpie divided by hairpin length (dH/hpl), entropy divided by hairpin length (dS/hpl), etc.

Probability based features derived from dinucleotide shuffling (dns); adjusted base pairing propensity (dP), adjusted Minimum Free Energy of folding (MFE) (dG), MFE index 1 (MFEI1), adjusted base pair distance (dD), adjusted shannon entropy (dQ), MFE index 2 (MFEI2), degree of compactness (dF).

2.3. Machine Learning

There are many data mining tools and most of them were used during the process at some points but Orange Canvas (Curk et al., 2005) was mainly used to perform filtering, sampling, ranking and data preprocessing steps, as well as to carry out SVM, Naive Bayes (NB), Random Forest (RF) and Logistic Regression (LR) classifications with 10 fold cross-validation. All the classifiers are used with their default settings.

After obtaining data from MiRBase, we needed to calculate the values of our features on this data as well as negative data sets. Subsequently running Java codes prepared specifically for this purpose, we had the desired outputs which would allow us to carry out a classification. Initial filtering on MiRBase data was achieved by removing

entries having multiple loops in structure or having more than two cleave site problems in either 3' or 5' ends. Later, both positive and negative data were loaded into Orange Canvas, stratified (if possible) random sampling applied according to the expected output (the same size, 10-fold validation etc.) combined and saved (*Figure* 6). The next step was the classification process for acquiring the best model capable of distinguishing the positive from negatives (*Figure* 7).

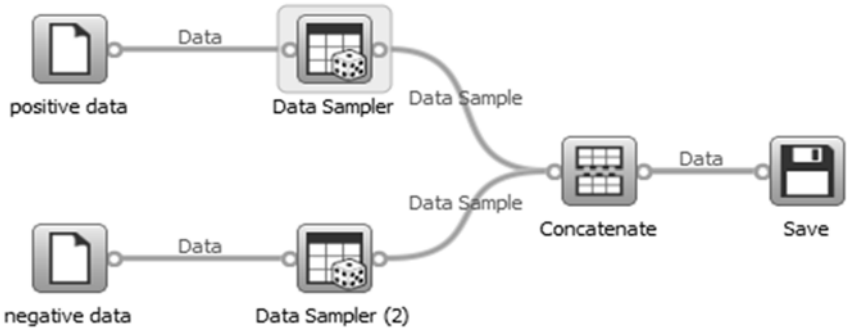

Figure 6. Sampling of data. Stratified (if possible) random sampling applied on data sets. According to the expected results Data Sampler node's parameters are used, e.g. if 1:2 ratio is needed between data sets, the number of samples can be chosen in this part.

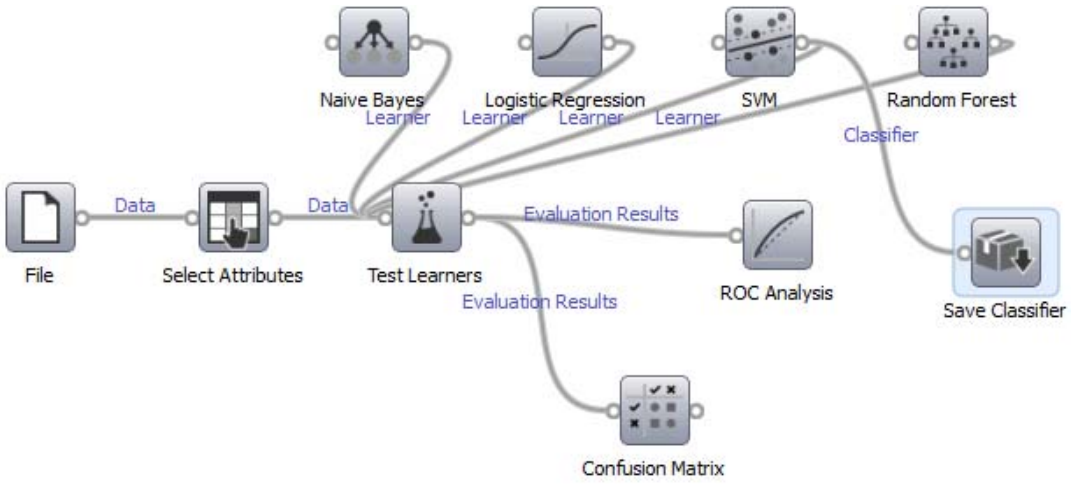

Figure 7. Classification with four different classifiers. After uploading the file including the data to be classified, the attribute defining the classes is chosen in the Select Attributes node. Test Learners is applied with four classifiers NB, LR, SVM and RF and the classifier producing the best accuracy is saved for the model. The results of Test Learners can be evaluated as ROC Analysis and Confusion Matrix.

In statistical analysis of classification, the F-score (F_1) is a measure used to evaluate the test's accuracy. It takes into account both the precision (p, the number of correct results divided by the number of all returned results) and the recall (r, the number of correct results divided by the number of results that should have been returned) of the test to compute the score (*Table 3*). In other terms, the F_1 can be elucidated as the harmonic mean of the precision and recall (Powers, 2011). Although, there are various formulas for accuracy calculation, we measured it as the proportion of true results in the overall outcome.

Table 3. Statistical measures

		Condition		
		Condition Positive	Condition Negative	
Test Outcome	Test Outcome Positive	True Positive	False Positive	Precision = $\frac{\Sigma\, True\ Positive}{\Sigma\, Test\ Outcome\ Positive}$
	Test Outcome Negative	False Negative	True Negative	Negative predictive value = $\frac{\Sigma\, True\ Negative}{\Sigma\, Test\ Outcome\ Negative}$
		Sensitivity (recall) = $\frac{\Sigma\, True\ Positive}{\Sigma\, Condition\ Positive}$	Specificity = $\frac{\Sigma \text{ True Negative}}{\Sigma \text{ Condition Negative}}$	Accuracy

2.4. Programs Included in the System

Even though we have designed and implemented the proposed method, we have also used some external code, especially during the calculation of features. Pre-implemented programs in our approach are RNAshapes (Giegerich, Voss, & Rehmsmeier, 2004) and RNAfold (Hofacker, 2003) for folding RNA sequences into stem-loop and hairpin structures and obtaining the minimum free energy scores. In addition, mfold (Zuker, 2003) was used for calculating features such as Gibbs free energy, entropy etc. DustMasker was used from the Blast+ package (Camacho C, Madden T, Ma N, et al. 2008), RNAEval was applied from Vienna RNA Secondary Structure package (Hofacker, 2003) and dc was accessed using an online server (http://www.biomath.nyu.edu/rag/rna_matrix_results.php).

CHAPTER 3

RESULTS AND DISCUSSION

3.1. Comparison of Four Tools

There are many the *ab inito* algorithms for miRNA gene prediction and most of them report an accuracy but these tools cannot be compared easily since they are trained on different data sets. Besides, most of the algorithms are not distributed publicly. Thus, we implemented the routines for calculating all features described in Ding et al., Jiang et al, Ng and Mishra, and Bentwich (Bentwich, 2008; Ding et al., 2010; Jiang et al., 2007; Ng & Mishra, 2007) and compared them on the same data set to investigate relative algorithm performance.

MiRBase human entries were used as positive data. A filtration procedure was applied as; we removed the samples that contained more than one hairpin when folded by RNAFold (Hofacker, 2003) or RNAShapes (Steffen, Voss, Rehmsmeier, Reeder, & Giegerich, 2006), and if there was no accurate link to Ensemble, those entries were removed as well. From the remaining samples, about 1000 miRNA examples we distributed into five random subsets containing 500 positive examples each. As for the negative data sets, we used ShuffledHuman and Pseudo (see *2.1 Data Sets*).

We generated five combined data sets comprising 60% training and 40% test samples from the overall data set which was used to train and test a SVM classifier using Orange Canvas. In order to have a fair comparison between these four studies (Bentwich, 2008; Ding et al., 2010; Jiang et al., 2007; Ng & Mishra, 2007), we needed to use this approach since fivefold cross validation could not be used with multiple studies at the same time; but had to be repeated individually, thus leading to different data sets.

We think that, the two datasets that were used (positive data – ShuffledHuman and positive data - Pseudo), are of different difficulty with the ShuffledHuman dataset being easier to solve than the Pseudo miRNA data set. This was also supported from the best results reported in Table *4* and Table *5*. The results for the ShuffledHuman - miRNAs in Table *4* lead to higher accuracy than the data in Table *5* which is achieved with Pseudo

- miRNAs. For both tables the best and the average accuracy are presented along with the standard deviation, calculated from fivefold cross validation.

Table 4. Accuracy of known human miRNAs and ShuffledHuman dataset.

Studies	Accuracy Values		
	Best	Average	Standard Deviation
(Ng & Mishra, 2007)	0.919	0.894	0.183
(Bentwich, 2008)	1.000	1.000	0
(Ding et al., 2010)	1.000	0.676	0.217
(Jiang et al., 2007)	0.954	0.952	0.003

Such a perfect result as achieved by Bentwich 2008 features in Table *4* is not expected for the pseudo miRNAs and Table *5* displays no such success. For the Pseudo miRNA dataset, the features used in (Ding et al., 2010) achieve the highest accuracy although with a high standard deviation among training sets.

Table 5. Accuracy of known human miRNAs and Pseudo dataset.

Studies	Accuracy Values		
	Best	Average	Standard Deviation
(Ng & Mishra, 2007)	0.930	0.895	0.060
(Bentwich, 2008)	0.986	0.983	0.002
(Ding et al., 2010)	0.996	0.599	0.198
(Jiang et al., 2007)	0.910	0.877	0.018

The best accuracy of 0.996, achieved by the features described in the (Ding et al., 2010) study, when used for 10 million putative hairpins in human would lead to 40000 false positive detections. Unfortunately, the number of putative hairpins in human is estimated to be quite large and the accuracy calculated here does not completely reflect the true accuracy since it is not entirely known what differentiates a true from a false hairpin. For real data, a much higher false positive rate is expected for real data and thus the high number of false positives may increase the cost of experimental validation of all predicted miRNAs drastically, making it almost impossible. This is further supported by showing that the accuracy strongly depends on the data set used for training and testing Figure *8* and Figure *9*.

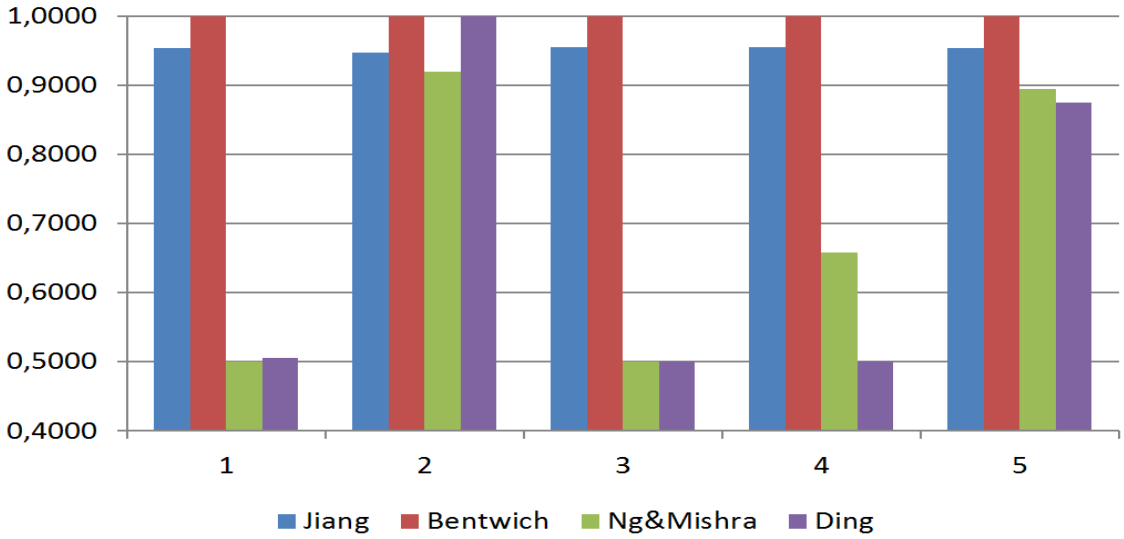

Figure 8. Accuracy of human miRNAs and ShuffledHuman miRNAs. All cross validation results are shown individually.

We considered Pseudo data set to be a harder negative data to handle for the classifier. Although (Ding et al., 2010) achieves the highest accuracy in one case, it fails in all other cases which indicates a strong dependence on the training and test data set and a poor generalization for the features from that study (Figure *9*). (Bentwich, 2008) does not show such generalization problems and outperforms all other studies on the remaining four data sets.

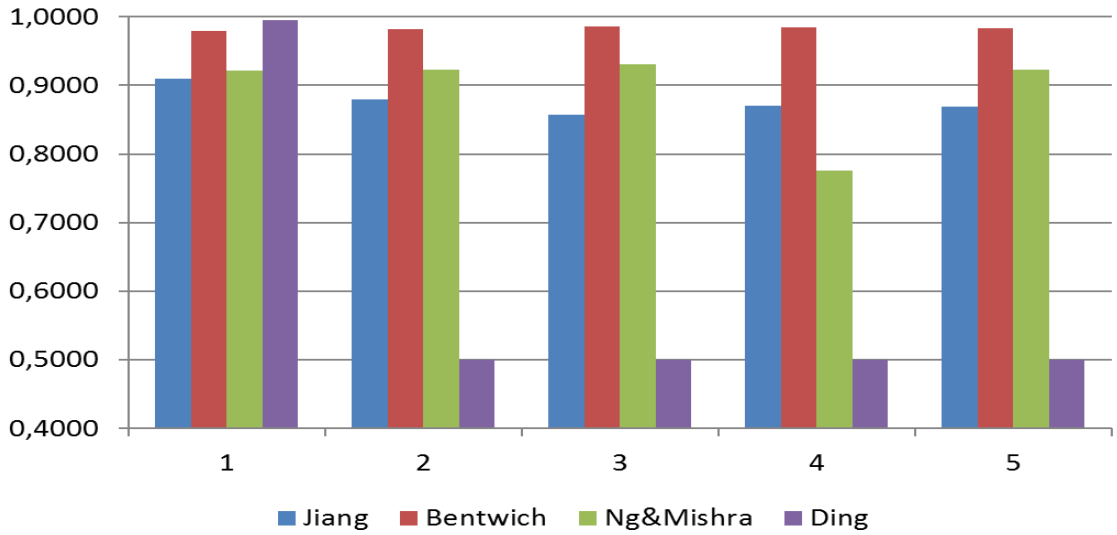

Figure 9. Accuracy of human miRNAs and Pseudo miRNAs. All cross validation results are shown individually.

As we expected, our results show that Pseudo data set is more difficult than ShuffledHuman negative data set. This directed us to think about the quality of the positive data used for classifications so we tested the quality of the MirBase which is the *de facto* standard repository for miRNAs.

3.2. Testing MiRBase

With the aim of analyzing whether entries in miRBase that are falsely annotated as miRNAs have an influence on classification accuracy, we performed classifications with three classifiers. Given that miRTarBase (Hsu et al., 2011) has only little data for human we used 180 available examples for the positive dataset from miRTarBase which claims that this data has strong experimental evidence showing miRNA-target interactions. MiRBase contains about 2000 human miRNAs, but in an attempt to keep the comparison fair, we randomly extracted 180 sequences. This guides to two positive datasets which both consist of 180 examples one from miRBase and one from miRTarBase which will be compared in. Both positive datasets were used in conjunction with the same negative dataset (pseudo hairpins) in order to ensure fair comparison.

In this study, we used the 10 most frequently used features in the 12 *ab initio* miRNA hairpin detection studies. The selected parameters are: hairpin loop length (hll), base pairing propensity (bpp), hairpin minimum free energy (hpmfe), dinucleotide

shuffling, p-value of hpmfe, and the frequencies of the following triplet structures U(((, U(.(, C(.(, A..., G(((.

Table 6. Comparing miRBase and miRTarBase as a positive data source.

	miRBase entries				**Proven miRNAs**			
Classifier	Sensitivity	Specificity	CA	F1	Sensitivity	Specificity	CA	F1
SVM	0.85	0.86	0.85	0.85	0.94	0.92	0.93	0.93
NB	0.90	0.82	0.86	0.87	0.93	0.90	0.91	0.92
LR	0.91	0.92	0.92	0.92	0.93	0.94	0.94	0.94

Table 6 shows, the classifications of miRNA hairpins and pseudo hairpins using SVM, Naïve Bayes (NB), and Logistic Regression (LR) to compare performance of miRBase as positive data and miRTarBase as positive data. The results of classifications for all three employed classifiers show that using the miRNAs having strong experimental evidence to interact with an mRNA as positive dataset (miRTarBase) provides a higher sensitivity, specificity, classification accuracy (CA), and F1 (calculation method as in Table *3*). Using positive examples derived from miRBase on the other hand leads to lower statistics (*Table 6*).

3.3. Class Imbalance

Here we investigate the impact of largely differing numbers of positive and negative examples on machine learning for *ab initio* miRNA prediction.

Positive examples for human miRNAs were obtained from miRBase. To examine class imbalance effect, the largest available negative dataset for human was used (Gudyś et al., 2013). Of this dataset about 50000 sequences were used for testing. For generalization the remaining about 18000 examples from the dataset were used. In addition to that, miRNA examples from ENSEMBLE (≈3200) (Flicek et al., 2013), the complete pseudo data set (≈9000) as described in (Ng & Mishra, 2007), and random sequences with the same length range with human miRNAs (1400) were used. The generalization dataset consists of approximately 32000 examples with most of them being negative.

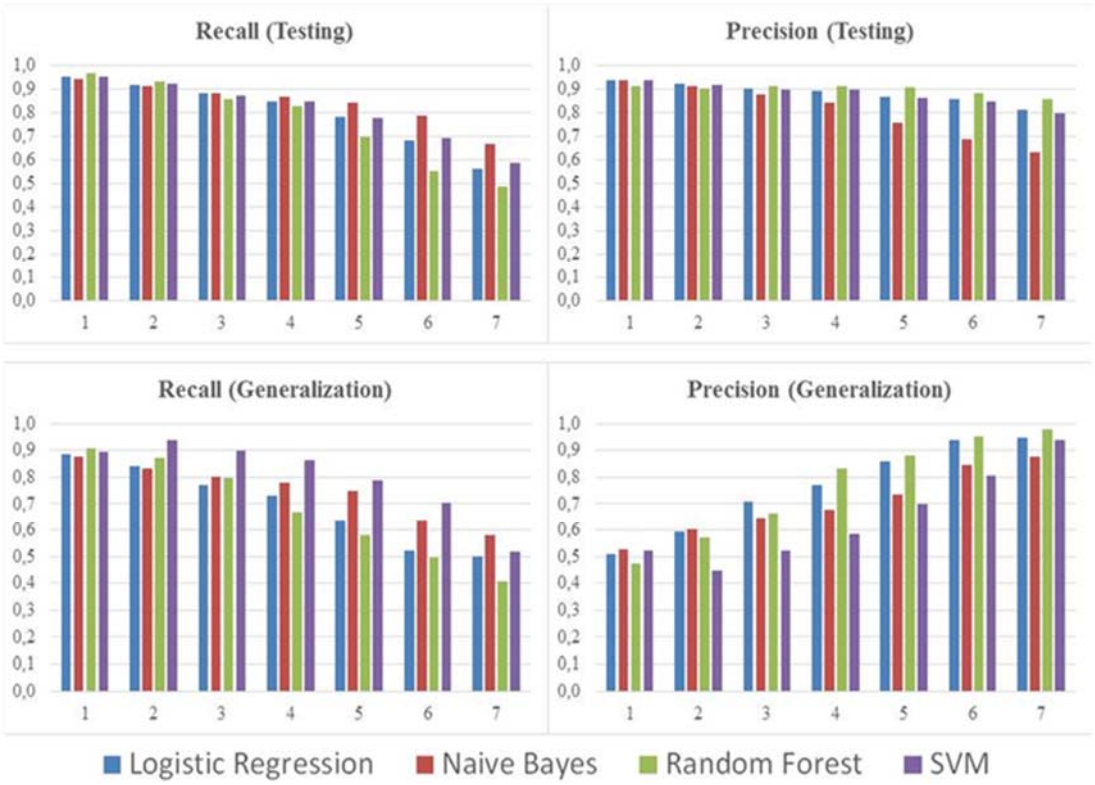

Figure 10. Class imbalance influence results. Recall and precision after training on the test dataset using 10 fold cross validation for four different classifiers (top left and right). Recall and precision for the best classifier tested on new examples that were not part of training or test set (bottom left and right). Numbers 1 through 7 correspond to different composition of positive and negative data during training. 1:1600 positive and 800 negative examples; 2 1600, 1600; 3: 1600-3200; 4: 1600-5000; 5: 1600-10000; 6: 1600-25000; 7: 1600-50000.

According to the results in *Figure 10*, during generalization (bottom panes) recall drops significantly (up to ~50%) while precision increases at the same time (up to ~50%). This implies that accuracy may not be a good measure to rely on for the performance of miRNA gene prediction based on machine learning and that it is a better choice to report precision and recall or sensitivity and specificity, instead.

3.4. Feature Number

Since we implemented a large number of features, we were interested to see whether increasing number of features on the best training dataset from Figure 10 would influence classification performance. Firstly, we ranked all features according to

information gain (Table *7*) and then started with the 5 most extreme features and included more features until all features were included in the classification. This leads to two curves for precision and recall (Figure *11*) one set which starts from low performance and one which starts with high performance. When features with the least information gain from the bottom are selected first, the trained classifiers never reach to recall or precision of classifiers trained starting with features with the largest information gain. Increasing number of features, makes the precision and recall values of the trained classifiers fall into the same range. They cannot be expected to reach to exactly the same values as an element of randomness is introduced through the use of 10 fold cross validation during classification.

Table 7. Feature ranking. Information gain for the 100 features with highest gain among all defined features.

Attribute	**Inf. gain**	**Attribute**	**Inf. gain**	**Attribute**	**Inf. gain**	**Attribute**	**Inf. gain**
#C.../sl	1,00	#U(./sl	1,00	st(G-C)/hpl	0,72	nl/sl	0,57
#U.../sl	1,00	#G../sl	1,00	l(lsr)/hpl	0,70	nl/hpl	0,56
#A../sl	1,00	#A(./sl	1,00	st(G-C)/sl	0,69	lscm/hpl	0,54
#A(../sl	1,00	#G(./sl	1,00	bpp/hpl	0,69	st(A-U)/hpl	0,52
#G(((/sl	1,00	Q	1,00	mwm/sl	0,69	mwmF/hpl	0,51
#A(((/sl	1,00	#C(./sl	0,99	bpp/nl	0,68	st(A-U)/sl	0,50
#U(((/sl	1,00	Tm	0,99	hpmfe_rf	0,68	*G(((	0,48
#U../sl	1,00	#U(./sl	0,99	l(lsr)/sl	0,68	*A...	0,47
#G.../sl	1,00	#A(./sl	0,98	dscs/nl	0,67	#A++#U/hpl	0,47
#C..(/sl	1,00	dH/sl	0,96	efq	0,67	#U++#A/hpl	0,47
#C((./sl	1,00	dS/sl	0,96	ediv	0,66	%U++%A	0,47
#A.../sl	1,00	dS/hpl	0,95	saln/hpl	0,66	%A++%U	0,47
#A((./sl	1,00	dH/hpl	0,95	bpp/sl	0,66	#G++#C/hpl	0,47
#U.((/sl	1,00	Tm/sl	0,94	mbs/sl	0,66	#C++#G/hpl	0,47
#U(../sl	1,00	hpmfe_rf/sl	0,94	mbs/hpl	0,66	%C++%G	0,47
#C(((/sl	1,00	hpmfe_rf_I1	0,94	lsr(%bp)/hpl	0,64	%G++%C	0,47
#U((./sl	1,00	dG/sl	0,93	lsr(%bp)/sl	0,62	#U++#A/sl	0,46
#G(../sl	1,00	hpmfe_rf/hpl	0,91	#nial_h/sl	0,61	*C(((	0,46
#C(./sl	1,00	Q/sl	0,91	#nial_h/hpl	0,61	#C++#G/sl	0,45
#C(../sl	1,00	dG/hpl	0,89	lscm/nl	0,60	adalr/hpl	0,45
#C.((/sl	1,00	efe	0,88	adal/hpl	0,59	%G-U	0,45
#G((./sl	1,00	Tm/hpl	0,88	bpp/saln	0,59	nl	0,44
#A.((/sl	1,00	bpd/sl	0,86	#gih/saln	0,58	#nisl_h/hpl	0,44
#G.((/sl	1,00	Q/hpl	0,81	#goh/saln	0,58	c#Ns/saln	0,43
#G(./sl	1,00	bpd/hpl	0,74	asal/hpl	0,58	#nisl_h/sl	0,43

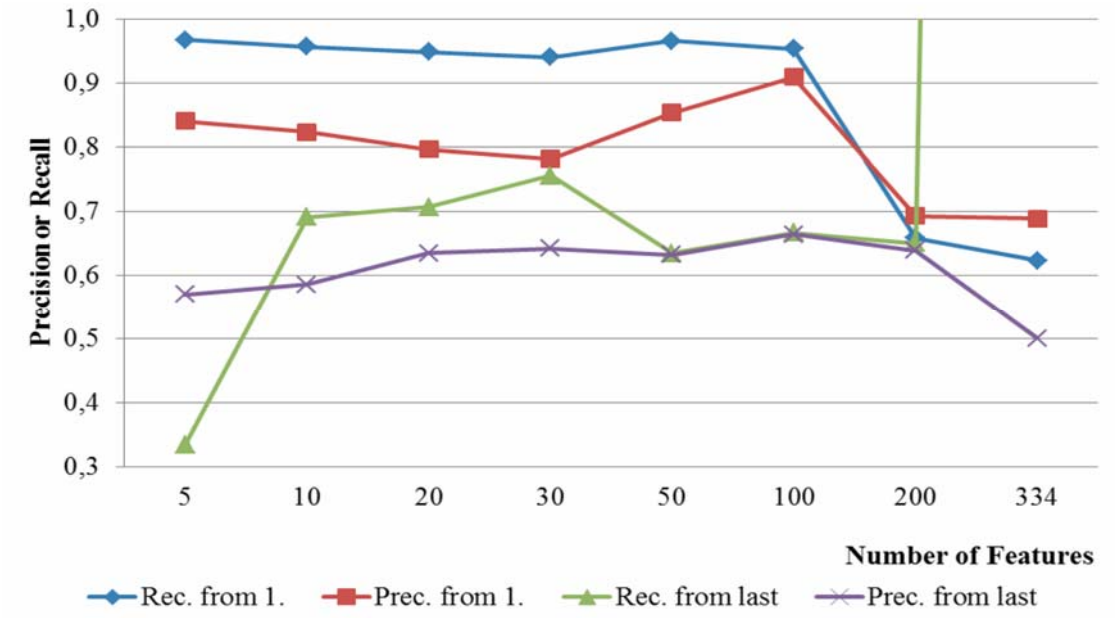

Figure 11. Feature number influence on classifier performance.

In Figure *11* "Rec. from 1." refers to recall for classifiers (in this case Naïve Bayes) trained with increasing number of features selected from the top of the list of ranked features according to information gain. "Prec. from 1." refers to the precision of classifiers trained in that manner. Rec. and Prec. from last refer to the recall and precision of classifiers with feature number increasing from the bottom of the ranked feature list. The calculation of the recall for 334 features from last was unsuccessful and therefore the value is out of bounds and not displayed in the figure.

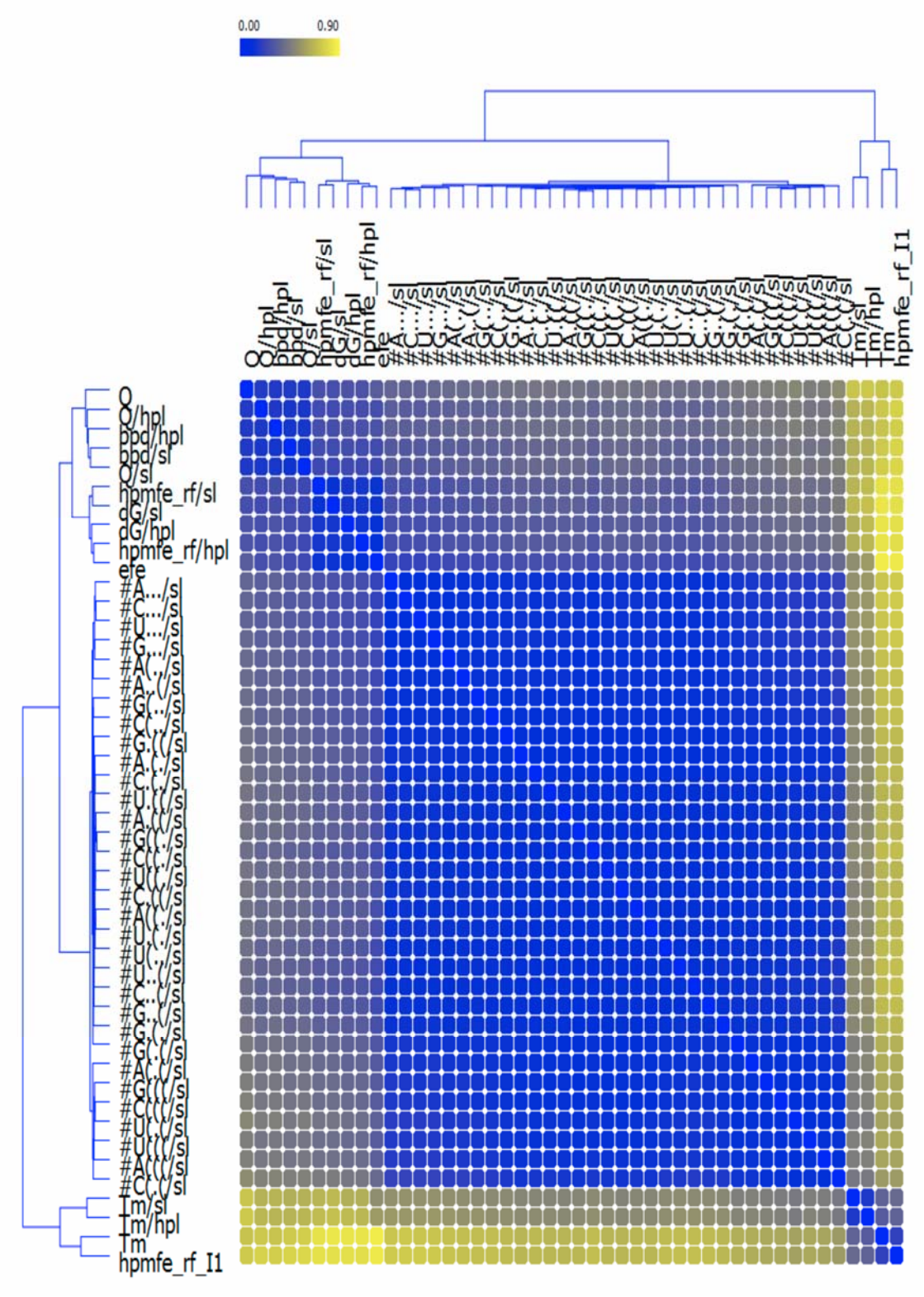

Figure 12. Correlation among the features with the largest information gain (see *Table 7*). Blue indicates high and yellow low correlation among the features. Features are further clustered hierarchically across the top and on the left hand side.

When we were dealing with such high numbers of features, another matter that came to mind was correlation among features *(Figure 12*). By solving this problem it is also possible to reduce the complexity of feature selection problem since by removing the highly correlated features we would also be reducing the number of features that can be selected from. In this case, we used attribute correlation as implemented in Orange Canvas, but it turns out that some of the features which are biologically strongly correlated were not reported to be correlated by the given method. This is also true for many of the features and their normalized versions e.g. Tm, Tm/sl, Tm/sl (*Figure 12*). Thus, we plan to manually review all attributes in the future and first remove all obviously correlated features before we try any other feature selection algorithm.

Based on these findings we suggest that in future studies the maximum number of features should not be more than 100 and it is also important to take into account that these features should not be highly correlated.

3.5. Classification

Among the all parameters available, 24 of them with the high information gain are selected; 10 triplet structures count normalized to stem length [#C.../sl, #U.../sl, #A..(/sl, #A(../sl, #G(((/sl, #A(((/sl, #U(((/sl, #U..(/sl, #G.../sl, #C..(/sl], Shannon Entropy [Q], melting temperature [Tm], Entropy normalized to hairpin length [dS/hpl], Enthalpy [dH/hpl], melting temperature normalized to stem length [Tm/sl], hairpin minimum free energy calculated by RNAfold normalized to stem length [hpmfe_rf/sl], minimum free energy index [hpmfe_rf_I1], Gibbs free energy normalized to stem length [dG/sl], hairpin minimum free energy calculated by RNAfold normalized to hairpin length [hpmfe_rf/hpl], Gibbs free energy normalized to hairpin length [dG/hpl], Ensemble Free Energy [efe], base pairing distance normalized o stem length [bpd/sl], GC in stem normalized to hairpin length [st(G-C)/hpl], length calculated over the longest symmetrical region (lsr) of the stem loop, i.e. the longest region without any asymmetrical loop normalized to hairpin length [l(lsr)/hpl].

After filtering the latest version of MiRBase for human miRNA hairpins by removing the entries having more than 2 overlaps in either 5' or 3' ends, we obtained 1196 samples out of 1872. All negative data sets mentioned before were used in this part and we decided to use the same size for both data sets after our observations in class

imbalance section. Sampling of negative data sets were achieved by using stratified random sampling in Orange Canvas.

Table 8. Classification results.

Classifier	CA	Sensitivity	Specificity	AUC	F1	Negative data sets
NB	0.9983	0.9983	0.9983	1	0.9983	NegHsa
LR	1	1	1	1	1	NegHsa
SVM	1	1	1	1	1	NegHsa
RF	1	1	1	1	1	NegHsa
NB	0.9398	0.9231	0.9565	0.9796	0.9550	ShuffledHuman
LR	0.9469	0.9691	0.9247	0.9893	0.9279	ShuffledHuman
SVM	0.9519	0.9440	0.9599	0.9898	0.9592	ShuffledHuman
RF	0.9603	0.9557	0.9649	0.9889	0.9646	ShuffledHuman
NB	0.8520	0.8344	0.8696	0.9251	0.8648	Pseudo
LR	0.8834	0.8487	0.9181	0.9545	0.9119	Pseudo
SVM	0.8800	0.8620	0.8980	0.9485	0.8942	Pseudo
RF	0.8821	0.8771	0.8871	0.9512	0.8860	Pseudo
NB	0.5084	0.5460	0.4707	0.5137	0.5078	NotBestFold
LR	0.5080	0.4724	0.5435	0.5255	0.5086	NotBestFold
SVM	0.5381	0.5769	0.4992	0.5357	0.5353	NotBestFold
RF	0.3344	0.3896	0.2793	0.2483	0.3509	NotBestFold

From the results shown in *Table 8* we can conclude that all of the classifiers were able to differentiate between miRNAs (positive) and NegHsa (negative) samples and have a perfect accuracy, while in NotBestFold negative dataset, NB, SVM and LR made almost random classifications and RF was even worse than arbitrary.

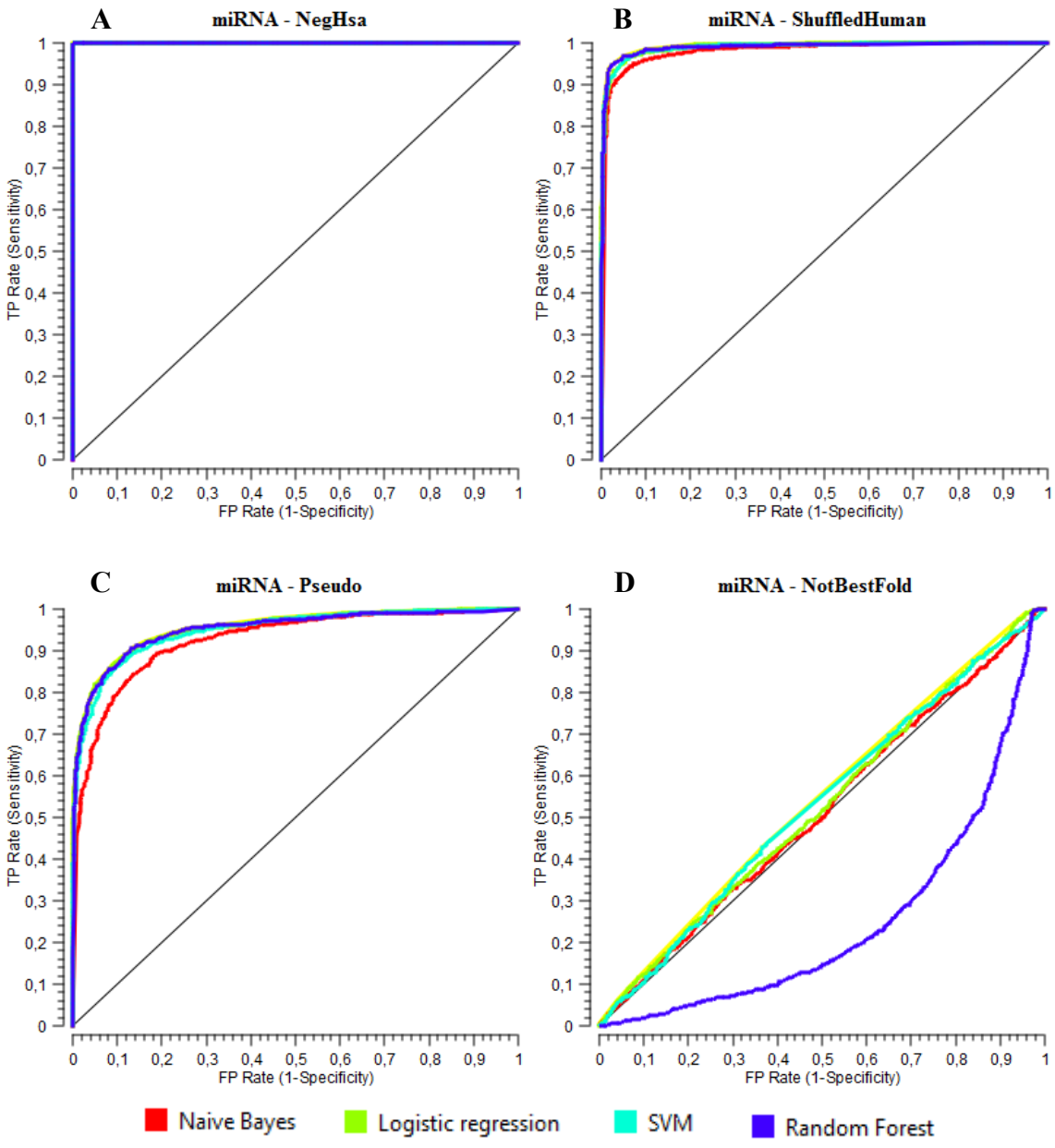

Figure 13. ROC analysis for four different classifiers on different data sets. Best accuracy is observed in miRNA – NegHsa data set (A) while the worst one is found in miRNA – NotBestFold (D).

CHAPTER 4

CONCLUSION

Regardless of the method chosen, the most vital challenge for miRNA gene prediction is the existence of high numbers of inverted repeats (IR) in most eukaryotic genomes making the transcripts of these IRs capable of forming strong hairpins (Lindow & Gorodkin, 2007). It has been indicated that there are about 11 million potential hairpins in the human genome (Bentwich et al., 2005) and these hairpins might originate from any part of the genome and take part in numerous processes, one of which may possibly be miRNA-mediated posttranscriptional regulation (Lindow & Gorodkin, 2007). Since not all hairpins are miRNAs, not only identifying the hairpins which would become functional miRNAs is a very difficult task but also the big number of possible hairpins makes increasing the accuracy of the prediction another challenge.

In recent years, machine learning approaches have become a widely used method for miRNA gene prediction studies. The requirements of machine learning methods such as positive datasets and parameters are available for miRNAs, since there are known miRNAs either experimentally validated or discovered through bioinformatics tools and there are also some rules defining a sequence as a miRNA (e.g., recognition and being processed by miRNA biogenesis pathway enzymes such as Dicer and Drosha), so the sequences that do not pass this criteria can be used as negative datasets. The abundance of machine learning processes used for miRNA gene prediction indicates that these methods are considered to be appropriate to handle this problem.

The main objective of this study was to form an integrative data mining approach which would include all the requirements to perform an efficient classification process to be able to detect possible miRNA hairpins in human data. At the end, we were able to achieve more than 95% accuracy, sensitivity, specificity, recall and precision values. With the help of this established approach, we will be capable of identifying potential miRNAs in any genome in future studies.

REFERENCES

Ambros, V., Lee, R. C., Lavanway, A., Williams, P. T., & Jewell, D. (2003). MicroRNAs and Other Tiny Endogenous RNAs in C . elegans, *13*, 807–818. doi:10.1016/S

Artzi, S., Kiezun, A., & Shomron, N. (2008). miRNAminer: a tool for homologous microRNA gene search. *BMC bioinformatics*, *9*(6), 39. doi:10.1186/1471-2105-9-39

Bar, M., Wyman, S. K., Fritz, B. R., Qi, J., Garg, K. S., Parkin, R. K., … Tewari, M. (2008). MicroRNA discovery and profiling in human embryonic stem cells by deep sequencing of small RNA libraries. *Stem cells (Dayton, Ohio)*, *26*(10), 2496–505. doi:10.1634/stemcells.2008-0356

Beezhold, K. J., Castranova, V., & Chen, F. (2010). Microprocessor of microRNAs: regulation and potential for therapeutic intervention. *Molecular cancer*, *9*, 134. doi:10.1186/1476-4598-9-134

Bentwich, I. (2008). Identifying human microRNAs. *Current topics in microbiology and immunology*, *320*, 257–69. http://www.ncbi.nlm.nih.gov/pubmed/18268848

Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., … Bentwich, Z. (2005). Identification of hundreds of conserved and nonconserved human microRNAs. *Nature genetics*, *37*(7), 766–70. doi:10.1038/ng1590

Berezikov, E., Cuppen, E., & Plasterk, R. H. A. (2006). Approaches to microRNA discovery. *Nat Genet*. 38, S2 - S7. doi:10.1038/ng1794

Bhaskar, H., Hoyle, D. C., & Singh, S. (2006). Machine learning in bioinformatics: a brief survey and recommendations for practitioners. *Computers in biology and medicine*, *36*(10), 1104–25. doi:10.1016/j.compbiomed.2005.09.002

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K. D., Ovcharenko, I., Pachter, L., & Rubin, E. M. (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science (New York, N.Y.)*, *299*(5611), 1391–4. doi:10.1126/science.1081331

Brameier, M., & Wiuf, C. (2007). Ab initio identification of human microRNAs based on structure motifs. *BMC bioinformatics*, *8*, 478. doi:10.1186/1471-2105-8-478

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

Cakir, M. V., & Allmer, J. (2010). Systematic computational analysis of potential RNAi regulation in Toxoplasma gondii. *2010 5th International Symposium on Health Informatics and Bioinformatics*, 31–38. doi:10.1109/HIBIT.2010.5478909

Camacho C, Madden T, Ma N. (2008). BLAST+ Command Line Applications User Manual. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK1763/

Chapman, E. J., & Carrington, J. C. (2007). Specialization and evolution of endogenous small RNA pathways. *Nat Rev Genet*, *8*(11), 884–896. Retrieved from http://dx.doi.org/10.1038/nrg2179

Curk, T., Demsar, J., Xu, Q., Leban, G., Petrovic, U., Bratko, I., … Zupan, B. (2005). Microarray data mining with visual programming. *Bioinformatics (Oxford, England)*, *21*(3), 396–8. doi:10.1093/bioinformatics/bth474

Deo, M., Yu, J.-Y., Chung, K.-H., Tippens, M., & Turner, D. L. (2006). Detection of mammalian microRNA expression by in situ hybridization with RNA oligonucleotides. *Developmental dynamics : an official publication of the American Association of Anatomists*, *235*(9), 2538–48. doi:10.1002/dvdy.20847

Ding, J., Zhou, S., & Guan, J. (2010). MiRenSVM: towards better prediction of microRNA precursors using an ensemble SVM classifier with multi-loop features. *BMC bioinformatics*, *11 Suppl 1*(Suppl 11), S11. doi:10.1186/1471-2105-11-S11-S11

Filipowicz, W., Bhattacharyya, S. N., & Sonenberg, N. (2008). Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nature reviews. Genetics*, *9*(2), 102–14. doi:10.1038/nrg2290

Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., … Fairley, S. (2013). Ensembl 2013. *Nucleic acids research*, *41*(Database issue), D48–55. doi:10.1093/nar/gks1236

Friedländer, M. R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., & Rajewsky, N. (2008). Discovering microRNAs from deep sequencing data using miRDeep. *Nature biotechnology*, *26*(4), 407–15. doi:10.1038/nbt1394

Gerlach, D., Kriventseva, E. V, Rahman, N., Vejnar, C. E., & Zdobnov, E. M. (2009). miROrtho: computational survey of microRNA genes. *Nucleic acids research*, *37*(Database issue), D111–7. doi:10.1093/nar/gkn707

Giegerich, R., Voss, B., & Rehmsmeier, M. (2004). Abstract shapes of RNA. *Nucleic acids research*, *32*(16), 4843–51. doi:10.1093/nar/gkh779

Gkirtzou, K., Tsamardinos, I., Tsakalides, P., & Poirazi, P. (2010). MatureBayes: a probabilistic algorithm for identifying the mature miRNA within novel precursors. *PloS one*, *5*(8), e11843. doi:10.1371/journal.pone.0011843

Gomes, C. P. C., Cho, J.-H., Hood, L., Franco, O. L., Pereira, R. W., & Wang, K. (2013). A Review of Computational Tools in microRNA Discovery. *Frontiers in genetics*, *4*(May), 81. doi:10.3389/fgene.2013.00081

Grundhoff, A., Sullivan, C. S., & Ganem, D. O. N. (2006). A combined computational and microarray-based approach identifies novel microRNAs encoded by human gamma-herpesviruses, 733–750. doi:10.1261/rna.2326106.3

Gudyś, A., Szcześniak, M. W., Sikora, M., & Makałowska, I. (2013). HuntMi: an efficient and taxon-specific approach in pre-miRNA identification. *BMC bioinformatics*, *14*, 83. doi:10.1186/1471-2105-14-83

Guerra-Assunção, J. A., & Enright, A. J. (2010). MapMi: automated mapping of microRNA loci. *BMC bioinformatics*, *11*, 133. doi:10.1186/1471-2105-11-133

Guzman, F., Almerão, M. P., Körbes, A. P., Loss-Morais, G., & Margis, R. (2012). Identification of microRNAs from Eugenia uniflora by high-throughput sequencing and bioinformatics analysis. *PloS one*, *7*(11), e49811. doi:10.1371/journal.pone.0049811

Hackenberg, M., Sturm, M., Langenberger, D., Falcón-Pérez, J. M., & Aransay, A. M. (2009). miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic acids research*, *37*(Web Server issue), W68–76. doi:10.1093/nar/gkp347

He, X., Zhang, Q., Liu, Y., & Pan, X. (2007). Cloning and Identification of Novel MicroRNAs from Rat Hippocampus. *Acta Biochimica et Biophysica Sinica*, *39*(9), 708–714. doi:10.1111/j.1745-7270.2007.00324.x

Hébert, S. S., Horré, K., Nicolaï, L., Bergmans, B., Papadopoulou, A. S., Delacourte, A., & De Strooper, B. (2009). MicroRNA regulation of Alzheimer's Amyloid precursor protein expression. *Neurobiology of disease*, *33*(3), 422–8. doi:10.1016/j.nbd.2008.11.009

Heikkinen, L., Kolehmainen, M., & Wong, G. (2011). Prediction of microRNA targets in Caenorhabditis elegans using a self-organizing map. *Bioinformatics (Oxford, England)*, *27*(9), 1247–54. doi:10.1093/bioinformatics/btr144

Hertel, J., & Stadler, P. F. (2006). Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data. *Bioinformatics (Oxford, England)*, *22*(14), e197–202. doi:10.1093/bioinformatics/btl257

Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Research*, *31*(13), 3429–3431. doi:10.1093/nar/gkg599

Hsu, S.-D., Lin, F.-M., Wu, W.-Y., Liang, C., Huang, W.-C., Chan, W.-L., … Huang, H.-D. (2011). miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic acids research*, *39*(Database issue), D163–9. doi:10.1093/nar/gkq1107

Huang, T.-H., Fan, B., Rothschild, M. F., Hu, Z.-L., Li, K., & Zhao, S.-H. (2007). MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. *BMC bioinformatics*, *8*, 341. doi:10.1186/1471-2105-8-341

Hutvágner, G., McLachlan, J., Pasquinelli, a E., Bálint, E., Tuschl, T., & Zamore, P. D. (2001). 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. doi:10.1126/science.1062961

Ibáñez-Ventoso, C., Vora, M., & Driscoll, M. (2008). Sequence relationships among C. elegans, D. melanogaster and human microRNAs highlight the extensive conservation of microRNAs in biology. *PloS one*, *3*(7), e2818. doi:10.1371/journal.pone.0002818

Jiang, P., Wu, H., Wang, W., Ma, W., Sun, X., & Lu, Z. (2007). 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. doi:10.1093/nar/gkm368

Jones-Rhoades, M. W., Bartel, D. P., & Bartel, B. (2006). MicroRNAS and their regulatory roles in plants. *Annual review of plant biology*, *57*, 19–53. doi:10.1146/annurev.arplant.57.032905.105218

Kadri, S., Hinman, V., & Benos, P. V. (2009). HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models. *BMC bioinformatics*, *10 Suppl 1*, S35. doi:10.1186/1471-2105-10-S1-S35

Khvorova, A., Reynolds, A., & Jayasena, S. D. (2003). Functional siRNAs and miRNAs exhibit strand bias. *Cell*, *115*(2), 209–16. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14567918

Kim, B., Yu, H.-J., Park, S.-G., Shin, J. Y., Oh, M., Kim, N., & Mun, J.-H. (2012). Identification and profiling of novel microRNAs in the Brassica rapa genome based on small RNA deep sequencing. *BMC plant biology*, *12*(1), 218. doi:10.1186/1471-2229-12-218

Kim, V. N., Han, J., & Siomi, M. C. (2009). Biogenesis of small RNAs in animals. *Nature reviews. Molecular cell biology*, *10*(2), 126–39. doi:10.1038/nrm2632

Kloosterman, W. P., Wienholds, E., de Bruijn, E., Kauppinen, S., & Plasterk, R. H. A. (2006). In situ detection of miRNAs in animal embryos using LNA-modified oligonucleotide probes. *Nat Meth*, *3*(1), 27–29. Retrieved from http://dx.doi.org/10.1038/nmeth843

Lagos-Quintana, M, Rauhut, R., Lendeckel, W., & Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. *Science (New York, N.Y.)*, *294*(5543), 853–8. doi:10.1126/science.1064921

Lagos-Quintana, Mariana, Rauhut, R., Yalcin, A., Meyer, J., Lendeckel, W., & Tuschl, T. (2002). Identification of tissue-specific microRNAs from mouse. *Current biology : CB*, *12*(9), 735–9. http://www.ncbi.nlm.nih.gov/pubmed/23719956

Lai, E. C., Tomancak, P., Williams, R. W., & Rubin, G. M. (2003). Computational identification of Drosophila microRNA genes. *Genome biology*, *4*(7), R42. doi:10.1186/gb-2003-4-7-r42

Larranaga, P. (2006). Machine learning in bioinformatics. *Briefings in Bioinformatics*, *7*(1), 86–112. doi:10.1093/bib/bbk007

Lau, N. C., Lim, L. P., Weinstein, E. G., & Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science (New York, N.Y.)*, *294*(5543), 858–62. doi:10.1126/science.1065062

Lee, R. C., Feinbaum, R. L., & Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, *75*(5), 843–54. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8252621

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., … Kim, V. N. (2003). The nuclear RNase III Drosha initiates microRNA processing. *Nature*, *425*(6956), 415–9. doi:10.1038/nature01957

Lee, Y., Jeon, K., Lee, J.-T., Kim, S., & Kim, V. N. (2002). MicroRNA maturation: stepwise processing and subcellular localization. *The EMBO journal*, *21*(17), 4663–70.http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=126204&tool=pmcentrez&rendertype=abstract

Lee, Y., Kim, M., Han, J., Yeom, K.-H., Lee, S., Baek, S. H., & Kim, V. N. (2004). MicroRNA genes are transcribed by RNA polymerase II. *The EMBO journal*, *23*(20), 4051–60. doi:10.1038/sj.emboj.7600385

Liang, H., & Li, W.-H. (2009). Lowly expressed human microRNA genes evolve rapidly. *Molecular biology and evolution*, *26*(6), 1195–8. doi:10.1093/molbev/msp053

Lim, L. P., Lau, N. C., Weinstein, E. G., Abdelhakim, A., Yekta, S., Rhoades, M. W., … Bartel, D. P. (2003). The microRNAs of Caenorhabditis elegans. *Genes & development*, *17*(8), 991–1008. doi:10.1101/gad.1074403

Lindow, M., & Gorodkin, J. (2007). Principles and limitations of computational microRNA gene and target finding. *DNA and cell biology*, *26*(5), 339–51. doi:10.1089/dna.2006.0551

Liu, X., & Theil, E. C. (2004). Ferritin reactions: direct identification of the site for the diferric peroxide reaction intermediate. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(23), 8557–62. doi:10.1073/pnas.0401146101

Long, J.-E., & Chen, H.-X. (2009). Identification and characteristics of cattle microRNAs by homology searching and small RNA cloning. *Biochemical genetics*, *47*(5-6), 329–43. doi:10.1007/s10528-009-9234-6

McGinnis, S., & Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic acids research*, *32*(Web Server issue), W20–5. doi:10.1093/nar/gkh435

Meng, F., Hackenberg, M., Li, Z., Yan, J., & Chen, T. (2012). Discovery of novel microRNAs in rat kidney using next generation sequencing and microarray validation. *PloS one*, *7*(3), e34394. doi:10.1371/journal.pone.0034394

Millar, A. a, & Waterhouse, P. M. (2005). Plant and animal microRNAs: similarities and differences. *Functional & integrative genomics*, *5*(3), 129–35. doi:10.1007/s10142-005-0145-2

Mor, E., & Shomron, N. (2013). Species-specific microRNA regulation influences phenotypic variability: perspectives on species-specific microRNA regulation. *BioEssays : news and reviews in molecular, cellular and developmental biology*, *35*(10), 881–8. doi:10.1002/bies.201200157

Morin, R. D., Aksay, G., Dolgosheina, E., Ebhardt, H. A., Magrini, V., Mardis, E. R., … Unrau, P. J. (2008). Comparative analysis of the small RNA transcriptomes of Pinus

contorta and Oryza sativa. *Genome research*, *18*(4), 571–84. doi:10.1101/gr.6897308

Nam, J.-W., Kim, J., Kim, S.-K., & Zhang, B.-T. (2006). ProMiR II: a web server for the probabilistic prediction of clustered, nonclustered, conserved and nonconserved microRNAs. *Nucleic acids research*, *34*(Web Server issue), W455–8. doi:10.1093/nar/gkl321

Nam, J.-W., Shin, K.-R., Han, J., Lee, Y., Kim, V. N., & Zhang, B.-T. (2005). Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic acids research*, *33*(11), 3570–81. doi:10.1093/nar/gki668

Nelson, P. T., Baldwin, D. O. N. A., Kloosterman, W. P., Kauppinen, S., Plasterk, R. H. A., & Mourelatos, Z. (2006). RAKE and LNA-ISH reveal microRNA expression and localization in RAKE and LNA-ISH reveal microRNA expression and localization in archival human brain, 187–191. doi:10.1261/rna.2258506.Expression

Ng, K. L. S., & Mishra, S. K. (2007). De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics (Oxford, England)*, *23*(11), 1321–30. doi:10.1093/bioinformatics/btm026

Oulas, A., Boutla, A., Gkirtzou, K., Reczko, M., Kalantidis, K., & Poirazi, P. (2009). Prediction of novel microRNA genes in cancer-associated genomic regions--a combined computational and experimental approach. *Nucleic acids research*, *37*(10), 3276–87. doi:10.1093/nar/gkp120

Pang, K. C., Frith, M. C., & Mattick, J. S. (2006). Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends in genetics : TIG*, *22*(1), 1–5. doi:10.1016/j.tig.2005.10.003

Pasquinelli, a E., Reinhart, B. J., Slack, F., Martindale, M. Q., Kuroda, M. I., Maller, B., … Ruvkun, G. (2000). Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, *408*(6808), 86–9. doi:10.1038/35040556

Pfeffer, S., Sewer, A., Lagos-Quintana, M., Sheridan, R., Sander, C., Grasser, F. A., … Tuschl, T. (2005). Identification of microRNAs of the herpesvirus family. *Nat Meth*, *2*(4), 269–276. Retrieved from http://dx.doi.org/10.1038/nmeth746

Powers, D. (2011). Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63.

Ritchie, W., Gao, D., & Rasko, J. E. J. (2012). Defining and providing robust controls for microRNA prediction. *Bioinformatics (Oxford, England)*, *28*(8), 1058–61. doi:10.1093/bioinformatics/bts114

Saçar, M. D., & Allmer, J. (2014). Machine Learning Methods for MicroRNA Gene Prediction. In M. Yousef & J. Allmer (Eds.), *miRNomics: MicroRNA Biology and Computational Analysis SE - 10* (Vol. 1107, pp. 177–187). Humana Press. doi:10.1007/978-1-62703-748-8_10

Saçar, M. D., Hamzeiy, H., & Allmer, J. (2013). Can MiRBase provide positive data for machine learning for the detection of MiRNA hairpins? *Journal of integrative bioinformatics*, *10*, 215. doi:10.2390/biecoll-jib-2013-215

Sewer, A., Paul, N., Landgraf, P., Aravin, A., Pfeffer, S., Brownstein, M. J., … Zavolan, M. (2005). Identification of clustered microRNAs using an ab initio prediction method. *BMC bioinformatics*, *6*, 267. doi:10.1186/1471-2105-6-267

Steffen, P., Voss, B., Rehmsmeier, M., Reeder, J., & Giegerich, R. (2006). RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics (Oxford, England)*, *22*(4), 500–3. doi:10.1093/bioinformatics/btk010

Tax, D. M. J. (2001). *One-class classification*.

Terai, G., Komori, T., Asai, K., & Kin, T. (2007). miRRim : A novel system to find conserved miRNAs with high sensitivity and specificity, 2081–2090. doi:10.1261/rna.655107.been

Van der Burgt, A., Fiers, M. W. J. E., Nap, J.-P., & van Ham, R. C. H. J. (2009). In silico miRNA prediction in metazoan genomes: balancing between sensitivity and specificity. *BMC genomics*, *10*, 204. doi:10.1186/1471-2164-10-204

Wang, G., Walt, J. M. Van Der, Mayhew, G., Li, Y., Zu, S., Scott, W. K., … Vance, J. M. (2008). Variation in the miRNA-433 Binding Site of FGF20 Confers Risk for Parkinson Disease by Overexpression of a -Synuclein, (February), 283–289. doi:10.1016/j.ajhg.2007.09.021.

Wang, X., Zhang, J., Li, F., Gu, J., He, T., Zhang, X., & Li, Y. (2005). MicroRNA identification based on sequence and structure alignment. *Bioinformatics (Oxford, England)*, *21*(18), 3610–4. doi:10.1093/bioinformatics/bti562

Wienholds, E., Kloosterman, W. P., Miska, E., Alvarez-saavedra, E., Berezikov, E., Bruijn, E. De, … Plasterk, R. H. A. (2005). MicroRNA Expression in Zebrafish Embryonic Development Ant Nestmate and Non-Nestmate Discrimination by a Chemosensory Sensillum, (July), 310–311.

Wu, Y., Wei, B., Liu, H., Li, T., & Rayner, S. (2011). MiRPara: a SVM-based software tool for prediction of most probable microRNA coding regions in genome scale sequences. *BMC bioinformatics*, *12*(1), 107. doi:10.1186/1471-2105-12-107

Xu, G., Zhang, Y., Jia, H., Li, J., Liu, X., Engelhardt, J. F., & Wang, Y. (2009). Cloning and identification of microRNAs in bovine alveolar macrophages. *Molecular and cellular biochemistry*, *332*(1-2), 9–16. doi:10.1007/s11010-009-0168-4

Xu, Y., Zhou, X., & Zhang, W. (2008). MicroRNA prediction with a novel ranking algorithm based on random walks. *Bioinformatics (Oxford, England)*, *24*(13), i50–8. doi:10.1093/bioinformatics/btn175

Xue, C., Li, F., He, T., Liu, G.-P., Li, Y., & Zhang, X. (2005). Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC bioinformatics*, *6*, 310. doi:10.1186/1471-2105-6-310

Yi, R., Qin, Y., Macara, I. G., & Cullen, B. R. (2003). Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs, 3011–3016. doi:10.1101/gad.1158803.miRNA

Yousef, M., Jung, S., Showe, L. C., & Showe, M. K. (2008). Learning from positive examples when the negative class is undetermined--microRNA gene identification. *Algorithms for Molecular Biology: AMB*, *3*, 2. doi:10.1186/1748-7188-3-2

Zeng, Y., & Cullen, B. R. (2004). Structural requirements for pre-microRNA binding and nuclear export by Exportin 5. *Nucleic acids research*, *32*(16), 4776–85. doi:10.1093/nar/gkh824

Zhang, B.-T., & Nam, J.-W. (2008). Supervised Learning Methods for MicroRNA Studies. In *Machine Learning in Bioinformatics* (pp. 339–365). John Wiley & Sons, Inc. doi:10.1002/9780470397428.ch16

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

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, *31*(13), 3406–3415. doi:10.1093/nar/gkg595